U.PORTO
FLUP FACULDADE DE LETRAS
UNIVERSIDADE DO PORTO

3º CICLO DE ESTUDOS EM CIÊNCIAS DA LINGUAGEM
TECNOLOGIAS DA LINGUAGEM

Native Language Influence Detection

Cross-Linguistic Analysis of Health Sciences Original Scientific Research Articles

Milaydis Sosa Napolskij

D

2021

Milaydis Sosa Napolskij

Native Language Influence Detection Cross-Linguistic Analysis of Health Sciences Original Scientific Research Articles

Tese realizada no âmbito do Doutoramento em Ciências da Linguagem, orientada pelo Professor Doutor Rui Manuel Sousa Silva

Faculdade de Letras da Universidade do Porto

2021

Milaydis Sosa Napolskij

Native Language Influence Detection Cross-Linguistic Analysis of Health Sciences Original Scientific Research Articles

Tese realizada no âmbito do Doutoramento em Ciências da Linguagem, orientada pelo Professor Doutor Rui Manuel Sousa Silva

Membros do Júri

Presidente:

Doutora Maria de Fátima Aires Pereira Marinho Saraiva, Professora Catedrática do Departamento de Estudos Portugueses e Estudos Românicos da Faculdade de letras da Universidade do Porto

Vogais:

Doutor Manuel Célio Conceição, Professor Associado da Universidade do Algarve

Doutora Karen Bennett, Professora Associada da Universidade Nova de Lisboa

Doutor Thomas Juan Carlos Husgen, Professor Associado da Faculdade de Letras da Universidade do Porto

Doutor Paulo Jorge de Sousa Oliveira Santos, Professor Auxiliar da Faculdade de Letras da Universidade do Porto

Doutora Elena Zagar da Cunha Galvão, Professora Auxiliar da Faculdade de Letras da Universidade do Porto

Doutora Belinda Mary Harper Sousa Maia, Especialista do domínio científico em que se insere a tese

Doutor Rui Manuel Sousa Silva, Professor Auxiliar da Faculdade de Letras da Universidade do Porto

A mis padres,

A mi hermana,

A mi amor, Carlos,

A Diego, mi caballero,

A mi familia presente y ausente,

A mis amigos de Cuba y de Portugal.

Table of Contents

Declaration of honor

I declare that this thesis is of my authorship and has not been previously used in another course or curricular unit of this or any other institution. References to other authors (statements, ideas, thoughts) scrupulously respect the rules of attribution, and are duly indicated in the text and in the bibliographical references, in accordance with the referencing rules. I am aware that the practice of plagiarism and self-plagiarism is an academic offence.

Declaro que a presente tese é de minha autoria e não foi utilizada previamente noutro curso ou unidade curricular, desta ou de outra instituição. As referências a outros autores (afirmações, ideias, pensamentos) respeitam escrupulosamente as regras da atribuição, e encontram-se devidamente indicadas no texto e nas referências bibliográficas, de acordo com as normas de referenciação. Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Porto, 20 de outubro de 2021

Milaydis Sosa Napolskij

Acknowledgements

Al profesor Rui Sousa Silva, impulsor de la lingüística forense en Portugal, por aceptar el desafío de supervisar este proyecto en primera persona. La imposibilidad de dedicarme a este estudio a tiempo completo hizo que el camino fuese el doble de extenso y trabajoso. Su paciencia, dedicación, conocimientos y confianza en mis ideas han permitido llevarlo a buen puerto. Espero poder retribuirle con creces.

A la profesora Belinda Maia, por su generosidad intelectual, por darme a conocer el mundo de la traducción especializada, la terminología y la lingüística de corpus; por mostrarme el valor de la interdisciplinaridad y por dedicar una parte de su tiempo a la revisión lingüística de esta tesis.

Agradezco al Profesor Altamiro da Costa Pereira, director de la Facultad de Medicina de la Universidad de Porto y Coordinador Científico de la Unidad de Investigación CINTESIS, por creer en mí, por dejarme ser y crecer; y al Doctor Antonio Soares, gestor ejecutivo de la misma unidad, mil gracias, por propiciarme una gestión autónoma del tiempo que me permitió muchas horas de estudio e investigación en tiempos neuronalmente útiles.

Estaré eternamente agradecida a mis colegas y amigos de CINTESIS. Muy especialmente a Olga Magalhães, por su ejemplo, sus muchas palabras de ánimo, su empatía y su liderazgo; a Claudia Azevedo, por su fuerza, por seguir el mismo camino difícil y por su dulzura endógena e infinita; a Bárbara Mota y a Pedro Sacadura –*the cool ones*– por el compañerismo y el buen gusto para todo; y a Ana Maria Gomes por su fuerza de carácter que me recuerda que vale la pena ser nosotros mismos y porque su fe en mí disipó muchas veces mis miedos de no estar a la altura de las circunstancias.

A mis antiguos colegas del Instituto de Ciencias Biomédicas de Abel Salazar de la Universidad de Porto (ICBAS-UP) y, en especial, del Departamento de Farmacología e Inmuno-Fisiología donde comenzó este viaje académico en 2013. Muy particularmente a Belmira Silva, Ana Patrícia Sousa, Margarida Araújo, Laura Oliveira, Graça Lobo y a la profesora Paula Ferreira da Silva por las palabras de ánimo, el cariño y por enseñar con las acciones.

A los profesores e investigadores de la Unidad Multidisciplinar de Investigación Biomédica (UMIB) del ICBAS-UP y muy especialmente a los profesores Mariana Monteiro,

Paula Jorge, Marco Alves, Pedro Oliveira, y Mario Sousa, y a las doctoras Sofia Pereira y Raquel Bernardino.

A Encarnação por su convicción en mí, por su fuerza y amor a la vida, su fe en el mundo, su alegría, su constancia y su disponibilidad.

A Isabel Silva por tanto cariño, palabras de apoyo y confianza, por estar siempre en las alturas ciertas, y su fuerza de carácter.

A Cristina Pinhão por las muchas conversaciones y discusiones sobre traducción de términos y otras cosas.

A Viviane Maia por el compañerismo, el intercambio constante de ideas y por aquella complicidad latinoamericana y femenina. "*Oh raça*!"

A mis amigos de muchos años Hugo Simões y Raquel Silva porque tenerlos en mi vida me hace querer ser mejor ser humano.

A mi amiga Yonay quien me dedicó algunas horas de su brevísimo tiempo y leyó parte de esta tesis. Sus comentarios ayudaron a disipar varias dudas y calmar mis desasosiegos. Quiero agradecerle por la disponibilidad, el cariño, la fuerza, y la eterna amistad.

La conclusión de este trabajo se la debo en gran medida a mi gran amiga Milena Paneque porque supo insuflarme entusiasmo, motivación y ganas de hacer cuando yo ya no tenía nada. Su inteligencia, claridad, humanidad, cariño, mesura y amparo permanentes hacen de cualquier derrotado un vencedor. "*Good things take time!*".

Esta tesis es sobre todo una prueba fehaciente de la confianza incuestionable que Carlos siempre ha tenido en mí. Una confianza mucho mayor que aquella que alguna vez he tenido en mí misma. Su amor por mí es uno de los pilares más sólidos e inquebrantables de mi existencia. Mi amor por él es para siempre.

Este trabajo también es prueba irrefutable de la serenidad y el amor inconmensurables de mi hijo Diego a quien le quité muchas, aunque no todas, de nuestras horas de convivencia familiar. Él es mi reparador de sueños, mi luz, mi alegría más dulce, mi paz. Todo este esfuerzo ha sido sobre todo por él.

A mis padres, Liubov Napolskij y José Antonio Sosa García, porque ya me admiraban cuando todavía no estaba ni cerca de cumplir esta meta. Les agradezco por su amor, su preocupación constante, su ejemplo, por ser los cimientos de mi carácter, de mis ganas de hacer y mi fuerza interior.

A mi hermana pequeña, Galina Sosa Napolskij, que parece haber nacido preparada pero que siempre ha trabajado mucho y bien para alcanzar lo que se propone. Su ejemplo de crecimiento personal y profesional me mostró, primero que yo a ella, que sí se puede, que lo valemos y nos lo merecemos.

A mi familia presente y a mi familia ausente porque son el marco de mis días pasados y mis días futuros.

A mis compañeros y profesores de la Facultad de Lenguas Extranjeras de la Universidad de La Habana, donde me inicié en la lingüística aplicada.

Resumo

As limitações impostas pela homogeneização linguística dos géneros científicos deveriam funcionar, previsivelmente, como uma barreira à distinção entre indivíduos e à sua língua materna. No entanto, os autores de artigos científicos utilizam cada vez mais características da sua língua nativa na sua escrita científica em inglês. Apesar de esta influência ter sido abordada em estudos culturais e sobre género textual, a perspetiva da autoria tem sido menos investigada. Esta tese contribui para o campo da deteção da influência da língua nativa, analisando os padrões de influência translinguística no texto científico. Discute-se se existem variáveis associadas à influência da língua nativa em artigos originais de investigação científica escritos em inglês por autores nativos das variedades europeias de português e de espanhol. Procura-se identificar essas variáveis e explicar a influência das línguas nativas, bem como as possíveis implicações dessa influência. O trabalho adota uma abordagem comparativa, assumindo um modelo especialmente concebido para estudar a influência da língua nativa e combinando estatística e linguística. As análises propostas baseiam-se no Corpora Comparativo de Artigos de Investigação – CORA, uma coleção de cinco corpora especializados com 825 403 tokens construídos especificamente para este estudo pela autora. Os resultados empíricos mostram que existem variáveis não linguísticas ou de estilo (por exemplo, frequência dos pronomes demonstrativos) e variáveis linguísticas ou de conteúdo (por exemplo, o uso da conjunção coordenativa "as well as" ou o advérbio "namely") que podem indicar a influência da língua nativa em autores portugueses/espanhóis de artigos científicos escritos em inglês. Da mesma forma, este estudo revela a associação de diversas variáveis linguísticas a estratégias utilizadas pelos autores não nativos para evitar o recurso a certas formas linguísticas na redação de artigos científicos em inglês. Este trabalho demonstra que a influência da língua materna também pode ser detetada em géneros altamente especializados, em particular se considerarmos aspetos sintáticos.

Palavras-chave: Análise de Autoria, Deteção da Influência da Língua Nativa, Escrita Científica, Texto Académico, Transferência Linguística.

Abstract

Although the constraints imposed by the linguistic homogenization of scientific genres should not allow any relevant linguistic distinction among individuals, scientific authors are increasingly using features of their L1 in their scientific writing in English. While this influence has been examined in genre and cultural studies, the authorship perspective has received less attention. This thesis contributes to the field of native language influence detection by examining the patterns of cross-linguistic influence on the scientific text. It discusses whether there are variables associated with the influence of the native language in original scientific research articles written in English by non-L1 English authors who are native users of the European varieties of Portuguese and Spanish. It attempts to identify these variables and to explain the influence of the native languages, and whether their existence has implications, for example, in teaching scientific English. The work adopts a comparison-based approach taking on a model specially designed for examining L1 influence and combining statistics and linguistics. The analyses are based on the Comparative Corpora of Research Articles – CORA, five specialized corpora with 825 403 tokens purposely built for this study by the author. The empirical results show that there are content-independent variables (e.g. frequency of demonstrative pronouns) and content-dependent variables (e.g. the use of the coordinative conjunction “as well as” or the adverb “namely”) that can indicate the influence of the Portuguese/Spanish authors’ L1 in the OSRAs they produced in English. Moreover, several content-dependent variables were associated with possible strategies of avoidance of use by these authors when writing in English. This work demonstrates that, besides texts like twitter posts, L1 influence can also be detected in highly specialized genres, especially if one takes syntactic features into consideration.

Keywords: Authorship Analysis, Native Language Influence Detection, Scientific Writing, Academic Text, Language Transfer.

List of Figures

List of Tables

1. Introduction

According to data on human resources working in Science, Technology and Innovation available in the website of the United Nations Educational, Scientific and Cultural Organization – UNESCO (http://data.uis.unesco.org/) and in their latest statistical report ("UNESCO Science Report: Towards 2030" 2016: 33) there are approximately 12 million full-time equivalent (FTE) research personnel in the world[1], of which near 7.8 million are researchers. About 75% of those 7.8 million are researchers from non-English speaking countries, making it clear that most scientific researchers in the world are native speakers of languages other than English.

Nonetheless, as is the case for many areas of today's societies, it is English that functions as the lingua franca of scientific communication. This predominance has increased over the last fifty years, mainly through globalization processes that affect all dimensions of society.

Communicating in one common language provides universality, which brings distinct advantages to the scientific community. It allows for an easier exchange of up-to-date scientific data and a better understanding of the scientific problems of each field. In other words, English functions as a common ground for dissemination of knowledge that, in the long term, serves the advancement of science.

However, communicating knowledge is difficult; doing so in a non-native language takes the process to a higher difficulty level. Besides using English to communicate in academic contexts as scientists, professors, experts, and entrepreneurs, non-native English-speaking scientific communities function in their respective societies and cultures, normally using their native languages in domestic and social settings, as citizens. As a result, a situation persists in which non-native English-speaking scholars have to alternate between their

[1] Refers to 116 countries, i.e., those in the EU, plus Algeria, Angola, Argentina, Australia, Bahrain, Bosnia and Herzegovina, Botswana, Brazil, Burundi, Cabo Verde, Cambodia, Canada, Chad, Chile, China, China, Hong Kong Special Administrative Region, China, Macao Special Administrative Region, Colombia, Costa Rica, Democratic Republic of the Congo, Ecuador, Egypt, El Salvador, Eswatini, Ethiopia, Gambia, Georgia, Ghana, Guatemala, Honduras, Iceland, India, Indonesia, Iran (Islamic Republic of), Iraq, Japan, Jordan, Kazakhstan, Kuwait, Lesotho, Madagascar, Malaysia, Mali, Mauritius, Mexico, Montenegro, Morocco, Mozambique, Myanmar, Namibia, New Zealand, Niger, North Macedonia, Norway, Oman, Pakistan, Palestine, Panama, Papua New Guinea, Paraguay, Philippines, Puerto Rico, Qatar, Republic of Korea, Republic of Moldova, Russian Federation, Rwanda, Senegal, Serbia, Singapore, South Africa, Sri Lanka, Switzerland, Syrian Arab Republic, Thailand, Togo, Trinidad and Tobago, Tunisia, Turkey, Uganda, Ukraine, United Arab Emirates, United Kingdom of Great Britain and Northern Ireland, United Republic of Tanzania, United States of America, Uruguay, Uzbekistan, Venezuela (Bolivarian Republic of), Viet Nam.

respective native languages and English as a professional language used to communicate their science.

The linguistic dichotomy that non-native English scientists face may impact their scientific communication, both oral and written. In oral scientific communication by non-native English speakers, one can expect certain imperfections in the linguistic output to occur due to the immediacy and spontaneity associated with conversations. However, written scientific communication in non-native English presents a linguistic setting that demands highly proficient language skills from non-native English authors who are expected to perform as if they were using their native language.

Additionally, written scientific communication takes place mainly within scientific genres. These genres carry the rhetoric heritage of the writing strategies of their Anglophone culture and language of birth, English (Swales 1990: 111-17). Non-native English authors have to comply with such a tradition, even though it is usually very different from theirs. When the conventions of a genre are combined with native writing, engagement, persuasion, and argumentation strategies of non-native English authors, the resulting linguistic output may be influenced by the native language of the authors (Mauranen, Pérez-Llantada, and Swales 2010: 642-46).

This rationale may seem unlikely since the stylistic conditioning imposed by the linguistic homogenization of scientific genres should not allow any relevant linguistic distinction among writers. The high level of textual standardization, with domain and field-specific rules and conventions established for scientific genres by journals, scientific societies, and even faculties, should operate as boundaries for the language authors use. It should also guarantee the employment of appropriate linguistic patterns, controlling language to the point of not allowing certain stylistic marks to pass through.

Nevertheless, authorship of scientific writing has changed significantly as the "postmodern era" has gradually transformed the initial assumptions that scientific genres and writing styles are a guarantee of the constraint of the authors' "authentic voice", and encourage concealment of references to national culture and dialectal makers of discourse (Pérez-Llantada 2012: 163).

In my professional context, I have been observing this influence for over ten years, while proofreading research articles written in English by European Portuguese researchers from the health sciences. This thesis results from my interest in researching native language influence in scientific writing.

Therefore, this research seeks to contribute to the field of native language influence detection (NLID) by adding to:

- the examination of the kind of texts profiled within authorship analysis;
- the linguistic viewpoint of analysis with support of quantitative data;
- the applications of NLID in general, and specifically, within translingual plagiarism detection,
- the description of less addressed languages within NLID, particularly of Portuguese and Spanish.

The overall purpose of this study is to examine the patterns of cross-linguistic influence on scientific text written in English by non-native authors (non-L1). It focuses specifically on one genre, original scientific research articles (OSRA), within the field of health sciences. The study is also circumscribed to non-L1 English speaking authors who are native (L1) speakers of the European varieties of two Romance languages, Portuguese (PT-EU) and Spanish (ES-EU).

1.1. Research Questions

Based on the overall purpose stated above, the present study seeks to answer the following research questions:

1. **Are there variables associated with L1 influence in OSRAs written in English by PT-EU and ES-EU L1 authors in the field of health sciences? If so,**

   1.1 what are those variables?

Scientific genres, and particularly research articles communicating new knowledge based on, for example, experimental results, i.e., original research, are known to follow very specific conventions and rhetorical organization. Therefore, OSRA authors are obliged to report their science constrained by the specific rules of OSRAs parts: Introduction, Methodology, Results, Discussion and the Conclusions, known by the acronym IMRAD (Swales 1990). One of the functions of all of these constraints is to modulate authorship markers so that research articles comply with the "basic purposes of scientific publications [that] are (1) to educate, (2) to inform, and (3) to record […] (4): to persuade" (Day, Sakaduski, and Day 2011: 1) and do so in a formal manner observing the correct use of English in relation to aspects that go from morpho-syntax to discourse. Thus, scientific authors are limited in their linguistic choices, such as metaphors (Day, Sakaduski, and Day 2011: 37) that do not comply with scientific writing. Besides this restriction in linguistic choices, authors who are non-L1 users of English are also expected not to leave linguistic traces of their native languages in the scientific text they produce in English.

This is the fundamental question of this empirical research. Knowing that the postmodern era has enabled non-L1 English authors to have a voice by allowing the combination of "normative" with "local" characteristics (Pérez-Llantada 2012), this study seeks to investigate if non-L1 English authors leave traces of their native languages that can be observed in certain language-related variables when using scientific English as a functional rather than an identity variety of the language (Pérez-Llantada 2012: 165). If that is the case, I then seek to determine which variables mark L1 influence. This investigation implies the comparison of OSRAs written by L1 English authors and authors who are L1 users of European Portuguese and European Spanish writing OSRAs in English. Also, the latter groups are compared with Portuguese and Spanish authors writing OSRAs in their respective L1s.

2. Is it possible to explain the absence/presence of L1 influence variables in OSRAs written in English by L1 authors of PT-EU and ES-EU L1?

Variationist sociolinguistics has demonstrated that linguistic change does not occur exclusively over a long period of time, but that it is possible to observe change in a linguistic sample collected over a short period of time (Labov 1963). Academic English has changed rapidly in the last sixty years (Pérez-Llantada 2012). Part of that change is justified by the participation in academic production of authors with diverse linguistic and cultural backgrounds. The influence of a native language in a foreign language has been addressed by researchers in the field of Second Language Acquisition (SLA) since the 1980s, and they have proposed different theoretical and conceptual frameworks to explain how this influence occurs. Special attention is given to the Theory of Interlanguage (Selinker 1972, 2014). The second research question proposes to provide the possible explanations of the variables that are found in the empirical study, and contribute to their linguistic description in relation to other relevant studies in the field of L1 influence.

3. **Are there implications associated with the absence/presence of L1 influence variables in OSRAs written in English by L1 authors of PT-EU and ES-EU L1?**

This question seeks to reflect on the implications of the absence/presence of variables that can indicate the influence of the L1 in authors who are non-L1 users of English when writing OSRAs in that language. Several implications of diverse natures can be anticipated. At the very least, this study can contribute to the characterization of scientific English. Another implication would be of an instructional character. A third implication could be related to direct professional significance for those proofreading and editing OSRAs in the health sciences in Portugal/Spain. Similarly, the study could have implications for translators working with Portuguese or Spanish and English in the health sciences.

1.2. Structure of the thesis

This work has two main parts. The first part is dedicated to the theoretical and conceptual background of the study (chapter 2), and the second contains empirical work (chapters 3 to 5).

Given the distinct interdisciplinary nature of the topic of this research work, the literature review attempts to address all the relevant concepts concerning native and other language influence detection applied to the scientific text. First, chapter 2 examines authorship analysis as the parent field of authorship profiling and native language influence detection. This is followed by a discussion of the theoretical and conceptual frameworks relevant to the topic, and an examination of language variation in the form of idiolect, dialect, genre, and style. Idiolect is examined in the light of the theoretical discussions about its existence, and their relevance for analyzing authorship. The next section discusses several theories and conceptual frameworks on language transfer, and then examines the intersection between native language influence detection and scientific writing. Scientific writing is then discussed in terms of genre and register while also detailing the discourse community and community of practice.

The second part of this thesis begins with chapter 3, which starts with a description of the corpora compilation process and examines the challenges of building own corpora and the reasons for assuming such a challenge. Then, there is a description of the final corpora and an explanation of the methods, procedures and tools used in the empirical work. The last section discusses the study design and outlines the operationalization of the research questions.

Chapter 2 and Chapter 3 end with summary sections that might be helpful for those who need to understand the main point of this research before reading the chapter on the findings and discussion.

Chapter 4 presents the analyses carried out with the corpora to detect native language influence in scientific writing while discussing their relevance and potential to be considered markers of native language influence. The chapter has a section discussing the results and summarizing the most relevant findings.

Finally, in Chapter 5, I outline the main contributions to native language influence detection, particularly in scientific texts, presenting the limitations to the study and proposals for future work.

2. NLID in Scientific Writing

This chapter discusses language background profiling, specifically native language influence detection (NLID), in scientific writing. First, I examine the concept of authorship analysis, and propose a working definition. Next, I address the development of authorship studies in literary, non-literary, and forensic contexts while considering some of the approaches adopted to solve authorship problems. Then, I present the concept of authorship profiling. I analyze the most relevant factors that can be determined by profiling authorship, and examine the concept of native language influence detection. This is followed by an exploration of the most important theoretical and conceptual frameworks in the field in terms of their usefulness to explain native language influence detection. After that, the discussion addresses the original scientific research article as a genre. I examine the scarcity of studies addressing the influence of the sociolinguistic variable of language background in scientific writing and demonstrate the pertinence of filling such a gap.

2.1. Authorship Analysis and NLID

The study of authorship has long been the research object of scholars from different fields. The investigation of authorship can be traced to almost 2500 years ago. There is evidence showing that the scholars of the Greek museum and library of Alexandria used to work on the systematic research of the writing style –including sentence structure and choice of words– of the work of celebrated poets like Homer for purposes of attribution or rejection of authorship (Love 2002: 14; Coulthard, Johnson, and Wright 2017: 152).

Authorship studies can be said to have been originated in Stylistics, which in turn was preceded by Rhetoric, a field dating back to the fifth century B.C. “concerned with the use of public speaking as a means of persuasion” (Bradford 2005: 2). Rhetoric opened the path to Stylistics inasmuch as modern literary studies began developing around 1850-1900, and the increasing specialization of the field led to a natural interest in literary authorship, which was approached through the analysis of “special expressions” and stylistic devices (Yllera 1979: 11-15).

The inception of stylistics as a contemporary field of applied linguistics and a method in literary studies is considered to have taken place at the beginning of the 20th century with Russian Formalism (around 1915-1930), followed by the Prague School (around 1926-1939). The first rejected the thought-provoking component of literature proposed by the impressionists and focused on the authors' words as the object whose analysis allowed the reader to study literature and, thus, language. The latter reformulated the formal approach of Russians, establishing that “language is the main sign system, but not the only one,” presenting literature as part of semiology and not purely of linguistics (Yllera 1979: 94); and including “context in textual meaning-making” which eventually gave rise to the functional approach for the study of language and authorship (Burke 2014: 2).

Authorship studies found in Stylistics a natural inaugural space for development, and as the whole field of Linguistics consolidated over the 20th century, authorship studies also developed, especially with regards to three main aspects. The first refers to the kinds of texts addressed within the field, which have expanded from complete literary or religious works to short pieces of writing published online, such as Twitter messages. The second focuses on the research methodologies used to analyze authorship, which have gone from qualitative approaches requiring extensive knowledge of the works and author(s) in analysis and academic training in linguistics, literature, cultural studies, and related fields, to purely quantitative methods based on statistics and carried out with sophisticated software; and, finally, to, a combination of both and consolidation of the interdisciplinary nature of authorship studies. Finally, the third aspect is the applications such analyses can have. Applications of authorship analysis have broadened from purely scholarly purposes of gaining knowledge or resolving historical authorship disputes or unknown authorship problems to more practical uses. Some of these uses are, for example, obtaining demographic information on consumers of a product to customize marketing campaigns, identifying deceptive customer reviews of products or services, providing evidence on the identity of individuals to solve criminal cases, or detecting plagiarism (Juola 2015: 22-23).

Before examining the evolution of the fields that deal with these aspects, let us explore the concept of authorship analysis.

2.1.1. Defining Authorship Analysis

Authorship can be a complex matter. Thus, it is important to emphasize that the authorship mentioned in this section and throughout this research study concerns first and foremost the agency of the act of producing text, but also the context(s) surrounding the production and its result (Love 2002: 32). Oral production can also be a target of authorship analysis (Love 2002: 32-39; Juola 2008: 6) since its complexity, as claimed by Goffman (1981: 144) almost four decades ago, is embodied in the different "roles of utterance production" of the speaker. However, in this study, the focus is on written text.

Another important factor in relation to authorship agency is that as with speech, written authorship is rarely a truly individual act as there are usually many authorship functions to fulfill in the writing process, and not all can be played by the actual writer (Foucault 1979; Love 2002: 39-50).

Love (2002) describes authorship as displaying four functions, i.e., precursory, executive, declarative, and revisionary. *Precursory authorship* is defined as "cases in which a significant contribution from an earlier writer is incorporated into the new work" (Love 2002: 40). *Executive authorship* refers to "the compiler of the verbal text up to the point where it is judged suitable for publication in one or another form (all subsequent alterations being classified as revisions)" and it is a type of authorship that can be performed as a single author or collaboratively (Love 2002: 43). *Declarative authorship* takes place when the author acts as a "validator […] placing [his/her] name upon the title-page [to] indicate a combination of precursory authorship and a form of sponsorship or fostering" of the content, but not the person who performs the actual writing (Love 2002: 44). Finally, *revisionary authorship* refers to cases of "editing" where "a second writer or editor remodels a work completed or in some cases abandoned by a first" author (Love 2002: 46-47). In this work, the agency of authorship refers to the writer, i.e., *executive authorship*, but it also considers the *precursory*, the *declarative*, and the *revisionary* functions of authorship.

The analysis of textual authorship has historically been associated with the need to answer questions concerning the agent who has created a given piece of writing, by addressing written style or "the recurrent [language] choices that the writer makes" as its main object of study (McMenamin 2002: 126). Authorship analysis has been the focus of interest of researchers working in different fields like literature (Miranda 2016; Calero 2006; Migueláñez 2019), forensic linguistics (Coulthard, Johnson, and Wright 2017; Kredens, Perkins, and Grant 2020; Grant and MacLeod 2018b; Sousa-Silva 2019), and computational sciences (Juola 2015; Malmasi and Dras 2018), and the term usually coexists with the term authorship attribution.

In the intersection of law studies, linguistics, and the forensic field, Coulthard, Johnson, and Wright (2017: 151) define authorship attribution as "the process in which linguists set out to identify the author(s) of disputed, anonymous or questioned texts". Because this is a definition designed for forensic linguistics, authorship analysis is centralized in the linguist, endowing him/her with an important share of professional and scholar responsibility within the field. Such foregrounding of the linguists' role in relation to authorship analysis bespeaks an intention of claiming the 'natural' space of linguists within the field of forensic linguistics. The definition also focuses on the operational character of the analysis and the 'forensic' characteristics of the texts.

Within the field of computer sciences and information retrieval, authorship attribution has been defined as "any attempt to infer the characteristics of the creator of a piece of linguistic data", "the science of inferring characteristics of the author from the characteristics of documents written by that author", or "the task of inferring characteristics of a document's author, including but not limited to identity, from the textual characteristics of the document itself" (Juola 2008: vii; 6; 2007: 120). Other authors describe authorship attribution as a field that "studies strategies for discriminating between the styles of different authors" (Raghavan, Kovashka, and Mooney 2010: 38) or "the process of examining the characteristics of a piece of work in order to draw conclusions on its authorship" (El Bouanani and Kassou 2014: 22). These definitions from the field of computational linguistics focus on three important aspects of authorship analysis: 1) its inferential or deducible nature, 2) its focus on the written text as its object of study 3) the fact that the outcome of the analysis could be only some of the author's characteristics and not necessarily the author himself/herself.

To the elements of the definitions described above, i.e., (1) written text as the object of study of authorship analysis, (2) seeking to identify the author or authorship traces, (3) using computational and linguistic methods and techniques, another aspect must be added that regards the number of authors of the written text, which should be limited to a number of “candidate authors” according to the characteristics of the case (Grant 2007: 6). Based on these elements authorship analysis is defined here as the act, process, or result of scientifically examining the style of a written text using linguistic and/or computational approaches to identify its author(s), or as many of his/her traces and/or contextual factors of the writing as possible, from a restricted number of candidates to provide a solution to an authorship problem or contribute to one.

This definition attempts to address all the main points described and alter or include others considered relevant. The main objective of the alterations proposed is to broaden the concept to fit virtually any authorship problem. The first of the elements incorporated refers to the inclusion of the nouns “act” and “result” and the verb “to examine” to characterize the actions performed in authorship analysis so that not only the task or the process of the analysis is encompassed, but also the act of analysis itself and the result of such an analysis, which can be, for example, the report a forensic linguistics expert produces on a given analysis. Another alteration regards the type of authorship problem as indicated by the authors in the field of forensic linguistics (i.e., disputed, anonymous, or questioned) and also as described by computer sciences (i.e., linguistic data, document, piece of work) since the way authorship problem is refer to in their definitions seem to go from one extreme to the other. In forensic linguistics, an authorship problem has, at the very least, legal implications, and in more complex situations, it may put those involved in the problem at the risk of imprisonment or even death, depending on the legal system. In the definitions from the computer sciences presented above (Juola 2008, 2007; Raghavan, Kovashka, and Mooney 2010; El Bouanani and Kassou 2014), the problematization of authorship issues is practically omitted, and the focus is mainly placed on the task of discriminating texts as belonging to one author or another. Therefore, in my definition, the expression ‘authorship problem’ refers to the object of analysis within written text –whether or not the problem is one of academic inquiry or disputed authorship with legal or life-threatening implications – while still placing some importance on the fact that it is a matter requiring a solution.

Having defined what authorship analysis is, the discussion now addresses the development of the field in relation to literary, religious, political, and forensic texts. The examination seeks to present briefly the approaches that have advanced knowledge within the field and allowed it to achieve its present state.

2.1.2. Authorship Analysis of Literary, Non-Literary, and Forensic Texts

The evolution of authorship analysis from the beginning of the 20th century until the present day can be observed in the diversification of the kind of texts addressed within the field. As I will show in this section, initially, the focus of authorship analysis was mainly on literary and religious pieces (Coulthard, Johnson, and Wright 2017). However, over time, political works, police statements, the law, and digital content, among many other kinds of texts, have been the object of study in this field (Mosteller and Wallace 1964; Grieve et al. 2018; Coulthard 1994; Coulthard, Johnson, and Wright 2017).

Authorship analysis of literary works has often been named traditional authorship analysis. It has been associated with the purposes of proving the authenticity of a piece of literary work or attributing the work to a potential/most likely author. Miranda (2016: 50; 2011: 157) refers to literary authorship analysis as an "assessment of the documented information on the genre's[2] […] profile[3]" and as a "defense of a position about the work's authorship upon careful and detailed analysis", respectively. As I show in the following paragraphs through the work of Malone (1787) and Calero (2006), the so-called traditional examination of authorship involves knowing the related genre, understanding literature as a body of works, and analyzing thoroughly and manually historical, biographical, and documental data and literary elements of writing style such as tone, narrative process, and the characters' attitude.

An early example of traditional authorship analysis concerns William Shakespeare's theater play, *King Henry VI*. The play's "visible inequalities" compared to other pieces known to be Shakespeare's were analyzed to show that the literary work had not been "originally and entirely composed by" Shakespeare, but by writers before him and then Shakespeare "formed" the play (Malone 1787: vii-1). This conclusion was reached upon examination of "manner and style", particularly concerning three aspects, "diction", "figures" –specifically, "allusions"- and "versification" (Malone 1787: 2-9). In relation to figures, it is shown that Shakespeare did not normally use allusions to "mythology", "classical authors", and "modern

[2] Refers to Galician-Portuguese poetry.

[3] My own translation from the orginal in Portuguese "nos propomos neste momento é equacionar o perfil inicial do género enquanto realidade documentada".

history" as were used in *King Henry VI* (Malone 1787: 2); and concerning versification, the play resembled works "produced before the time of Shakespeare" in terms of "a certain stately march", "[uniform] pauses at the end of every line", and the absence of a "redundant syllable" in the verse (Malone 1787: 4). The analysis of certain unique words, such as "proditor" and "immanity", are shown not to ever had been used in other plays indisputably authored by Shakespeare, and also phraseology, historical facts, and documental evidence were analyzed to prove that Shakespeare was not the sole author, but rather the adapter-author of a piece based on previous works (Malone 1787: 3).

A more contemporary example of *traditional authorship analysis* has been published by Calero (2006) on the anonymous Spanish novel *La Vida del Lazarillo de Tormes*. The author contrasts his authorship hypothesis –the humanist Joan Lluís Vives March– to other authors of the time, and specifically to one author –the also humanist Alfonso de Valdés – after eliminating the other authors on historical and stylistic arguments. He then uses contemporary testimonies about A. de Valdés' low Latin proficiency and lack of talent for writing such a masterpiece as *La Vida del Lazarillo de Tormes*. Additionally, he refers to interpretations of certain passages of the novel that can also be found in Vives' work but not in Valdés'. Calero (2006: 3) indicates that attribution can only occur if there is a correspondence between *La Vida del Lazarillo de Tormes* and the potential authors' previous and subsequent work. The author continues with an analysis of three aspects. The first is the themes presented in the work, which are as diverse as poverty, famine, anticlericalism, charity, piety, spirituality, morality and education, nobility, adulation, virtue, honor, hostility, philosophy, fortune, rights, Judaism, and many others (Calero 2006: 4-27). The second is the expression, which refers to the fact that the work is written in an epistolary form (Calero 2006: 31). Finally, the author analyses the style of the anonymous work and the work of both potential authors at different levels of language. At the phonological level, the use of alliteration is addressed. At the grammatical level, the author refers to the use of certain syncopated forms like '*do'* instead of '*adonde*' (in English 'what place' or 'where') and the frequency of exclamation marks. At the syntactical level, the author refers to the use of hyperbaton. At the lexical level, the preference for certain words like '*alumbrar'* ('light', 'light up' or 'illuminate'), '*tomar'* ('take') or '*recio'* ('strong', 'tough', or 'robust') are highlighted. Finally, at the phraseological level, the author refers to the recurrent use of expressions such as ‘*a la sazón*’ (‘at that time’, ‘at the time’, or ‘at the occasion’), ‘*no sé qué* (‘I don’t know what else’) or ‘*por no ser prolijo*’ (‘for not being verbose’).

As can be seen from the description above, traditional authorship analysis can be very comprehensive and time-consuming. It may also be deemed subjective, regardless of the argumentative power of the elements analyzed, because the analysis is heavily based on the researcher's perception and knowledge (Holmes 1994). Scholars in the field have addressed this aspect by turning to quantitative methods to complement qualitative results and address criticism of subjectivity (Holmes 1998).

Early references to using the quantitative approach to analyze authorship of literary works are those attempted by scholars like Fleary (1874), Ingram (1874), and Furnival (1887) from the New Shakespearean Society. These authors provided quantitative evidence showing a “steady change” in Shakespeare’s style over the 22 years (from 1589 to 1612) of playwriting (Tuldava 2004: 145).

The quantitative perspective of authorship studies in literature continued to develop. During the first half of the twentieth century, other works were published, such as a study of the utility for authorship attribution of the relative numbers of nouns, according to their frequency of occurrence (Yule 1944), and works on the relevance of computational methods like those by Erdman and Fogel (1966) and Williams (1970), all preceding the onset of computers (Love 2002: 133-34).

The quantitative perspective is still a complementary tool of analysis, which has proved to be of great relevance to attribute authorship in literature. For example, in 2013, there were many headlines about a case independently investigated by Patrick Juola and Peter Millican at the request of the press who wanted to verify their suspicion that the actual author of *The Cuckoo’s Calling,* a crime fiction novel published the same year by a debutant novelist named Robert Galbraith, was actually J.K. Rowling writing under the referred pseudonym (BBCNews 2013). Patrick Juola and Peter Millican applied computational linguistics techniques and were able to verify that J.K. Rowling was indeed very likely to be the author of *The Cuckoo’s Calling* based on similarities with her previous work. The discovery, which was later confirmed, revealed nothing more than J.K. Rowling’s wish to use a pseudonym to be associated with a genre of novel different to what she had been writing, i.e., the collection of fantasy Harry Potter novels, but it much discussion (Juola 2013).

Moreover, recently, the ETSO project: *Estilometría Aplicada al Teatro del Siglo de Oro* (Cuéllar-González and García-Luengos 2017) used the quantitative approach to analyze hundreds of Spanish plays from the Golden Years period (approximately 1492 – 1700). From such an analysis, the authorship of many plays is confirmed, and the alteration of the author of others is proposed. One of the works whose authorship is altered is *La monja alférez*, historically attributed to Pérez de Montalbán, but seemingly authored by Juan Ruiz de Alarcón, according to the analysis and as confirmed by historical documentation and literary analysis, such as metric analysis (Migueláñez 2019).

Besides literature, early authorship scholars used to focus typically on issues of disputed authorship concerning Bible-related works. It was precisely the analysis of a biblical text that caused the inception of the quantitative perspective of authorship studies, which emerged in 1851. In that year, the British mathematician Augustus De Morgan (1806-1871) suggested a method to solve the wrongly attributed authorship of the Epistle to the Hebrews to Paul the Apostle. The proposal was to compare those texts to other Pauline Epistles resorting to the average length of words calculated in characters, which De Morgan argued would allow a person to identify the authentic author of the disputed texts. De Morgan's ground-breaking suggestion to solve the authorship problem of the Epistle to the Hebrews was seconded by other authors such as Mascol (1888), who also tried to solve the authorship attribution problem of the Pauline letters with a mathematical proposal (Pavelec et al. 2008: 414).

However, one of the most important early contributors to the quantitative perspective was Mendenhall (1887, 1901), who improved De Morgan's proposed method and presented an approach with which it was possible to obtain "a graphic representation of an arrangement of words according to their length and to the relative frequency of their occurrence" (Mendenhall 1887: 238). Mendenhall hypothesized that such an arrangement could function as a distinguishing feature of an author's writing style and offered the model for validation to his peers; though the challenge was never really pursued by other scholars nor by Mendenhall (Lord 1958: 282; Love 2002: 133; Coulthard, Johnson, and Wright 2017: 153).

Eventually, it was not only literary and biblical texts that caused scholars to address authorship problems. Researchers of the nineteenth and twentieth centuries also focused on political texts. Mosteller and Wallace (1963) carried out a study of *The Federalist Papers,* a collection of 85 essays addressed to the citizens of New York to convince them "to adopt the new American Constitution" (Coulthard, Johnson, and Wright 2017: 153). The texts were published between 1787 and 1788 as anonymous essays but were later known to have been authored by three statesmen of the time: Alexandre Hamilton, James Madison, and John Jay. The authors were supposed to have each written several essays, but Hamilton and Madison claimed to be the sole writer of 12 of the texts. Many years later, Mosteller and Wallace aimed to solve the problem of the disputed authorship by conducting research using stylometric techniques, and proved that features like the frequencies of grammatical items work as *idiolectal markers* of an author's style, identifying Madison as the one likely to have authored most of the writings (Mosteller and Wallace 1963: 306). *The Federalist Papers* has been one of the most studied cases of authorship attribution, and its texts are used for educational purposes and to demonstrate de functioning of authorship attribution software like *The Signature Stylometric System* available online freeware[4].

Another famous case involving politicians is that of the Bixby Letter. Contrary to *The Federalist Papers*, the Bixby Letter is a short piece of writing authored by President Abraham Lincoln in 1864 and sent to Lydia Bixby of Boston to present his condolences, and that of a whole nation, for the death of, supposedly, her five sons in the American Civil War (Grieve et al. 2018). Doubts on the authenticity of the letter were raised around 1925 when it was discovered that the original had never been where it had been said to be, i.e., in the University of Oxford in England, and that several copies had been fabricated for profit-making purposes (Barton 1926). The authorship of the Bixby letter has been said to belong to John Hay (Burlingame 1999), who was, then, Lincoln's assistant, but it has equally been attributed to Lincoln (Emerson 2006). Recently, Grieve et al. (2018: 6) used a new method named *n-gram tracing* to discern between the two potential authors. The method aims to compare a short piece of writing with a corpus of texts known to have been produced by the candidate authors by first extracting all the n-grams (tokens, i.e., running words in a text (Scott 2018a: 536), and/or characters) of a given length (e.g., 1, 2, 3 n-gram) from the disputed text, then taking

[4] (http://www.philocomp.net/texts/signature.htm)

“a random sample of texts of equal size from each possible author” (p.6), and calculating the “percentage of n-gram types” found in the disputed text that are also found in each sample. The writer with the “highest percentage of these n-grams” (p.6) is the most probable author of the disputed text.

Until 1968, the study of authorship problems, such as deceiving authorship, was much rarer, and the participation of linguists in the field was also non-existent (Coulthard, Johnson, and Wright 2017: 152). It was only in that year that linguistics played an important role in a legal case of deceiving authorship. Such a contribution refers to a study by Svartvik (1968) pointing out discrepancies found in the confession statements attributed to Timothy John Evans –a man accused of killing his wife and infant daughter– that called the authorship of the crime and the responsibility of Mr. Evans into question. The analysis provided additional evidence that at least two parts of the statements by Timothy Evans –an illiterate person– had been heavily edited by police and not just “voluntarily and spontaneously […] dictated […] without any preliminary questioning and virtually without interruption” (Svartvik 1968: 22). The evidence referred primarily to the distribution of finite verb clauses and the usage of clauses containing the adverbs of time “then” and “also” that were more prominent in parts of the statements which Timothy Evans did not recognize as corresponding to the actual course of the events (Svartvik 1968: 45-46). As a consequence, Mr. Evans was posthumously determined innocent and granted pardon by the Queen nearly 16 years after his execution in 1950 (Coulthard, Johnson, and Wright 2017: 215; Gudjonsson 1993: 117).

This demonstration of the practical utility of a linguistic analysis marked the inception of forensic linguistics, defined as “the scientific study of language as applied to forensic purposes and contexts” (McMenamin 2002: 86) or “the interface between language and the law” (Gibbons and Turell 2008: 1) and whose interdisciplinary/multidisciplinary character has always been recognized (Rieber and Stewart 1990: 2; Gibbons and Turell 2008: 1).

Within the forensic context, cases like that of Derek Bentley's have been solved thanks to the participation of linguists (Coulthard 1994). Derek Bentley was found guilty of murdering a police officer and was executed by hanging in 1953. About forty years after the events of the Derek Bentley case took place, Coulthard was asked to analyze Bentley’s confession statement with regards to its authenticity. Upon using a mixed-approach examination combining discourse analysis and corpus linguistics, Coulthard showed that

Bentley’s confession statement was likely to have been altered by the police and not just simply dictated by Bentley as officially declared by policemen involved in the case. Eventually, the analysis supported the allowance of an appeal against the conviction and the subsequent granting of posthumous pardon in 1998 (Coulthard, Johnson, and Wright 2017: 170; Coulthard 1992).

Another case is that of Jenny Nicholl, also solved by resorting to forensic authorship analysis. In Jenny Nicholl’s case a set of four text messages were submitted for forensic analysis for attribution of authorship to one of two candidate authors. One was the victim herself (the messages were sent from her mobile phone), and the other was her ex-boyfriend, suspected of murdering her. The analysis, based on comparison, showed that the stylistic choices of the author of the questioned messages were consistent with the ex-boyfriend’s writing style, and distinct from the stylistic preferences of the victim in pre-crime text messages. (Coulthard, Johnson, and Wright 2017: 158-60).

Forensic linguistics is conceived as an applied subfield of linguistics, “informed” by other “linguistic sub-disciplines” and addressing all levels of language, i.e., the phonetic, phonological, grammatical, lexical, syntactic, semantic, and pragmatic levels (Coulthard, Johnson, and Wright 2017: 14). Forensic linguistics comprises three main dimensions of research that aim to cover the broader range of settings in which forensic linguistics may play a relevant role. As explained by Johnson and Coulthard (2010: 7), Coulthard, Johnson, and Wright (2017: 14), and May, Sousa-Silva, and Coulthard (2021: 2) these are:

- The study of the written language of the law;
- The study of interaction in the legal process, which in criminal cases includes everything from an initial call to the emergency services to the sentencing of someone who has been found guilty; and
- The description of the work of the forensic linguist when acting as an expert witness.

However, the term forensic linguistics is also understood as what has been called the *lato sensu* of the term (Sousa-Silva and Abreu 2015: 111; Coulthard and Sousa-Silva 2016), i.e., the “scientific study of language […] used in Court of law or public discussion and debate” or that is of “public interest” (Turell 2013: 8).

Although authorship analysis is not exclusive to forensic linguistics, it is within the frameworks of forensic linguistics that authorship analysis has grown exponentially. The emergence of forensic linguistics as a field opened the path to expanding research in the intersection of the legal fields and areas like computational linguistics, applied linguistics, or stylistics. Thenceforth, the study of authorship gained greater relevance and breadth, producing works that can be circumscribed in the second and third dimensions of the field.

In the last thirty years, the advent of the digital era and the subsequent increase of access of populations to communicating in environments characterized by speed, ubiquity, and the possibility of anonymity have resulted in a rapid increase of research focusing on authorship analysis. Some of these works are concerned, for example, with cybercrime (Sousa-Silva Forthcoming 2021) online identities (Marko 2021; Amuchi et al. 2012), hate speech (Qian et al. 2018; Carney 2014), offensive language (Methven 2017), deceptive language (Fornaciari and Poesio 2011, 2012; Bond and Lee 2005), politics-related language (Clarke and Grieve 2019), the language of courtroom interaction (Eades 2008), and plagiarism (Sousa-Silva 2013). As a consequence of the publication of so many works, the field has grown steadily. These works do not always imply direct forensic usefulness since they are not carried out in response to specific forensic problem. However, they invariably carry knowledge with the potential to contribute to the resolution or prevention of future forensic situations, ethical problems such as academic dishonesty and plagiarism, or even social or mental health situations like aggressive advertising or suicidal speech in digital environments.

2.1.3. Current State of Development of Authorship Analysis

Since authorship studies have historically involved scholars mainly from the fields of humanities and researchers from the computer sciences, two main perspectives of authorship analysis have evolved, which can be considered either antagonist if applied in isolation, or compatible, if understood as interdependent (Wright 2014: 11-12; Grant 2008: 225). For practical reasons, these perspectives are expressed here as the linguistic and the quantitative paradigms.

Within the linguistic paradigm, authorship analysis can be subdivided into two major types, i.e., authorship problems that cannot be approached by comparison with other texts because they simply do not exist or are unavailable, and authorship problems that can be addressed by comparison with other texts. This division has also been described as Sociolinguistic Profiling and Authorship Attribution (Coulthard and Sousa-Silva 2016). In the second group, three types of problems have been defined, resulting, altogether, in four typologies as described below according to Grant (2008) for item 1) and McMenamin (2002: 93) for items from 2 to 4:

1. Single-text problems**:** addressed by the question "what kind of person wrote the text" and that refer to a question of authorship "where there is no realistic possibility" of comparing the questioned texts to any other text. This typology comprises sociolinguistic profiling and psycholinguistic profiling of authorship. (pp 222-24)
2. Determination of authorship consistency: in cases where "particular writing, which may or may not be already accepted as part of a body (canon) of known writings, is consistent with the rest of the known writings." (p. 93)
3. Authorship Comparison: comparing "a questioned writing with the writings of a large number of possible authors if there are no obvious suspect authors." (p. 93)
4. Assessment of Authorship Resemblance: usually presented after "possible suspect authors can be identified by external (non-linguistic) means." It applies to cases where the linguist has to look for similarities between the "questioned writing" and the writing of another "author or a small number of candidate authors." (p. 93)

Within the quantitative paradigm, authorship analysis has been described as consisting of three major fields. According to Reddy, Vardhan, and Reddy (2016) these fields are:

1. Authorship Profiling: defined as "the task of determining demographic features of authors like native language, education, gender, age and personality traits of an author by understanding their writing styles." (p. 3092)
2. Authorship Identification is presented by these authors as being subdivided into:
    a. authorship attribution, which "determines the author of a given anonymous text from known writings of many authors" and

    b. authorship verification, which “finds whether the given texts were written by the particular author or not by considering the writings of a same author.” (p. 3092)

3. Plagiarism Detection, which is dedicated to detecting “whether a given document is original or not. This approach is broadly categorized as text alignment and source retrieval. Text alignment is a process of matching the contents in terms of passages between two documents. Source retrieval is a process of searching for the similar sources of a suspicious document.” (p. 3092)

The linguistic paradigm, authorship analysis refers to types of problems, whereas, in the quantitative paradigm, authorship analysis is presented as different types of tasks. However, despite the differences in designations and approaches, it can be said that there is a correspondence between the major types of problems/fields within each paradigm. Table 1 shows a representation of such correspondence. In the linguistic paradigm, the categories within multiple-text problems (Grant 2008: 224) or comparative authorship analysis which is “the task of comparing texts of known authorship with one or more anonymous texts with a view to potential attribution” (Grant and MacLeod 2018b: 82) are intentionally not divided to represent the breadth of applicability of these analyses, which can be used to address many different multiple-text types of problems.

<table>
<tr><th colspan="2">QUANTITATIVE PARADIGM</th><th>LINGUISTIC PARADIGM</th></tr>
<tr><td colspan="2">Authorship Profiling</td><td>Single-text problems
(Sociolinguistic/psycholinguistic profiling of authorship)</td></tr>
<tr><td rowspan="2">Authorship Identification</td><td>Attribution</td><td rowspan="3">Multiple-text Problems or Comparative Authorship Analysis

Authorship comparison
Determination of authorship consistency
Assessment of authorship resemblance</td></tr>
<tr><td>Verification</td></tr>
<tr><td colspan="2">Plagiarism Detection</td></tr>
</table>

Table 1 – Correspondence of authorship analysis fields/types of problems according to the general paradigm adopted

The computational and the linguistic paradigms work well as complementary dimensions to authorship analysis. The techniques and types of analyses used in each approach are distinct. The computational approach analyzes data using computers and relies, for example, on statistical tests, mathematical formulas, and algorithms tested for their replicability, validity, and reliability (Solan 2013: 574; Wright 2014: 20). Despite works like those by Argamon et al. (2009) and Argamon and Koppel (2012) employing Systemic Functional Linguistics (SFL) as the theoretical language foundation to address authorship analysis problems, research developed within the quantitative paradigm usually does not refer to linguistic theories that can explain why certain features perform better than other in identifying authorship in a given text type, register, or genre (Wright 2014: 21).

The linguistic paradigm, on the other hand, is usually regarded as one that would entail the analysis of an authorship problem relying on the linguist, and conducted on a case-by-case logic, which makes it difficult to replicate (Nini and Grant 2013: 2; Wright 2014: 20); but, contrary to the computational approach, this perspective is based on linguistic theories concerning language variation that can describe and explain differences between authors, and examines authorship in all its depth, that is, not only morpho-syntactically, as approached by the quantitative perspective, but also semantically and pragmatically (Wright 2014).

Some of the works combining quantitative and qualitative methods that approach authorship issues comprehensively and provide quantitative support to linguistic evidence are those by Grant (2013) and Johnson and Wright (2014). The first proposes an analytical framework based on “vocabulary choices and morphological features” (p. 472) to address short texts resorting to a statistical approach known as Jaccard’s coefficient, a correlation for binary values that can indicate similarity between two groups. The second presents a corpus linguistics approach using the Enron emails corpus to describe a methodology of combined systematic approaches to address authorship analysis and authorship profiling. Ultimately, linguistic and quantitative paradigms are two avenues reaching the same final point, the attainment of results that can serve as evidence for authorship attribution, joining quantitative data, and theoretical insight into the data.

2.1.4. Authorship Profiling

Authorship profiling is a subfield of authorship analysis. Recent definitions of authorship profiling describe it as the most initial level of analysis at which a linguist may be confronted because no other texts are available for comparison, and therefore, only the author(s)' characterization(s) is possible (Queralt 2014: 37); as the type of analysis usually requested by the police when the clues on the identity of the author(s) of whatever crime are weak (Coulthard and Sousa-Silva 2016); as "the task of determining the characteristics of an anonymous author, such as their demographic details, from the way they use language" (Nini 2018: 39); and as a subfield that "distinguishes between classes of authors by studying how language is shared by people" (Bevendorff et al. 2020: 509).

The definition by Queralt (2014) and (Coulthard and Sousa-Silva 2016) recalls the first level of authorship problem types described above, i.e., single-text problems (Grant 2008: 222). In these terms, authorship profiling aims to characterize authorship based on language-related aspects relevant to the analysis. Any language-related aspect of authorship as understood in section 2.1.1, i.e., as encompassing the author(s) of the text, the context(s) surrounding the production of the text(s), and the text(s) itself, contributes to characterizing authorship. This means that sometimes authorship analysis does not focus on who produced the text but rather on the context of the text production and the text itself (Grant 2008). For example, in the forensic context, a written confession may be questioned with regards to its "mode of production" (Grant 2008: 221). That was partly the situation in Derek Bentley's case described previously (Coulthard, Johnson, and Wright 2017: 163) in which the differences between what was in the text of the confession and the circumstances of the confession production, demonstrate the participation of "multiple authors" and not of only one author.

However, authorship profiling refers, more particularly, to what is described in the second and third definitions, focusing on linguistic output to describe the writer of the text in terms of "socio-collective traits" (Turell and Gavaldà 2013: 498). These traits refer to sociodemographic variables, variables that connect languages and users, are connected among each other, and influence one another (Coulthard, Johnson, and Wright 2017: 14). Gender, age, ethnicity, geographical location, education level, language background, and profession are some of the sociodemographic variables that can be studied to understand an author’s profile (Grant 2008: 222-23).

Authorship Profiling has a brief history of roughly two decades. It is deeply rooted in Sociolinguistics as the field interested in studying “the social uses of language” (Chambers 2013: 1) and specifically in Variationist Sociolinguistics, a subfield of Sociolinguistics and Linguistics (Chambers 2013: 2).

In its origin, Sociolinguistics was very much influenced by research in dialect geography and historical linguistics dating from the late nineteenth and early twentieth centuries that reported on diachronic language changes and phonetic variation concerning social aspects in French, German and American English varieties (Koerner 1991: 59-60). Other works related, for example, to bilingualism in Switzerland and India (Weinreich 1951, 1957) or to the examination of bilingualism as a field of research (Weinreich 1953) also contributed to shaping Sociolinguistics (Koerner 1991: 61).

Although the term “sociolinguistics” had been used many years earlier by the researchers Hodson (1939) in India and Currie (1952) in the United States of America, the formal inception of the field was in 1962 at “the 37th Annual Meeting of the Linguistic Society of America in New York City on December 29, 1962” (Chambers 2013: 2; Labov 1963: 273; Mohan 2004: 261) where Labov presented “an abbreviated version” of his publication “The social motivation of a sound change” (Labov 1963: 273).

Labov (1963) presented the results of empirical research that was considered to be pioneering for demonstrating that linguistic change could be observed over a short period and not only in diachronic studies as previously defended. It could also be used for correlating linguistic variants of the speech community of the island of Martha's Vineyard (Massachusetts, United States of America) to social factors like age, social class, sex, occupation, geographical distribution within the island, and ethnic origin of the informants, as well as for operationalizing style (e.g., articulatory style) as an independent variable (Chambers 2013: 2; Hazen 2007: 73).

However, Labov (1966 [2006]) and Weinreich, Labov, and Herzog (1968) were the works that initiated a tradition in studies focusing on the description of social factors and their correspondence with linguistic patterns (Chambers 2013: 2; Wright 2014: 34). They presented the rationale of “linguistic heterogeneity” as the object of and orderly study that would describe speech/performance as the realization of *langue*/competence, an idea that questioned the conventional belief that only the “homogenous” and “abstract” system of *langue*/competence could be the true object of study in linguistics (Chambers 2013: 7; Hazen 2007: 74).

Following the methodological approaches of Labov (1966 [2006]) and Weinreich, Labov, and Herzog (1968), other works describing the stratification of language were published in sociolinguistics circles that examined, for example, multilingualism and variations within Euskera – the language of the Basque Country in Spain– in the city of Bilbao, capital of the Basque Country (Arostegui and Etxebarria 1985); urban speech variation concerning a specific phonetic alternation called *yeísmo* (Martín 1983) with regards to social factors like occupation, social class or ethnicity; and the phonetic alternation according to the formality in the urban speech of the city of Belo Horizonte, Brazil (Veado 1983).

Sociolinguistics has developed greatly in the last fifty years and is currently divided into Variationist Sociolinguistics and Interactional Sociolinguistics, although both terms overlap with Sociolinguistics without distinction. According to the words on the home page of the scientific journal *Language Variation and Change*, Variationist Sociolinguistics is:

> “The study of linguistic variation and the capacity to deal with systematic and inherent variation in synchronic and diachronic linguistics. Sociolinguistics involves analyzing the interaction of language, culture, and society; the more specific study of variation is concerned with the impact of this interaction on the structures and processes of traditional linguistics. Language Variation and Change concentrates on the details of linguistic structure in actual speech production and processing (or writing), including contemporary or historical sources.”
>
> https://www.cambridge.org/core/journals/language-variation-and-change#

Like any authorship analyst, researchers working in authorship profiling have drawn on the assumption of Variationist Sociolinguistics that language variation is systematic and observable, leading to the development of research seeking to determine the linguistic variables that better predict certain sociodemographic variables.

Together with social class and geographical region, gender and age are among the most studied variables in Sociolinguistics (Labov 1990; Raidt 1993; Eckert 1989; Eckert and McConnell-Ginet 2003). Previous knowledge from Sociolinguistics on gender and age has motivated that these two variables are also the most studied factors in authorship profiling, given their usefulness to discriminate between users, especially anonymous or disguised users of the digital media, for forensics, marketing, or security purposes (Ferro and Peters 2019: 474). Specifically, gender has been interpreted and discussed in the computer sciences as the biological sex and a so-called simpler factor to address, given its binary nature and the fact that it is easier to collect texts authored by a person of a given sex (Savoy 2020: 11).

As a result of these features' informative power, several authors from computational linguistics have contributed to empirical research in authorship profiling. For example, Argamon et al. (2003) examined a large corpus of formal written text from the British National Corpus (BNC) to determine differences between male and female authors when writing in English. They found that pronouns and certain types of noun modifiers are more "prominent" in "female-authored documents" than in texts penned by men and can predict gender in these types of documents with an accuracy of about 80%.

In an annotated corpus of text from Dutch Twitter users, Nguyen et al. (2013) found that 'tweet length' increases with age but does not seem to change between users who are male and those who are female, while self-reference, i.e., use of the pronoun 'I,' is more frequent in younger users and in females more than in older people and males. Peersman, Daelemans, and Van Vaerenbergh (2011) studied a corpus of Flemish Dutch posts from the Belgian social networking site Netlog and found that unigrams such as 'bro' (brother) or "grts" (greetings) are useful for classifying users by age as they correlate with younger users with an accuracy of 71.3%.

Goswami, Sarkar, and Rustagi (2009) used corpora of texts from blogs to study slang words and the average length of sentences as features to classify texts by age groups and gender. They did not find significant differences in the length of sentences produced by males and females or according to the age groups (10s, 20s, 30s, or higher). However, they found that the usage of slang words predicts gender with 77.39 % and age with 89.68 % accuracy.

Schler et al. (2006) also used corpora of texts from blogs to study gender and age. They found that their selected style-related features (parts-of-speech, function words, and blog-specific features) performed better than content-related features (theme-related words classified under categories such as ‘money,’ ‘family,’ friends, ‘sports’) in predicting gender. However, the combination of both (502 features) correctly classified authors’ gender of unidentified texts in 80.1% of the cases. As to age, content-related features performed better than style-related features, but again, the combination of both identified authors in their 10s from authors in their 20s and authors in their 30s with 87.3% and 96% accuracy, respectively; whereas the distinction of authors in their 20s from authors in their 30s was considered less successful at 76.2%.

However, many of these empirical studies have been criticized for addressing gender and age from the biological perspective only, disregarding their sociological dimension (Nini 2014: 19; 40). The omission of the social dimension of gender and age contributes to overlooking linguistic instances reflecting adaptation to communicative situations rather than biological sex, i.e., the modification of language resulting, for example, from communicative interaction with another user (Nini 2014; Bamman, Eisenstein, and Schnoebelen 2012).

Another factor of authorship profiling that has also been addressed is personality, which given its nature, has been closely related to research in the field of Social Psychology. It aims at understanding, for example, the relationship between function words and social behaviors (Pennebaker 2013), or linguistic markers of personality disorders such as narcissism, known to be linked to suicidal tendencies (Holtzman et al. 2019; Ansell et al. 2015).

In this regard, knowing that people’s emotional state may reflect in their choice of words (Pennebaker, Mehl, and Niederhoffer 2003), researchers, mainly from the field of psycholinguistics working on authorship profiling, seek to determine personality traits based on text. For example, Litvinova et al. (2016) examine models based on parameters that can be quantified such as readability indexes (i.e., Flesch readability index, Hanning Index or index of complex words, average sentence length in words), lexical diversity, and frequencies of part-of-speech to predicting behaviors implying self-harm; and Liu, Perez, and Nowson (2016:

6) successfully determine Twitter users' personality traits like "extroversion, emotional stability […], agreeableness, conscientiousness and openness", using deep learning approaches based on vectorial representations of different parts of the text, i.e., characters, words, and sentences .

However, other factors, such as the linguistic background of authors, have been less explored in the context of authorship profiling. This could be due to a possible lower demand for solutions to actual forensic problems or commercial or security challenges related to the language(s) used by people to communicate. Also, the complexities associated with the morphology, syntax, semantics, and pragmatics of the many languages currently used in the cyber world may play a role in the number of studies of the analysis of the authors' linguistic background. However, the linguistic background of authors is of utmost importance if we consider that, as social beings, language is our main and most primary form of conveying ideas whether in oral or written form, face-to-face or at a distance. Therefore, language is one of the most basic expression of who we are. The most long-standing term for this authorship profiling type of analysis is native language identification (NLI). However, as shown later in this section, other terms such as native language influence detection (NLID) or other language influence detection (OLID) are also used.

2.1.5. Native Language Identification (NLI)

Language background is one of the sociodemographic variables addressed within authorship profiling. The profiling of authors' language background has more frequently been defined within the literature of computer sciences than in works situated within linguistics. In computer sciences, LPB has been described as a task and is usually called *native language identification - NLI*. A usual working definition of NLI found in research articles from the field of computational linguistics is that of a "task of identifying the native language (L1) [also mother tongue or first language] of a writer based solely on a sample of their writing in another language." (Tetreault, Blanchard, and Cahill 2013: 48).

The concept, however, is grounded within the domain of second language acquisition (SLA), namely within "language transfer", "cross-linguistic influence" or "cross-linguistic effects", that all refer to the study of the "direct and indirect consequences" of a speaker's native language (L1) on the use of a language that he/she learned later in life (L2) (Jarvis 2012b: 1).

According to Jarvis and Crossley (2012: 19-21) the fields of computational linguistics and SLA converged in NLI about the same time when researchers from artificial intelligence were beginning to work on automated text classification with the aim of profiling authorship. The authors refer to several early works addressing the classification of texts according to genre (Stamatatos, Fakotakis, and Kokkinakis 2000; Santini 2004) or focusing on the presentation of novel and better-performing techniques of machine learning (Sebastiani 2002; Alpaydin 2004). However, none of these works presented a theoretical framework that could explain the success of their quantitative findings.

Within SLA, Jarvis and Paquot (2015: 605) define NLI as “the task of automatically identifying the first language (L1) of a language user based on the person’s production of the target language”. This definition assumes the hypothesis that it is possible to detect the influence of the L1 of a person on his/her L2 production by analyzing the use of language patterns common to the speakers of the person’s L1 (Malmasi and Dras 2017). Likewise, NLI works on three premises, as explained by Kyle, Crossley, and Kim (2015: 188), i.e. 1) having corpora for comparison (L1 vs. L2); 2) using a set of previously determined linguistic features relevant to such a comparison, and 3) working with a statistical or computational approach. It follows from these assumptions and premises that NLI works with written text, even if the text is a transcription from oral interactions, as can happen in cases of authorship analysis in forensic contexts (Grant 2008: 216).

As will be shown later in this section, language background profiling can have different applications. An important application of this authorship analysis task is within the forensic field. For this reason, the next section examines the differences and similarities of NLI and the related language analysis field within forensic linguistics known as LADO - Language Analysis for the Determination of Origin.

2.1.6. NLI and LADO

Within the forensic context and as noted previously by Perkins (2014: 44-46), NLI is related to LADO –, a field dedicated to the analysis of spoken language for purposes of granting asylum based on a claim of origin or ethnicity, when the claim cannot be sustained with identification documents like passports (Patrick 2019: 1-2). According to the Convention Relating to the Status of Refugees (UNHCR 1951), persons fleeing from conflict or war zones in different parts of the world have the human right to seek asylum to protect themselves from persecution "for reasons of race, religion, nationality, membership of a particular social group or political opinion"(p.3) , provided that they are "unable" or "unwilling" (because of fear) "to avail themselves of the protection of their country"(p.14). However, to claim such a right, they have to demonstrate they were born and raised in their country of origin or belong to a specific ethnicity.

NLI works as a supplementary mechanism for cases of authorship profiling, for example, in a forensic context (Perkins 2014; Perkins and Grant 2018). Likewise, LADO is a type of profiling analysis of a speaker (Foulkes, French, and Wilson 2019: 92-93) used to supplement the task of demonstrating the origin of the asylum seekers by determining if they are native speakers of the language of the country in conflict or from a given ethnic group, which is often done using interviews "to test their speech [in the language] they claim as mother tongue" (Patrick 2019: 2).

These two areas converge in three main points. One is the ultimate objective of determining an individual's native language based on that person's linguistic output. Both LADO and NLI operate based on the relation between the individual's native language (L1) and other languages the person learns later in life (L2/L3/Ln). For this reason, LADO and NLI seek to establish whether an individual is using L1 discourse. The second similarity refers to the type of task they addressed, which is typically one of verification, in cases of a claim referring specifically to the speaker's language background; or classification, when there are "no specific claim, but instead an open question of what information can be gleaned about the speaker" (Foulkes, French, and Wilson 2019: 95). Lastly, the third similarity refers to the fact that both LADO and NLI are considered supplementary analyses, i.e., analyses that add to the core evidence.

Despite these similarities, LADO and NLI do differ significantly. LADO and NLI differ in the degree of supplementarity they represent. In LADO cases, the supplementary nature of the analysis is fundamental to the case, while in NLI, the results obtained are truly complementary to other pieces of evidence. That is, in cases of asylum seekers, LADO does not play an actual accessorial function. It has more of a fundamental role in asylum granting or denying since the lack of identification documents of the asylum seekers does not leave space for other types of analyses (Eades et al. 2003: 45). NLI, on the other hand, usually works as an auxiliary of forensic investigations involving written language (Perkins 2014: 45).

Another important difference already mentioned concerns the type of discourse they address, which is written in NLI and oral for LADO cases. This second difference has implications in the techniques used to analyze the linguistic data. LADO resorts to techniques within the fields of phonology and phonetics, dialectology, or sociolinguistics (Foulkes, French, and Wilson 2019: 93), while NLI employs techniques from the fields of grammar, lexicology, or computational linguistics (Kredens, Perkins, and Grant 2020; Perkins 2014, 2015; Perkins and Grant 2018).

NLI and LADO also differ in the type of data addressed in these types of analyses. While NLI works with corpora of texts that were not produced with the specific intention of future analysis, LADO uses oral discourse that is narrated or induced through an interview carried out with the primary purpose of generating data that is later analyzed to determine the speaker's origin. The written linguistic material used in NLI occurs independently of any analysis for which such a material may be used later; however, the linguistic material used in LADO is elicited or induced purposely (Perkins 2014: 45) and can be influenced by factors like the following:

> "the power differential between people seeking asylum and those involved in judging their claims, dislocation of the speaker in time and space, multilingualism and linguistic accommodation, cross-cultural misunderstanding, the ability of narration and other modes of speech to yield appropriate data, varying levels of understanding of interview goals, the scope for linguistic imitation (non-authentic speech), and test-awareness, among others" (Patrick 2019: 6).

Therefore, linguistic material used in NLI tends to be less context-dependent (Grant 2008: 216) than linguistic material used in LADO.

Finally, there is a difference in the specialists that carry out the work in these two fields. While linguists, and specifically forensic linguists, participate in both, in NLI it is rather specialists concerned with written text that take part in the analyses, i.e., grammarians, lexicologists, computational linguists, artificial intelligence specialists, psycholinguists or translators; while in LADO, it is usually specialists from the fields of phonology and phonetics, NENS (non-expert native speaker) or Government agents that participate in the analyses. Table *2* below shows a summary of these similarities and differences.

	NLI	**LADO**
Similarities		
Objective of the analysis	Native Language Determination	
Type of task	Verification or classification	
Differences		
Type of analysis	Complementary	Supplementary
Type of discourse addressed	Written	Oral
What linguistic data is addressed	Syntax, vocabulary, phrases	Sounds (phonemes, accents, place of articulation)
Type of data used	Collected or deduced. Uses corpora built with a discourse that was not produced for the analysis	Elicited, invoked, narrated, or induced. Uses discourse produced during an interview for analysis
Who performs the analysis	Forensic Linguists; Grammarians, lexicologists; Computational linguists; Artificial Intelligence Specialists; Psycholinguists	Forensic Linguists, Phonologists, Phoneticians, NENS (non-expert native speaker), Government Agents

Table 2 – Similarities and differences between NLI and LADO

It can be said that both NLI and LADO are profiling tasks that rely on language –written and oral, respectively– to determine identity-related features or sociolinguistic characteristics. However, while LADO is a forensic type of task with currently well-established procedures in most countries (Eades et al. 2003), NLI may or may not be used in forensic cases. This difference grants NLI a greater breadth of applicability as has been shown, for example, by Tomokiyo and Jones (2001), who proposed to use textual features instead of acoustic ones of a speech sequence to identify the native language of a speaker.

2.1.7. NLI, NLID, and OLID

Works approaching NLI from the linguistic perspective, and particularly with a forensic focus, have also used the terms native language influence detection – NLID (Perkins and Grant 2018) and other language influence detection – OLID (Kredens, Perkins, and Grant 2020) to convey conceptual aspects concerning this type of authorship profiling.

The use of the expression "influence detection" instead of "identification" proposes that the authorship profiling problem concerning the effects of a native language (L1) on a second language (L2) based on written textual production cannot always unequivocally "identify" the native language of the author of the production, but rather indicate the likelihood that certain features in the realization of the L2 are affected by an L1. This expression also speaks of the possibility of any given individual having more than one L1 influencing the output in a non-L1 language.

The expression "other language" also assumes these proposals and extends the concept in terms of the language influencing another language and the direction of such an influence. Based on what was previously discussed by Pavlenko (2000), the notion of 'other language' proposes to broaden the L1 concept from "nativeness" to dominance; and the notion of the directionality of the influence from unidirectional to bidirectional (Pavlenko and Jarvis 2002). In other words, not only can a dominant language or languages - be it native or not- influence a non-dominant language, but also a non-dominant language may influence a dominant language, or for that matter, a native language. In other words, the realization of a non-dominant language may be affected by one or more dominant language(s), and vice versa (Perkins and Grant 2018: 2; Kredens, Perkins, and Grant 2020: 11).

In this work, the perspective adopted on the profiling of authorship concerning language background corresponds with the forensic linguistics approach. In other words, rather than identifying authorship, the analysis seeks to indicate the likelihood that linguistic influence originates in a given native language. Thus, the perspective of a non-native language being influenced by a native language is adopted. Accordingly, the term native language influence detection (NLID) is preferred over other options.

2.1.8. Applications of NLID

The profiling of an individual's language background based on his/her linguistic written output has at least three major types of applications: educational, marketing and business intelligence, and forensic.

In the educational aspect, the profiling of the language background (NLI/NLID/OLID) can be of great assistance to foreign language professors in identifying phonological, morpho-syntactic, grammatical, lexical, and discursive difficulties that learners may experience as a result of the influence of their native language or a dominant language even if not native. This identification can be instrumental in elaborating teaching material adapted to the students' specific learning needs or developing computer-aided language learning (CALL) software (Tetreault, Blanchard, and Cahill 2013: 48; Malmasi and Dras 2018: 404). As mentioned earlier, second language acquisition (SLA) has been the natural ground where the identification of native language has developed. For this reason, it is also in this field where the first applications have been registered, the first methods and techniques have been applied, and the first and most relevant theoretical proposals have taken place. Although applications in this field concern especially English taught as a foreign language, other languages are currently being registered and described from this perspective. For example, Gayo, Zampieri, and Malmasi (2018) presented “the first Portuguese dataset compiled for Native Language Identification (NLI)” (p.295) containing essays authored by learners of Portuguese from fifteen nationalities. The dataset was shown to help identify the influence of native language at the lexical level of Chinese, English, German, Italian and Spanish students when writing in Portuguese.

In the era of the internet, an important part of human life takes place online, and the degree of customization of almost anything users experience on the internet is increasingly higher. This form of digital way-of-being includes, among many others, the commercialization of goods and services or the implementation of e-businesses or platform business for which information about actual and potential customers is essential to improve products or develop new ones. The utility of NLI in this area is usually aggregated with other profiled features of customers very much related to sentiment analysis, content analysis and personality classification (Oberlander and Nowson 2006). Sentiment analysis or opinion mining is defined as “the task of finding the opinions of authors about specific entities” (Feldman 2013: 82). Content analysis is understood as “the systematic study of 'manifest content' in all forms of communication [...] in news media, speeches, advertisements, and campaigns, [...] social media and blogs, [using] text analysis, the systematic study of written text or transcribed speech, as well as techniques that focus on nontextual message content, including pictorial images, graphical elements, moving images, nonverbal behaviors, music, and sounds." (Neuendorf and Kumar 2015: 1). Lastly, personality classification can be described as the task of labelling individuals as, for example, extraverted, agreeable, conscientious, neurotic, or open based on “any form of observable behavior that can be perceived by others” (Vinciarelli and Mohammadi 2014: 276). The aggregation of NLI and these other areas happen because language background alone does not provide enough information for purposes, for example, of marketing (Oberlander and Nowson 2006; Glance et al. 2005).

In relation to the forensic applications, NLI can help “to glean information about the discriminant L1 cues in an anonymous text”, thus contributing to cybersecurity and online safety (Malmasi and Dras 2018: 404). For example, Wong and Dras (2011) studied a set of syntactic features to verify if their use in automatic classification tasks could improve the results of authorship profiling of phishing emails. The authors obtained the syntactic trees of a training set of 490 statistically parsed essays extracted from the International Corpus of Learner English (ICLE) for seven languages. The tree cross-sections were used to characterize non-native speaker errors and then to classify another 175 essays written in the same 7 languages also from the same corpus. Their study showed that the approach can improve the results obtained by other authors in native language identification by reducing by 30% “the error in the cross-validation evaluation with significance testing” (Wong and Dras 2011: 1608-09).

However, texts do not need to be anonymous for NLID/OLID to be instrumental in resolving an authorship problem. Translingual plagiarism, as examined by Sousa-Silva (2013, 2014, 2019) is this type of problem. As defined by the author, translingual plagiarism is “another case of plagiarism of ideas […] where the plagiarists lift the text from one language, have it translated into another language, and subsequently reuse it as their own” (Sousa-Silva 2014: 72). In this context, translingual plagiarism is deemed a linguistic type of problem (Sousa-Silva 2014: 74) in which, as suggested by the term, at least two language systems are present. Therefore, translingual plagiarism has been approached from the perspective of the relationship between languages or, more accurately, the effects of language A on language B (Sousa-Silva 2014: 79-81). Since authorship problems in the field of NLID/OLID also imply the presence of a minimum of two languages and the understanding of how one affects the other, progress in this field may apply to and advance research in translingual plagiarism detection by providing insights on linguistic features that denounce the influence of another language, and vice versa.

Investigations on NLI have focused mostly on English produced by non-native users in informal contexts such as users’ interaction in online or learning environments. Comparative analysis for NLID has resorted to corpora of texts produced in this type of context. For example, Perkins (2014) has studied NLID in two corpora of texts from online blogs produced in English by native and non-native bloggers who are L1 Persian speakers; and authors like Argamon et al. (2009) have examined the writing of non-L1 English authors from Russia, the Czech Republic, Bulgaria, France, and Spain, using a corpus of texts produced by users at a learner level of language acquisition/instruction. Only a few authors have used corpora of texts produced by “highly-advanced” non-native users (Goldin, Rabinovich, and Wintner 2018: 3591) and so-called non-learner writers (Kredens, Perkins, and Grant 2020). This research seeks to add to works addressing NLID of advanced English users, specifically in scientific writing produced by Portuguese and Spanish.

2.2. Theoretical and Conceptual Frameworks for NLID

The study of authorship from the linguistic viewpoint has consolidated methodological approaches and conceptual frameworks that can consistently account for an author or group of authors' idiosyncrasies. As mentioned above, authorship profiling, and therefore NLID, have in variational sociolinguistics and second language acquisition their main theoretical basis. Since language is not homogenous among the speakers (Biber 1995: 1; Labov 1972), the one aspect that can connect all of them is the relative "lack of homogeneity" they show (Marquilhas 2013: 17). Therefore, NLID conceptual frameworks draw on the variation of language and on bilingualism/multilingualism. Given its relevance for authorship studies, authorship profiling, and specifically for language background profiling and NLID, the next sections discuss language variation at the individual and group levels; and SLA concepts concerning bilingualism/multilingualism and theories on cross-linguistic influence.

2.2.1. Language Variation: Idiolect

Language is definitely a social event. Variation concerns not only differences but also the similarities individuals display to be able to communicate. Although the description of variation has focused on linguistic communities as the space where changes occur, it is the individual that acts as the most elemental agent of language use. Halliday, McIntosh, and Strevens (1964: 156) affirm that:

> "it is the individual who speaks and writes; and in his language activity dialect and register combine. In the dialect range, the finer the distinctions that are recognized, the smaller, in terms of number of speakers, the unit which we postulate as the dialect community becomes. Eventually we reach the individual. The individual is, so to speak, the smallest dialect unit: each speaker has his own IDIOLECT."

From this perspective, the individual occupies a central and primary role in language realization since it is the individual who acts as the most elemental source of linguistic output, and, thus, of linguistic variation. Thus, studying the particular way individuals use language, i.e., idiolect, becomes paramount.

The assumption underlying authorship studies is the existence of idiolect (Coulthard 2004: 431), a term defined at the time as “the totality of the possible utterances of one speaker at one time in using a language to interact with one other speaker” (Bloch 1948: 7). The first ideas on individual language are credited to linguist Hermann Paul who said:

> “every linguistic creation is always the work of one single individual only. Several no doubt may create similar products, but neither the act of creation nor the product is affected by that.” (1890: xliii).

The distinction of a linguistic individual is also present in the work of Sapir (1927) when he affirms that “we all have our individual styles […], and they are never the arbitrary and casual things we think them to be” (p.903); and of Bloomfield (1933) when referring to the capacity of individuals to express “great differences even among the native members of a […] relatively uniform group” (p.45). Similar notions were also proposed by Benedetto Croce in 1921, Otto Jespersen in 1935 and K. Rogger in 1941 when referring to *lingua individuale*, individual language habits, and *Individualsprache*, respectively (in Coseriu 1978: 63-64).

The meaning of the term comes from the Greek expressions *ἴδιος* /idios/, meaning "one's own, personal, private", and *λεκτος* /lektos/, a derivation of the verb *λέγω* /légɔ/*,* meaning *“*to say, to tell, to speak, to recite, to say something that is written, to narrate in a manner that implies care or choice”, and probably appeared in the Greek language associated with the word *διάλεκτος* /dialektos/, meaning “dialect” or way of expressing themselves of a specific group of persons (Dicciogriego n.d.-a, n.d.-b; DGE n.d.).

Idiolect has been a matter of debate. As far as the essence of the connection between language and thought is concerned, idiolect has been defined as either a deviation from a standard language that is common to individuals, or as the way that language is used by individuals in a given context. In the first case, language comes first, and idiolect can be identified by contrast to language; in the second case, the totality of the usage made by individuals making up language, in which case the individual linguistic creation/interpretation precedes language (Penco 2007: 2; Johnstone 1996: 12).

Barthes (1986 [1964]: 21) affirmed that “since language is always socialized, even at the individual level” there is no such thing as an individual’s language, which is why he presents idiolect as mostly an “illusion”. He, however, mentioned three “realities” for which the concept may be useful. The first is for cases of aphasia, i.e., language impairment resulting from some form of brain damage which will prevent the individual from communicating and understanding others clearly. The second is when there is a need to name the particular form of writing of a person, i.e., the writer’s style, although writers will always be influenced by “patterns coming from […] the community” (p.21), he affirmed. Finally, he admitted the concept of idiolect may be useful to refer to the language of a linguistic community who “all interpret in the same way all linguistic statements” (p.21). Still, Barthes calls for the need of an intermediate concept between Saussure’s *langue* and *parole* that is not idiolect but rather *parole* that is structured but not formal or official.

Jakobson (1971) also positioned himself against the notion of idiolect as proposed by within synchronic dialectology by Hockett (1958) who defines idiolect as “the totality of speech habits of a single person at a given time” (p.321). Jakobson does not recognize private property in language, and his rejection of the concept of idiolect is justified by the fact that speaking habits proposed in Hockett’s definition do not include the individual’s “habits of understanding the speech of others” (Jakobson 1971: 559) and since communication is bidirectional, talking about idiolect is, in his words, “fiction”.

Petrenko (2006: xi), among others, argued “against an idiolect conception of language” following Kripke’s (1982) interpretation of philosopher Ludwig Wittgenstein’s proposal against “private language” defined as “what can only be known to the person speaking; to his immediate private sensations” (Wittgenstein 1986 [1953]: 89e). These authors also view language as preceding idiolect and idiolect as a rather fabricated concept since understanding a “private language” requires the mediation of the community of speakers, and this mediation prevents individuality.

In the context of phonological variation and change, Labov (1989: 1) poses the question of “where to find the most systematic view of the linguistic system—in the individual who carries the genetic mechanism, or in the community that exerts the stimulus and control”, and uses the description of the short “a” of the Philadelphia dialect to support his position in favor of the priority of language over idiolect. He affirms, “language is not a property of the individual, but of the community. Any description of a language must take the speech community as its object if it is to do justice to the elegance and regularity of linguistic structure" (Labov 1989: 52).

However, the problem of priority between language and idiolect is apparent in this case. Idiolect has been considered a form of diaphasic variation of language, i.e., a variation of language according to stylistic factors, including individual variation (Coseriu 1982: 19-20; Ferreira et al. 1996: 481). The concept of idiolect has also been defined as the dialect of one person in a specific time of his/her life (Crystal and Ivić 2014: 63; McMenamin 2002). In this perspective, an idiolect is described as a restriction or a previous stage of a dialect to explain the priority issue by proposing a more integrated view of the concept. The concept of idiolect adopted in this perspective considers synergy of an idiolect with language and dialect and implicitly with other variations of language, as described by McMenamin (2002):

> Language can only be observed in individuals whose idiolectal features are very important for applications related to authorship identification. However, such individual characteristics become unimportant for the description of the speaker's dialect or language (the usual goal of linguistic analysis), wherein the focus is on group characteristics shared by all speakers or writers of the speech community. Dialects are not simply large collections of individual idiolects but are a synthesis of shared elements. Since language variation and change within a dialect or language are group phenomena, the idiolect is less the source of variation and more its reflection in the individual. When language changes over time, there are periods when "competing" new and old forms exist side by side in the whole speech community. Multiple forms will also be found in the language of an individual speaker, i.e., in his or her idiolect. Such individual variation is due to changes going on in the speech community, as well as to changes occurring in the person's own process of language acquisition and use. (p.67)

As can be seen from this description, idiolect does have a complex abstract nature, and although it concerns the individual, idiolect is not truly completely private since it can be interpreted by other speakers, and this interpretation occurs because meaning is negotiated among individuals. Still, the actual 'maker' of language is the individual since it is the individual who "uses language so as to locate [himself] in a multi-dimensional social space" (Hudson

1996: 29) and is capable of expressing himself even in the most constrained registers such as scientific discourse (Johnstone 1996: 89).

Coulthard (2004: 431-32) defines idiolect as "every native [speaker's] own distinct and individual version of the language they speak and write [which] will manifest itself through distinctive and idiosyncratic choices". This premise establishes that it is possible to identify the author of a text –or traces of him/her– based on the analysis of the language he/she uses in his/her writing. What supports this argument is the assumption that the use of language is distinctive of the individual. Such characteristic use of language by an individual is automatic and systemic as it is "usually unconscious" (McMenamin 2010: 488) and it affects all levels of language, i.e., from the smallest units, like speech sounds or letters to the larger structures, such as conversations or texts (Coulthard, Johnson, and Wright 2017: 155; Sapir 1927: 904).

According to Dittmar (1996: 111) idiolect is "the language of the individual, which because of the acquired habitus and the stylistic features of the personality differs from that of other individuals and in different life phases shows, as a rule, different or differently weighted [communicative means]." So, because idiolect is in direct correspondence with the individual, it may change with age, according to life experiences that take place across time or in occasional episodes, or even as a result of some type of professional or leisure activity such as medical practice or football playing.

The study of idiolect posits practical problems pointed out by Coulthard (2004: 432) as concerning "how much and what kind of data would be needed to uniquely characterize idiolect" or "how the data, once collected, would be analyzed and stored". These problems associated with the "most common position in linguistics [of viewing] language as an abstract social construct" (Barlow 2010: 2) are at the base of authorship researchers' attempts to approach the analysis of the linguistic output by proposing functional concepts of the reflection of the idiolect in an individual's language realization like the concept of idiolectal style (Turell and Gavaldà 2013; Turell 2010).

Turell (2010: 217) acknowledges the difficulties in determining idiolect, since "countless amounts of data from each individual" would be needed to complete such a task. Therefore, she examines the usefulness of the idiolectal notion and proposes the concept of idiolectal style. She defines idiolectal style as largely associated with the mechanisms through which a system like language or a dialect that is shared by many people "is used in a distinctive way by a particular individual" and with the results obtained from such use, i.e., "the speaker/writer's production" which is – she affirms– "individual" and "unique" as described by Coulthard (2004), and optional or selected as posed by Halliday (1989) Halliday (1989) (p.217).

In her description, Turell goes from the concept of idiolect as an "idealized model of language" (Turell 2010: 216) to the notion of idiolectal style as one already in use, especially by forensic linguists, but not much discussed. That is, if idiolect is an individualized version of the language, idiolectal style is speech or text as realized by the individual. In this context, if the linguist's object of study is not the whole system but a specific linguistic occurrence, the amount of data needed to determine or verify individual style acquires a finite character. In forensic linguistics, for example, this finiteness is determined by the "information or clues which massively restrict the number of possible authors" in a given authorship problem, which reduces the number of texts and authors to analyze to a much manageable number in comparison to the whole system of language (Coulthard 2004: 432).

Turell (2010: 240) defined this finiteness in terms of "the populations involved and of many others". The data obtained from such a comparison of language usage is the researcher's Base Rate Knowledge. However, she admits that the task was "impossible" and in a later publication proposes that the amount of data needed for the analysis of a given authorship problem can be delimited by "a relevant population, or group of language users from the same linguistic community, with which the specific behavior of the speakers or writers under comparison can be compared" (Turell and Gavaldà 2013: 499).

This approach views the notion of idiolect in its practical dimension and suggests that the linguistic realization of an individual is compared to a reference sample of a relevant population to determine consistency and variation in language production. The proposal articulates what researchers had been doing ever since the first authorship problems emerged, i.e., focusing on the linguistic output and comparing it to similar outputs as realized by equivalent linguistic groups to obtain references of what is standard and what diverges from standard in specific contexts. Most importantly, the concept of idiolectal style opened the possibility of providing explanations of authorship phenomena based on the authors' sociolinguistic background.

2.2.2. Explaining Idiolectal Output

Idiolectal data obtained from empirical analysis to attribute authorship has been explained by resorting to aspects of sociolinguistic and cognitive explanations, as well as psycholinguistics and functional linguistics (Grant 2010; Grant and MacLeod 2018a).

The resort to different explanations in the context of authorship analysis, and specifically in profiling, has been more obvious at the lexical level, probably due to the relevance of words and their usage to determine meaning (Coulthard, Johnson, and Wright 2017: 110-22). It has been stated above that individuals acquire and apprehend language in their own terms and that such linguistic knowledge is realized in the choices they make when speaking or writing (Coulthard 2004: 432; Halliday, McIntosh, and Strevens 1964: 156). It is due to such personal apprehension of language that individuals "make typical and individualizing co-selection of preferred words" despite having the possibility of choosing any word or word combinations from language to express themselves (Coulthard 2004: 432).

The linguistic occurrence of collocations was first described by (Firth 1962) while establishing the differences between "context", "citation" and "collocation" for the determination of meaning in the context of lexicography. But, the co-selection of words following complex linguistic processes was later defined by Sinclair (1991) as the open-choice principal. This principle establishes that any linguistic output is abstractly constituted by empty places organized in units which may correspond to "a word, a phrase or a clause" and users of the language choose from a "wide range of [linguistic] choices" (p.109) to fill in such places and materialize a linguistic output while respecting the constraints associated with the respective place. The open-choice principle operates in contrast to the idiom principle which presupposes that the first does not impose enough constraints on a linguistic output, not even after dialect or register-related constraints have been applied (Sinclair 1991: 114). The idiom principle refers to "semi-preconstructed phrases that constitute single choices even though they might appear to be analyzable into segments" (Sinclair 1991: 110).

If there are pre-established words or combinations of words that can be used in the empty spaces that emerge while users produce linguistic output, then the choices available to the users for each unit they have to build are not so open as they can appear. Still, language users manage to build idiosyncratic linguistic output, since even idiomatic choices allow for grammatical, syntactic, and lexical variation (Sinclair 1991: 112).

Ever since the first theoretical approaches to the study of collocations appeared, the topic has been the focus of research of many linguists. Collocation has been generally defined as the "co-occurrence" of words, but as its understanding is not consistent among researchers in the field, it usually converges with the study of phraseology and formulaic language (Gries 2013: 138). In this regard, Christiansen and Arnon (2017: 543) refer to these co-occurrences as "multiword sequences" and acknowledge the existence of "different terms" to name the same phenomenon despite the "breadth of theoretical perspectives and backgrounds of the contributing authors".

The interest of authorship studies in multiword sequences concerns their potential to indicate traits of the user's identity and be of relevance to deciding semantic and pragmatic content of a linguistic output. This is because multiword sequences "mark out speech community members from outsiders" and their use and meaning is agreed in the "speech community" Wray (2017: 572)

Another explanation of how language users learn, store and use vocabulary has been proposed through the theoretical notion of lexical priming, introduced by Hoey (2005). Lexical priming is a theory that views collocations as a psycholinguistic phenomenon, describing it as the "psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution" (Hoey 2005: 5). According to the author "every word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word" (Hoey 2005: 9). These encounters include all the circumstances present in the context in which a word was learnt and that users of the language recall when using that word. Likewise, when primed words are used in word sequences, all the information around and about these words is "loaded" in the sequence, which becomes embedded or nested in the words comprising the sequence. Then the sequence is primed "in ways that do not apply to the individual words making up the combination" (Hoey 2005: 8). This theory suggests that when speakers choose words or word sequences, they also select the loads of information accompanying those selections. Thus, the way each person constructs the language is obtained "out of the primings acquired from a unique set of data", so their use of language is unique, “because all [their] lexical items are inevitably primed differently as a result of different encounters, spoken and written” (Hoey 2005: 181)

The uniqueness of the users’ linguistic output and their simultaneous belonging to a homogenous language system has also been explained by resorting to linguistic, social and cognitive arguments. According to Hudson (1996: 11) “no two speakers have the same language, because no two speakers have the same experience of language”.

In this respect, De Beaugrande (1999) describes how real language (as observed in large corpora) is governed by a “continual process of interaction among constraints” (p.131). These constraints can be “standing” (systemic) or “emergent” (discoursal) (p.131) and can have a linguistic, a social or a cognitive nature, ordering the heterogeneity of linguistic output as produced by different users. The author suggests a dialectic perspective for the relationship between linguistic dichotomies, such as “langue versus parole, competence versus performance, homogenous versus heterogeneous, general versus specific, social versus individual, regularity versus innovation, grammar versus lexicon, syntax versus semantics, and so on” (p.132). Within such a dialectic relation each side constantly contributes to the “evolving order of the other” (p.132).

More recently, the cognitive perspective served as foundation for a theoretical proposal on linguistic identity: the model on resources and constraints for authorship analysis (Grant and MacLeod 2018b). This is a conceptual framework conceived to account for a “new forensic authorship task – that of authorship synthesis” (p.82), but that allows for its application to authorship profiling and multiple-text problems or comparative authorship analysis since these type of tasks also seek to explain “the causes of consistency and variation in language production” (p.82).

Authorship synthesis is the “assumption of alternative identities in order to apprehend offenders in the context of the online sexual abuse and grooming of children” (MacLeod and Grant 2017: 157). In other words, it is the task of an authorized person, such as a police officer, “posing” online as a minor in order to uncover the anonymous offender who is contacting the minor and bring them to justice (Grant and MacLeod 2018a: 82). Because such “posing” takes place online, what the officer carrying out the task does is basically to incorporate the linguistic persona of another individual while suppressing his/her own (Grant and MacLeod 2018a: 91-92).

This theory examines “the relationship between language and identity” and proposes that identity is not solely “the result of externally imposed social categories” (Grant and MacLeod 2018b: p.81) expressed as sociolect or idiolect, nor exclusively the result of the employment of linguistic resources available in the context of linguistic interactions. In the authors’ proposal identity is a combination of both. That is, there are elements of the “linguistic persona” (p.85) that remain stable and can be described in terms of sociolect/idiolect while others are built upon linguistic interaction and will be characterized in relation to resources available to the users in the moment of the linguistic realization (p.86).

The novelty of the resources and constraints proposal (Grant and MacLeod 2018b) consists of the combination of notions of sociolinguistic background and discursive interaction to explain authorship realization. It subscribes to the dynamic character of the linguistic output of the individuals, establishing that it results from stable and dynamic resources that can be used together or that are mutually exclusive depending on the availability of the resources and/or context. So, the idiolect of any individual is the result of the interplay of such elements.

The theoretical notions examined above in relation to idiomatic expressions, collocations or formulaic phrases, and linguistic identity regard the individual as the center of the linguistic action. Thus, these notions contribute to supporting the concept of idiolect and explain idiolectal output which are at the basis of authorship attribution. Also, these concepts have been discussed because, although the OSRAs I analyze in chapter 4 are produced by multiple authors, the executive authorship (Love 2002) of the texts may be under the responsibility of only one author –usually the first – and therefore, understanding this can explain a given linguistic choice that may be essential for the analyses.

However, the focus of authorship profiling and therefore, of NLID, is on shared language, which is why variation at the group level is crucial for this research and thus is discussed next.

2.2.3. Language Variation: Dialects, Registers, Genres, and Styles

Of course, human language varies a lot. Anyone can observe this variation in their daily interaction with other individuals. However, the variation is not chaotic; it occurs in an orderly manner by individuals making choices at all levels of language, from pronunciation to word choice and order (Biber 1995: 1). For variation to be observed, it needs to be studied with reference to something with some degree of stability and homogeneity (Ferreira et al. 1996: 479). That "something" has been described as the linguistic norm or standardized language, standard variety, or even the correct or adequate variety of a language (Alfajarín 2013).

The linguistic norm may be characterized as a politically and institutionally accepted variety of a language, i.e., a variety of the language which due to "historical, economic and social reasons acquired functional and psychological independence among the speakers" and which is contained in "instruments such as grammar books, dictionaries, handbooks..." (Ferreira et al. 1996: 482)[5]. This is very similar to Trudgill's (1999: 117) definition:

> a language one of whose varieties has undergone standardization. Standardization [consists] of the processes of language determination, codification, and stabilization. Language determination refers to decisions that have to be taken concerning the selection of particular languages or varieties of language for particular purposes in the society or nation in question. Codification is the process whereby a language variety acquires a publicly recognized and fixed form. The results of codification are usually enshrined in dictionaries and grammar books. Stabilization is a process whereby a formerly diffuse variety (in the sense of Le Page and Tabouret-Keller (1985: 70)) undergoes focusing and takes on a more fixed and stable form.

The notion of correct language is associated with these concepts. The author also defines standard variety as "a dialect", i.e., "one variety [of a language] among many" that is "unusual […] in a number of ways" because "it is […] by far the most important dialect […]from a social, intellectual and cultural point of view; it does not have an associated accent" and it is "not a set of prescriptive rules" (Trudgill 1999: 123;25). This notion is also understood as a

[5] I translated and adapted from Portuguese. The original text reads "*Língua, no uso mais comum, é uma noção político-institucional. Corresponde a um sistema linguístico abstracto que, por razões políticas, económicas e sociais, adquiriu independência tanto funcional como psicológica para os seus falantes. Dão conta do funcionamento desse sistema instrumentos próprios, tais como gramáticas, dicionários, prontuários* (…)."

"model of reference for interdialectal communication" that is "neutral, common, or general" or as a register that is "somewhere between the colloquial and the more elaborate or specialized register" (Alfajarín 2013: 129). The concept of adequate language concerns the relativization of the relation between the norm or standardized language as the correct language and the standard variety to meet the needs of specific communicative situations (Alfajarín 2013: 130). Given that this study focuses on scientific writing, the concepts of linguistic norm or standardized language are taken as references for the analysis of the research corpora whenever necessary, in order to determine, for example, if a given use which can be accepted as normal in scientific writing, can also be verified in general language.

Variation has traditionally been described in reference to time, location, social organization, and discursive situation. From this perspective, variation can be labeled as diachronic, diatopic, diastratic, and diaphasic, respectively; and it can operate at all levels of language, from the phonetic to the discursive level (Ferreira et al. 1996: 480; Coseriu 1982: 19-20).

Diachronic or historical variation refers to changes in language across time, addressed within historical linguistics as the evolution of language at all its levels, but which has especially focused on the semantic level (Adler 2014: 15). Because this work considers linguistic occurrences from a synchronic perspective, this form of variation will not be further examined.

According to Ferreira et al. (1996: 480-81), diatopic variation describes language varieties according to geographical regions. This type of variation is also called geographical variation and it describes regional language varieties. Diastratic variation concerns variation according to the many aspects of speakers' sociodemographic dimension, such as age, profession, gender etc. This form of variation is also called social variation. Within this variation, a language variety that is shared by a social group and by means of which the group can be identified is called sociolect. Finally, diaphasic variation concerns the variation that takes place according to pragmatic and discursive contexts and which imply the knowledge and use of certain registers or ways of using language according to the formality of the situation.

However, in corpus linguistics and based on the systematic analysis of texts (Biber 1995: 1;7), variation has been divided into two main kinds: dialects and registers. From this approach, dialects include both diatopic and diastratic variations as described above. Therefore, there are geographic dialects or social dialects. Thus, geographic dialects refer to "varieties associated with speakers living in a particular location", and social dialects refer to "varieties associated with speakers belonging to a given demographic group [like] women versus men, or different social classes". That is, dialects are, as in the more traditional view explained above, the variation of language as exhibited by the users (Halliday, McIntosh, and Strevens 1964).

Registers, in turn, are defined by Biber (1995: 7) as referring to any "situationally defined variety" or a variety associated with a "particular context or purpose". According to the author, registers differ in their non-linguistic and linguistic characteristics; registers can be "named" within cultures, and registers "can be defined at any level of generality." Similarly, for Halliday, McIntosh, and Strevens (1964: 141) 'registers' are the variation of language according to its intended use. That is, language varies as the purpose of its use varies. In other words, individuals do not make the same linguistic choices to communicate about a health problem to a friend that they use to talk about their professional experience to a recruiter; or the language they employ to write an email to their parents or to a hierarchical superior.

Registers respond to a great extent to conventions concerning the appropriateness of language (Halliday, McIntosh, and Strevens 1964: 150). However, registers also correlate to contextual aspects that "operate across dialects" (Biber 1995: 4-5). Context may include several non-linguistic aspects that affect registers, such as the communicative objective, the relation among the speakers and the way they interact or the communicative situation (Biber 1995: 7).

Biber (1995) explains that the combination of non-linguistic factors influences the breadth of linguistic options that can be accessed within a given register. That is, some registers are more constrained than others. This is the case of the scientific register in comparison, for example, to the literary register. A biomedical researcher does not have more than one option for naming a given cell, but a novel writer may name an object within a story using many different words or word combinations. In a similar vein, Halliday, McIntosh, and Strevens (1964: 151) explain that “language is realized as the activity of people in situations, as linguistic events which are manifested in a particular dialect and register.” For this reason, language can only be explained by considering “various situations and situation types in which language is used”.

Biber (1995: 7-8) states that since situations may be of a very general and very specific nature, registers can also be defined in different levels of generality/specificity. Therefore, register differences may be seen as a continuum. Within such a continuum, registers are defined according to different aspects. The fewer aspects defining a register, the broader that register is, and vice versa. That is, in one extreme of the continuum, the registers found are differentiated by only one aspect, for example, the form of realizing the register, in writing or speech. At the other extreme of the continuum, “highly specified registers” are found. An example of the latter is the scientific article, defined not only by how it is realized, or the formality of the language used, but also by elements like the target audience, the topic, or the communicative function.

Some of the registers mentioned in the above continuum levels may be distinct enough in relation to purpose, intended audience, text conventions, and others, to be considered a genre (Biber 1995; Halliday, McIntosh, and Strevens 1964: 154). Genres are categories created by the human (Farrel 2003). Humans seem to be prone to organizing things, whether abstract or concrete, into categories, types, groups, subgroups, etc.(Rosch 1978). Because we are a species capable of thought and speech, we can also perceive patterns, events, or relations and translate them into rationalized arrangements, building systems of categories that help us manage cognitive load and information processing (Sweller 1988). Most readers would recognize the patterns of the quote below as a poem, and specifically as an ode:

SIR—All cases complete, the study was over *a*
the data were entered, lost once, and recovered. *a*
Results were greeted with considerable glee *b*
p value (two-tailed) equalling 0.0493. *b*
The severity of illness, oh what a discovery, *c*
was inversely proportional to the chance of recovery. *c*
When the paper's first draft had only begun *d*
the wannabe authors lined up one by one. *d*

(Fragment of the Poem "Ode to multiauthorship: a multicentre, prospective random poem", Horowitz et al. 1996)

The perception that this is an ode is based on its compliance with common characteristics attributed to this type of poems, such as rhyme (over-recover / glee-three / discovery-recovery / begun-one), stanza pattern (*aa, bb, cc, dd, ee*), or exaltation tone (i.e. "oh what a discovery"). These features enable the reader to recognize the text above as a poem, despite the word choice and theme being typical of scientific writing (e.g. "severity of illness", "*p*-value (two-tailed) equalling 0.0493", "inversely proportional"), and even though it was published in the scientific medical journal *The Lancet*. These features are common to the odes "family". The reader recognizes it because it has been formally established and conveniently taught that this is what the genre poem is supposed to be and sound like.

Genre typifies texts, and in recent understandings of the concept, genre also relates texts to social actions, takes part in cultural expression and plays a critical role in "meaning-making" (Bawarshi and Reiff 2010: 3; Miller 1984). For example, in the poem shown above, the authors are expressing their feelings and opinion about a topic, which hitherto remains controversial: (multi)authorship of scientific research articles and the conflicts derived from deciding who should be considered an author of a given research work. After an earlier history of single-authored publications, around the 1950s, research articles initiated a process of change in the number of authors signing one single research article (Cronin 2001). Multi-authorship was a natural response to a growth in collaborative work, especially in long-established fields like Physics and Medicine. However, at some point, multi-authorship also became an expression of an increasingly worrying phenomenon of opportunistic and unethical behaviors in relation to authorship crediting, which was and still is very closely related to publishing competitiveness and academic career advancement (King 2012). Culturally, the authors of this poem found what is probably the cleverest way to convey their opinions about the phenomenon within their discourse community. An ode is expected to depict a serious matter, and one worth praising using a dignified tone ([ode] 2019). Here, the genre is used to talk about a very serious concern for the scientific class, encouraging thoughtfulness, but also adding lightness.

Genre is one of those constructs that humans have created over time — something conceived by humans for the sake of organization of whatever social or cultural object or subject. There are various "systems" of genre classification: artistic, media, musical, literary, and, of course, language related. Current linguistic, rhetorical, and sociological traditions of genre were preceded by attempts at some sort of genre classification according to literary creations in Greece (Farrel 2003: 384). The Aristotelian classification of two types of poetry: (1) hymns/eulogies; (2) satire quickly developed into ancient literary criticism, which was followed by others in the "Hellenistic and Roman Period" (Farrel 2003: 391). The main division of "serious and elevated" in contrast to "less exalted" continue in the Roman period, but experienced some "hybridism" or "crossing" (Wilhelm Kroll in Farrel 2003: 392) creating poetry with aspects from both genres. The Roman Empire saw its end at the very beginning of the 13th century. Literary genres were then overtaken by Scholasticism, which took it back to Plato and remained underdeveloped and undiscussed until the 18th Century with Neoclassicism and a "number of factors contributed to the centrality of genre" (Prince 2003: 454).

From the late 18th century on, several theories were developed that contributed to knowledge on genre. According to Bawarshi and Reiff (2010), there are six main literary traditions or schools of genre: neoclassical; structuralist or literary-historical; romantic; post-romantic; reader response; and cultural studies. These literary traditions contributed to non-literary schools of genres, i.e. linguistic, rhetorical and sociological traditions which in turn have developed in their own terms, undergoing significant advancement in the last fifty years, beginning in the 1960s and opening new paths of applications and uses in fields like education and discourse analysis.

The connecting element among non-literary approaches to genre is the understanding that genres "reflect and coordinate social ways of knowing and acting in the world, and hence provide valuable means of researching how texts function in various contexts" (Bawarshi and Reiff 2010: 5). The focus of interest of this research work is in the scientific genre OSRA, described within English for Specific Purposes (ESP).

Another concept that is very closely related to register and genre is style. Halliday, McIntosh, and Strevens (1964: 154) defined style as the dimension of registers that concerns the "relations among participants" and the way in which these influence the linguistic choices of the language users. The authors described some basic categories such as "colloquial", "polite", "causal", "intimate", or "deferential", but assume that 'styles' are better seen as a "cline" in which linguistic features vary according to the "degree of permanence" which includes duration and hierarchy of the relation. The authors refer to relations like those established when individuals meet in public transportation, the relation between parents and children, or the relation between students and professors. The relevance of the relation and the influence it may have on the linguistic output "depend on the language concerned" (Halliday, McIntosh, and Strevens 1964: 155). For example, in Portugal, addressing professors in the university context, as a student, usually requires the use of titles to establish a certain distance of personal treatment which is culturally appropriate. The same approach, however, is not conventional in Spain or Brazil, and can even be considered an exaggeration.

Style has also been commonly understood as the distinctive, peculiar or idiosyncratic way of using language, which on the one hand, must be recurrent (McMenamin 2002: 110), and on the other, may be partly conscious and partly automatic (Olsson 2008: 30). This perspective defines style rather in terms of the individual language and less in terms of the language used by a community. Authorship analysis views style precisely as the individual linguistic choices made by users.

Given that establishing differences between registers, genres and style can be difficult, authors like Biber (1995) use the term register for only the more general levels of the register continuum. Therefore, the term genre is used for particular language varieties defined in relation to aspects linguistically more constrained. Some of these aspects are related to conventions of the organization of the text (Biber and Conrad 2019: 34) and the "message type that recurs regularly in a community (in terms of semantic content, participants, occasions of use, and so on)" such as an obituary (Ferguson 1994: 21). These language varieties may be the language of literary works such as novels or poems, the language of a recipe, or the language of an academic report (Biber 1995: 8). Finally, the term style is left to nominate more specific registers with regards to their linguistic features since the combination of these features brings the whole of the linguistic output into the foreground (Biber 1995: 9).

The different definitions of the terms register, genre and style convey the idea that these are really variations of the same phenomenon. In fact, the terms register, genre and style have been defined as ways of describing text varieties from different perspectives (Biber and Conrad 2009: 2), which reflects the co-existence and overlapping of linguistic and non-linguistic aspects surrounding these concepts and the complexity of language variability.

In this work the concepts of dialect, register, genre and style assumed are those in use within corpus linguistics with aspects taken from functional linguistics and sociolinguistics. Although the variation of the language analyzed in the present study is contained in a specific genre, i.e., the OSRA, the analysis regards variation more in terms of register and style than in terms of genre, attempting to pursue a functional approach.

2.2.4. Explaining Linguistic Output Influenced by (An)Other Language(s)

For historical, social and economic reasons many individuals of modern societies are exposed since birth to more than one language. Europe is a good example of a territory within which many languages co-exist, and within Europe there are territories where individuals will certainly learn more than one language since birth. In reference to a similar context, Hockett (1958: 321) wrote:

> "[…] someone born of English-speaking parents in Germany, who learns the one language from his family and the other from his playmates, possesses two idiolects rather than one. […] In some cases, it is impossible to decide whether a speaker has two rather similar idiolects or just one relatively flexible idiolect; fortunately, such marginal cases are not numerous enough to impair the practical utility of the approach."

Nowadays, the cases designated in the above quotation are not so marginal anymore. Language learning in Europe has increased exponentially in the last six decades (CE. 2014). Language learning has not always taken place in formal school-like environments. In the digital era, as can be easily observed, exposure to other languages and cultures has also occurred via software applications, social media, television and other audiovisual media. This exposure has brought about an increasingly bilingual or multilingual population. In the current global world, the idiolect of many people may comprise a "distinct and individual version" of the languages "they speak and write", which will become "manifest through distinctive and idiosyncratic choices in texts" (Coulthard 2004: 432). The knowledge of more than one language adds to the linguistic resources available to the individual as proposed by Grant and MacLeod (2018a). Understanding how this addition occurs and manifests itself in speech and writing has been one of the purposes of SLA studies.

The next section discusses the concepts of first language, mother tongue and native language, as well as the notions of second, foreign and additional languages.

2.2.4.1. Native Language, Foreign Language

Due to the unavoidable exposure to a number of mainstream languages associated with the globalization processes of modern society, many individuals nowadays are likely to be exposed to more than one natural language from the early stages of their lives. Many may learn a foreign language in interactions involving formal instruction or through informal means. Taking as a reference 2018, French, German, Spanish, Russian and Italian were the most frequently taught languages in Europe, with English, of course, at the top of the list in relation to upper secondary level (Eurostat 2018).

Provided that the foreign language instruction or the informal acquisition experience is maintained through the school years and that the language is studied and practiced systematically, the learning process can lead to language proficiency. With time, if individuals reach adulthood with a good knowledge of the foreign language, a state of bilingualism or multilingualism –if more than one foreign language is learned– can be achieved (Baker 2001).

Human linguistic behavior in relation to the knowledge and usage of more than one language, i.e., bilingualism or multilingualism, has received plenty of attention during the last 60 years (Dörnyei 2005: 6). Most of the research concerning this matter is related mainly to the fields of Linguistics, Education, Psychology, and Sociology or an intersection of those (e.g., Psycholinguistics) (Cenoz 2013).

Although bilingualism or multilingualism apply to any pair or group of languages an individual can speak and use with an acceptable degree of proficiency, it is in the context of English teaching/learning that these concepts are most frequently discussed (Baker 2001). For this reason, an important part of the past and current research on the wide-ranging topic of foreign languages teaching/learning is centered on English as a second, foreign, or additional language.

The notions of 'second' and 'foreign', when used in the framework of English language teaching/learning, refer to the sociodemographic and pedagogical contexts in which the language knowledge is acquired. 'Foreign' was the label some academics began to use by the end of the 1940s to name the teaching/learning of English to/by non-English speaking individuals in an attempt to differentiate the "English language use and learning that was felt to be different "from the native situation and that was physically outside the native speaking countries" (Nayar 1997: 14). The label "second" developed especially after World War II, following demographic movements to countries that had not suffered the direct devastating impact of the war (e.g., the United States of America), and further to advancements in the field of Structural Linguistics concerning the influence of the first language in second-language learning (Nayar 1997: 11-12).

Currently, English as a Foreign Language (EFL) refers to a context of teaching/learning in which a person has a first language/mother-tongue and is learning a foreign language in his/her own country or in a country where the language is not a native and official language (Nayar 1997). This would be the case of Portuguese and Spanish students who learn English in their respective countries in schools, sometimes from an early age or in other institutions at later stages of their lives.

On the other hand, English as Second Language (ESL) is concerned with a teaching/learning context in which a person has a first language/mother-tongue and is learning another language in a country where he/she lives permanently or in some long-term temporary situation, and where that foreign language is official and used in regular instruction and society in general. The concept can apply to at least two types of situations. One refers, typically, to immigrants in the United States, the United Kingdom, Australia, Canada, or New Zealand, who do not speak English and learn the language in order to integrate and function in society. The other perspective is concerned with English-speaking countries like Bangladesh, Ghana, India, Kenya or Malaysia where English is an official –but not an endemic language– thus used widely in education and to communicate at all levels of society by a population who does not actually require English to communicate due to its multilingualism (Nayar 1997: 15). English as an Additional Language or as an Associate Language (EAL) refers to the latter description of an ESL situation (Nayar 1997: 19).

Both ESL and EFL emerged and have traditionally been presented from the perspective of native speakers, that is, English as a Native Language (ENL). This view takes 'nativeness' as a reference and assesses non-native speakers' speech realization based on native speakers' competence and performance of the language (Nayar 1997: 14). Other ethnolinguistic and more literal terminologies that paralleled this triad were those proposed by Strevens (1982: 420): "English-speaking, English-using, and non-English-using countries." These correspond to countries such as Great Britain, The United Stated of America, or Canada for English-speaking; countries such as India, Bangladesh, or Jamaica, for English-using; and countries such as Portugal, Spain, or Brazil for non-English-using.

Ethnolinguistic views usually focus on what ESL/EFL speakers do wrong in relation to ENL speakers, putting the native variety of the language as an ideal of realization of the language non-natives are expected to achieve. Ethnolinguistic views have been gradually replaced by sociolinguistic understandings of English-speaking users like that first discussed by Kachru (1985), aiming to provide "fresh conceptualization" of "world Englishes" (Kachru 1992: 3) and recognize "historical, educational, and functional distinctiveness" (Kachru 1997: 68). His perspective is that of a "stratification" of English contained in "concentric circles," referring to three main groups of English users in the world according to "the types of spread, the patterns of acquisition and the functional domains" (Kachru 1985: 12). These circles are named by the author "the inner circle," "the outer circle (or extended circle), and "the expanding circle," and are in fairly direct correspondence with the concepts of ENL, ESL, and

EFL respectively, and with the groups described by Strevens (1982). However, Kachru (1985, 1992) offers a perspective that takes into account the motivation behind the need to learn English and its uses "across cultures and languages" (Kachru 1985: 12).

These groups of English users also differ in their "speech fellowships," Despite their differences, these groups belong to a "wider speech community" of English users, which, in turn, is contained in a language "community" (Kachru 1985: 15-16). Table 3 below summarizes Kachru's view of English-speaking users in the world.

STRATUM	TYPES OF SPEECH FELLOWSHIPS	COUNTRIES OF REFERENCE
The Inner Circle	Norm-providing varieties	USA, UK, Canada, Australia, New Zealand.
The Outer or Extended Circle	Norm-developing varieties	Nigeria, Pakistan, Philippines, Singapore, Sri Lanka, Tanzania, Zambia etc.
The Expanding Circle	Norm-dependent varieties	China, Egypt, Israel, Japan, Saudi Arabia, Portugal, Russia, Spain etc.

Table 3 – Stratification of World English according to Kachru (1985, 1992)

Both ethnolinguistic and sociolinguistic perspectives of English teaching/learning agree that the separation among groups of English speakers/users is not clearly delimited or definite in the outer (ESL users) and the expanding (EFL users) circles. Fluctuations in this regard may depend on the users' language knowledge, use of language, acceptance by a wider speech community, and linguistic policies (Kachru 1985: 13-14).

More recent theoretical approaches on the relationship between English and other languages worldwide include the need to consider 'multilinguism' and 'translinguism'. These paradigms view the more traditional divisions of World Englishes as "ideological inventions" which do not reflect "everyday language practice" and implicitly refer to some Englishes (native speakers', for example) as better than others (Lee and Canagarajah 2021: 99). Despite the importance of these new paradigms, a comparative study will require references to whatever aspect is being compared between groups. Therefore, in the framework of the present work, users are divided into two main groups. The first group belongs to or is from the so-called inner circle, i.e., native speakers of English, and native speakers of Portuguese or Spanish. The second group would typically belong or would be from the expanding circle

(Kachru 1985, 1992), i.e., users of English who are native speakers of the European varieties of Portuguese/Spanish. However, this division does not seek to project any form of superiority of native speakers. It aims at establishing references for the comparisons. Furthermore, it is admitted that the groups belonging to the expanding circle could also be identified with characteristics of the outer circle as explained in section 3.1.2.5. Finally, the label L1 is assumed to refer to the first group; and the label non-L1 is assumed to denote speakers from the second group.

2.2.4.2. Theories on Cross-Linguistic Influence

Both the ethnolinguistic and sociolinguistic views of the language learning process have naturally generated the need to understand the effects of the learners' first language on the foreign language they acquire, and ultimately, of one language on another. This field of studies is designated as cross-linguistic influence (CLI) or transfer studies, although the first term is more wide-ranging and covers more cross-linguistic phenomena, such as "overuse, underuse, and avoidance of language forms, functions, and structures in one language due to the influence of another language, as well as cross-linguistic effects at the level of conceptualization and mental processing"; transfer studies, in turn, refers to "transfer of a form, structure, or meaning from a person's knowledge of one language to their use of another" (Kellerman & Sharwood Smith, 1986 in Jarvis 2012a)

Within CLI, several theoretical assumptions have developed in the last decades that attempt to explain the relation between an L1 and a non-L1 user. Among the most important is the theory of markedness which suggests that language parts correlate in pairs of "least distributed" (marked) versus "more distributed" (unmarked) elements (Isurin 2005: 1115). The designations marked/unmarked seemed to have been first used in 1930 by the linguists Trubetzkoy and Jakobson, but the notion of markedness had been noticed earlier in 1815 by G. M. Roth (Henning 1989: 21; 15). Moreover, the concept has evolved significantly since it was initially proposed, becoming one of the most discussed in linguistics due to the many interpretations it has had (Henning 1989: 11).

In the context of cross-linguistic analysis, this theory is used to try to anticipate the structures of non-L1 linguistic output that are more likely to be replaced with the corresponding structures of the users' L1, based on the assumption that "those linguistic phenomena in the target language which are more marked than the corresponding phenomena in the native language will be more difficult to learn" (Isurin 2005: 1115). In a similar vein, markedness in texts written by authors in a non-L1 is expected to be found in whatever structure is least distributed, i.e., more marked in the non-L1 than in the L1. The concept of markedness was used, for example, by Turell (2010: 215) "to establish the rarity in the frequency of use of two grammatical variables" with different distributions in Spanish and Catalan, although the case she refers is not one of a person writing in a non-L1, but rather of a person writing in his L1 after prolonged contact with a non-L1.

Another similarly important theory is the conceptual transfer hypothesis (CTH). This theory connects with cognitive linguistics, and its significance rests in the fact that it tries to connect a user's experiences in one language to the acquisition and development of another (Jarvis 2012a: 1556). Conceptual transfer is described by Odlin (2005: 6) in terms of another CLI hypothesis – that of linguistic relativity or the assumption that language influences thought. Thus, conceptual transfer is defined by these authors as "those cases of linguistic relativity involving, most typically, a second language". A more detailed definition is offered by Jarvis (2012a: 1555), who specifies that conceptual transfer refers to "language behavior or language-related behavior [exhibiting] CLI effects that are interpreted as having taken place in the person's conceptual system before the conversion of his or her preverbal message into language".

Jarvis (2007: 52) distinguishes between the "concept transfer" and "conceptualization", both included in CTH. Concept transfer is "conceptual transfer related to a person's conceptual inventory", while conceptualization refers to "conceptual transfer stemming from a person's patterns of conceptualization". The first concerns "the makeup of the inventory of concepts in a person's long-term memory" and the latter regards "the process of selecting specific concepts from long-term memory, calling them up into working memory, and combining them dynamically in various orders, structures, and configurations in order to construct temporary representations of various types of phenomena (e.g., smells, sounds, tastes, feelings, relationships, and dynamic visual images of objects, events, scenes, situations, episodes), whether real or imagined" (Jarvis 2007: 54).

Other relevant theories attempting to explain the direction of the cross-linguistic influence are the structural overlap hypothesis and the language dominance hypothesis. In a user with knowledge of two language systems, such as an L1 and a non-L1, CLI due to structural overlap can take place when the user's non-L1 language system has more than one form for a given structure, but only one of those forms is similar to the form that structure takes in the user's L1. This overlap may cause the non-L1 user to favor his/her L1-like form of the structure more frequently than native speakers of his/her non-L1 language would (Foroodi-Nejad and Paradis 2009: 411). Similarly, the language dominance hypothesis establishes that the language a person uses with greater proficiency is also the language at the base of the user's patterns or structures choices (Foroodi-Nejad and Paradis 2009: 412). The difference among these two theories is that, in the overlapping proposal, the direction of the CLI can go both ways, although usually between languages with similar degrees of development concerning the user; and the dominance hypothesis assumes that the CLI is unidirectional, i.e., always from dominant to non-dominant.

Finally, the contrastive analysis hypothesis (CAH) refers to the differences that exist between two languages, specifically between an L1 and non-L1 of a person. This theory claims that the characteristics of a person's L1 that make that language different from the person's non-L1 can potentiate errors in the person's non-L1 performance (Dulay et al., 1982 in Sinha et al. 2009: 118).

The conceptual and theoretical contributions described above have advanced cross-linguistic studies significantly. However, one that has been most relevant for explaining the effects of one language on another is that of interlanguage (Selinker 1972). The research developed in the ESL/EFL fields has been usually situated within an educational perspective, focusing on the teaching/learning processes, strategies, mechanisms, techniques, methods etc. However, the theoretical proposals developed around bilingualism and multilingualism can help explain linguistic phenomena from contexts like that approached in this study, i.e., authorship of scientific writing produced in English by L1 and non-L1 users of the language. In this regard, the theory of interlanguage and the concepts of hybridization and glocalization are discussed next.

The theories of cross-linguistic influence examined above may be useful in explaining possible effects of L1 influence in the scientific text. The theory of Interlanguage is taken as the reference to explain possible interlingual instances.

2.2.4.3. Interlanguage, Hybridization, and Glocalization

Interlanguage is a concept from a psycholinguist theory of the field of Second Language Acquisition (SLA), a theory that originated in the need to understand, from the psychological point of view, the learning processes experienced by learners of a second language, typically English, and explain their errors when producing utterances and sentences. The ultimate objective was to improve language instruction and contribute to "isolate relevant data of second language learning"(Selinker 1972: 210-11). The Interlanguage Theory (IT) is based on the concept of "interlingual identification" of speech units, which examined the need to understand bilingual learners' mental processes when in a situation of language contact (Weinreich 1953: 7).

Interlanguage applies only to what Selinker (1972) calls "meaningful performance situations" (p.210), or situations in which individuals over the age of 12 attempt to produce meaning they probably already have, in the language they are learning. The author establishes that only a limited percentage of second language learners (likely, 5%) will achieve "native-speaker competence" because they are able to "reactivate" the latent language structure as proposed by Lenneberg (1967: 374-379 in Selinker 1972: 212), that is, a brain structure all individuals are born with and that is first activated when exposed to what will eventually become their mother tongue. For the rest of the vast majority of second-language learners, who most probably will not achieve "absolute" success on the second language production and understanding, he proposes an "already formulated arrangement" (also located in the brain) he calls the latent psychological structure (Selinker 1972: 212).

The utterances produced by this second type of learner in the language they are learning is what Selinker calls "the observable data" to which theoretical predictions can be associated. Such utterances are produced in what he calls Interlanguage: "a separate linguistic system based on the observable output which results from a learner's attempted production of a TL [target language] norm." (Selinker 1972: 213-14); or as recently stated by the same author “Interlanguage is that linguistic/cognitive space that exists between the native language and the language that one is learning. Interlanguages are non-native languages which are created and spoken whenever there is language contact” (Selinker 2014: 223). So, Interlanguage is not the linguistic system of the native language of the learner, nor is it the system of the language he/she is acquiring, better known as target language. It is a system in-between the native language and the target language and it is presented as a normal, structured, systemic and dynamic outcome of the learners’ language development, containing new/novel forms (Selinker 1972, 2014).

Because interlanguage utterances need to be observable, Selinker establishes that only three types of utterances can be used to identify interlanguage. Taking as a reference an individual who is learning a second language these would be:

- NL – Utterances that individual produces in his/her native language;
- IL – Utterances that individual produces in the language he/she is learning;
- TL – Utterances produced by a native speaker of the language being learned by the individual.

These utterances, also called “behavioural events,” represent the data that is relevant to second language learning, and the information that can be obtained from the IL events are the observable “surface structures” of the psycholinguistic processes underlying second language learning. One mechanism presented by Selinker as being very relevant to the five processes underlying interlanguage is fossilization: “linguistic items, rules or subsystems” that the learner continues to produce in his/her interlanguage in contrast to what would be appropriate in the target language, regardless of his/her age or “amount of explanation” received in the target language. The “linguistic items, rules or subsystems” that are “fossilizable” are those that will become the “most interesting” linguistic phenomena of IL performance.

According to Selinker, five main processes can explain an IL event. These processes, Selinker says, can occur in isolation or in combination. Table 4 presents such processes (Selinker 1972: 215-16).

PROCESSES UNDERLYING IL		DEFINITION
1.	Language transfer	An IL performance, or part(s) of it, that derives from the learner's native language
2.	Transfer-of-training	An IL performance, or part(s) of it, that derives from the training practices used to teach the second language
3.	Strategies of second-language learning	An IL performance, or part(s) of it, that derives from the approach(es) used to learn the second language
4.	Strategies of second-language communication	An IL performance, or part(s) of it, that derives from the approach(es) used to communicate with a native speaker in the second language
5.	Overgeneralization of TL linguistic material	An IL performance, or part(s) of it, that derives from extending "rules and semantic features" of the target languages to second language structures under a certain logic.

Table 4 – Psycholinguistic processes underlying Interlanguage as proposed by Selinker (1972: 215-21).

These psycholinguistic processes are directly related to "fossilizable items, rules and subsystems" of the interlanguage, that is, structures that have the potential to remain in a certain form "no matter what the age of the learner or amount of explanation and instruction received in the target language" (Selinker 1972: 215). These forms deviate from what a native speaker of the language would produce in the same communicative situation. In other words, these processes are the means used by the non-native user of the language to realize fossilizable structures in IL utterances/sentences. When fossilizable structures are realized using combinations of these processes, "fossilized IL competences" take place. The occurrence of "fossilized IL competences" in a group of individuals can result in a new dialect in which such competences would be the norm (Selinker 1972: 217).

The Interlanguage Theory offers an excellent conceptual ground to explore the realization of English in scientific writing by L1 and non-L1 users of the language. Among the five processes described by Selinker, language transfer or CLI emerges as one of the most studied and central to the theory and to empirical research like this.

Two other concepts that may help understand the relationship between an individual's L1 and non-L1 are hybridization and glocalization.

The hybridity theory was first borrowed from biology and applied to language by Bakhtin (Sanchez-Stockhammer 2012: 134). Hybridization has been defined as "a process whereby separate and disparate entities or processes generate another entity or process (the hybrid), which shares certain features with each of its sources, but which is not purely compositional". Moreover, hybridization can occur at different language levels, from more complex to more basic, i.e., at the communication level, at the languages levels, in text types, texts, sentences, clauses, phrases, idioms, collocations, words, morphemes, and sounds (Sanchez-Stockhammer 2012: 134).

It is not possible to address in this work all forms of hybridization, nor is that the objective of this research. However, some of the above-mentioned levels are worthy of note to explain how hybridization takes place. The level of "individual languages" (Sanchez-Stockhammer 2012: 145), for example, is particularly relevant since it can translate what could occur in an interaction between an L1 and a non-L1 in scientific writing in the context of English as the lingua franca of science. This level is described by Sanchez-Stockhammer (2012) in literary terms as what takes place in the "hybrid novel" in which "Western and post-colonial (native) writing traditions creatively interact" (Fludernik et al. 2005, 227 in Sanchez-Stockhammer 2012: 145). Similarly, traditions of scientific writing may be blended together, resulting in hybrids, i.e. "texts that have features of more than one style [that could represent] unintentional intrusion of features from the 'traditional' style into a discourse that is attempting to be modern" as described by Bennett (2008: 206) referring to her own corpus of study.

Within the level of individual languages, interlanguage is described by Sanchez-Stockhammer (2012: 148) as a hybrid language since users "fill gaps" they may have in their non-L1 by resorting to elements from their L1. It is in this perspective that the concept of hybridization may be useful in describing non-L1 realization of scientific English, especially at more advanced levels of proficiency where interlanguage may easily be considered as being within the norm, because hybridization occurs at more complex levels of language such as discourse.

Finally, the perspective of hybridization connects with the notion of glocalization in terms of the merging and interdependence that occurs, in many societies, of global and local practices or demands within a given area (Frello 2013). Although to the best of my knowledge, there is no theory of glocalization, there are proposals explaining the concept in terms of globalization involving on the one hand, the specifics of universal values, and on the other, the commonness of values that are specific to a given society, i.e., a way of incorporating the local in the global, and vice versa (Roudometof 2016: 2). In relation to this, Pérez-Llantada

(2012: 164) discusses “emerging ‘glocal’ discourses that hybridize the Anglophone standardized norms with their unique rhetorical traits”, showing that glocal practices are evident, for example, in the Spanish–English Research Article Corpus (SERAC) where “different culture-specific linguistic preferences, rhetorical traits, and intellectual styles” (Pérez-Llantada 2012: 175) are maintained in texts produced by non-L1 users of English. The connection of hybridization and glocalization with the concept of interlanguage is interpreted here as the recurrent participation of one system in another. In this case, the systems are language systems and the participation may encompass all level of language at which the influence of the scientific authors’ L1 writing in English can be observed, with this influence being continuous, i.e., it is not sporadic but frequent.

2.3. NLID in Research Articles

Most of current non-native English-speaking scholar communities use English as their standard means of international and sometimes even national scientific communication. This fact holds to be especially indisputable for written scientific communication. More and more academic work – like dissertations, reports, or protocols – is produced entirely and directly in English, often avoiding native languages during the whole writing process and assuming a completely foreign language system as a natural means of communication.

Scientific articles are probably the best example of this. According to data from Ulrich’s Web Directory[6] from 2018, about 78% of all current scientific articles in the world are published in English (Johnson, Watkinson, and Mabe 2018: 25). Scientific journals, especially those holding the “peer-reviewed” and “indexed” badges, are more than ever produced directly and exclusively in English, even when complete editorial teams are based in non-English speaking countries and consist of a majority of non-native English speaking affiliates (González-Alcaide, Valderrama-Zurián, and Aleixandre-Benavent 2012: 4). This is the case, for example, of the former *GE - Jornal Português de Gastrenterologia*, currently known as “GE - Portuguese Journal of Gastroenterology” ('GE Port J Gastroenterol' 2019), which only accepts

[6] “Ulrich's™ is the authoritative source of bibliographic and publisher information on more than 300,00 periodicals of all types academic and scholarly journals, Open Access publications, peer-reviewed titles, popular magazines, newspapers, newsletters and more from around the world. It covers all subjects, and includes publications that are published regularly or irregularly and that are circulated free of charge or by paid subscription.” ('Ulrich's Periodicals Directory - Ulrichsweb' consulted on Oct 25, 2019).

articles written in English, leaving most of its authors' native Portuguese to be used in the counterpart of the articles' abstracts, the so-called "resumo."

The relevance of English in publishing scientific discoveries and in researchers' scientific careers depends on different factors (Pérez-Llantada 2012). Some factors concern the pursuit of international recognition and professional promotion in scientific careers (p. 5). Other are related with the need to communicate scientific research results to larger peer audiences that can access, discuss and validate science (p. 50). Finally, there are factors like the pertinence and relevance of the findings being communicated (84-85). However, in many scientific fields like engineering, computing, physics, natural sciences, and health sciences, most of the journals considered to be of high impact factor (HIF), and thus, of better quality, are available only in English (Hamel 2007: 58; Benfield and Feak 2006), making it necessary to resort to that language to disseminate any type of scientific discovery.

The research article is currently one of the most important sources of scientific knowledge communication. The relevance of the scientific article is given by the main purposes it serves of being a primary source of scientific data and news; a space of knowledge claiming; and a form of priority establishment that provides researchers with the recognition and the acceptance of their peers (Holmes 1987: 220; Gross et al. 2002: viii). Moreover, given its rather short extension in comparison to other scientific genres, the research article is an excellent resource for researchers to respond to the need for almost instant sharing of scientific discoveries fostered by the competitiveness for professional and scientific career advancement that the rapid growth of information and knowledge societies has provoked (Johnson, Watkinson, and Mabe 2018: 13).

2.3.1. ESP, Discourse Community and Communicative Purpose

Within the field of Foreign/Second Language Teaching, the interest in the study of languages like Russian, German, French, Portuguese, and of course, English, for specific instead of general purposes, began gaining relevance around the 1970s (Strevens 1977: 145). In the next 20 years, English for Specific Purposes (ESP) developed in parallel with the increasing relevance of English as an international language of communication in science, technology, and trade (Johns and Dudley-Evans 1991: 297). The 'specific' in ESP refers directly to the “students’ own specific language learning purposes” (Belcher 2009: 1), which can be diverse.

Currently, some of the most prominent branches of ESP are English for Academic Purposes (EAP) and English for Occupational Purposes (EOP) (Bawarshi and Reiff 2010). Within the latter category, more specific branches are contained such as English for Business Purposes (EBP), English for Legal Purposes (ELP), and English for Medical Purposes (EMP); and then some branches are a combination of EAP and EOP such as English for Academic Medical Purposes (EAMP), English for Academic Business Purposes (EABP), and English for Academic Legal Purposes (EALP) (Bhatia 2014).

In all these cases, the individuals’ purpose in learning English is closely related to the need to perform within a ‘discourse community’ in the context of specific situations (Belcher 2009: 3-5). Accordingly, ESP is a field of research and instruction of “specialized varieties of English,” typically taught to “non-native speakers of English, in advanced and professional settings” (Bawarshi and Reiff 2010: 41).

In ESP, a “discourse community” is defined by Swales (1990: 21-29) as a group of individuals who: (1) has “a broadly agreed set of common goals [that are] tacit or formally inscribed”; (2) has “mechanisms of intercommunication” that can be physical spaces like “meeting rooms” or communication means like a periodic “newsletter”; (3) “uses its participatory mechanisms […] to provide information and feedback”, i.e., the members of the group interact through such mechanisms; (4) uses genres that are specific to the group in the “communicative furtherance of [the community’s] aims” (5) has “specific lexis,” which can eventually evolve to “shared and specialized terminology” (6) has a “threshold level of members with a suitable degree of relevant content and discoursal expertise” who are in the position to transfer knowledge on “shared goals” and “communicative purposes” to members entering the community (Swales 1990: 24-27). Typically, researchers of any given scientific area, usually organized in research groups, constitute a discourse community. That is the case, for example, of medical researchers in the field of gastroenterology or metabolic diseases who meet all of the six characteristics described by Swales (1990).

Given the demand for specific language objectives to be met, ESP has to focus on specific occurrences of the language used by discourse communities to respond to specific situations that ESP’s target audience wants to access. To understand these occurrences of language, ESP instructors/researchers have to focus on genres and the “contexts in which they function and interact with other genres: how one genre responds to others (intertextuality or interdiscursivity)" (Belcher 2009: 4). For example, in the context of EAP, the applications submitted in reply to research position calls will, in turn, give rise to responses from research selection boards, and the latter can eventually give place to selection process revision upon request from at least one applicant.

Therefore, around the time ESP grew in importance, genre analysis also gained relevance as a research topic and pedagogical tool (Bawarshi and Reiff 2010: 41). However, it was Swales (1990) that "most fully" provided the field with theory and methodological development, bringing genre analysis and ESP to such a common ground that they are often consider equivalent (Bawarshi and Reiff 2010: 41).

According to Swales (1990: 1), ESP uses genre analysis for applied purposes in the field of "academic and research English." The development of the ESP perspective on genre analysis was initially more focused on descriptive "quantitative studies of linguistic properties of language varieties" (Bawarshi and Reiff 2010: 42). However, as ESP studies have increasingly focused on more specific academic areas carrying out more and more sophisticated analyses, natural evolution has occurred that has gradually led ESP genre analyses towards rhetorical investigation, connecting "linguistic and rhetorical studies of genre" and broadening interests to "communicative functions" (Bawarshi and Reiff 2010: 42).

In the context of ESP, genre works like a tool that helps members of the discourse community "achieve and further their [own] goals," but also and eventually, the goals shared by the community (Bawarshi and Reiff 2010: 45). Ultimately, genre is the means by which discourse community members accomplish their own and shared goals. Thus, realizing the role of the genre in an adequate manner is extremely relevant to guarantee and maintain permanence and belongingness in the community. As a consequence, ESP has concentrated efforts on the analysis of genre with the main purpose of instructing (essentially non-native users of English) on genre. Therefore, genre is a central concept to ESP defined by Swales (1990: 58) as:

> “a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community and thereby constitute the rationale of the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style.”

Thereafter, the highlight of the importance of “communicative purpose” is given by Swales when he affirms that it is a “privileged property of genre” and one that regulates the “scope of a genre” so as to maintain it within a “comparable rhetorical action” (Swales 1990: 58). Furthermore, other properties of the genre (e.g., form structure) work as regulatory mechanisms of the “prototypicality” of the genre. The closer the genre realization is to the communicative purposes and other properties, the more “prototypical” it is of that genre (Swales 1990: 52).

This understanding of “communicative purpose” as a key-player of genre has led ESP to gradually pay closer attention to the rhetorical side of genre, focusing on rhetorical moves and then finding and characterizing the linguistic features that correspond and realize such moves. It may, therefore, be said that ESP usually follows a top-down approach to genre analysis, or from “context to text” (Bawarshi and Reiff 2010: 47). Although this type of approach can be operationalized in more or less detailed steps (Bhatia 1993), an analyst following an ESP method would always consider at least “identifying purpose” in relation to a “discourse community” and then analyzing the realization of the genre in terms of “rhetorical moves” and its corresponding “textual” and “linguistic” features (Bawarshi and Reiff 2010: 48).

Researchers in the field of ESP have significantly contributed to the advancement of the study of “discipline-specific” genres, especially of different parts of the research article like the introduction and the abstract. Studies on the totality of a research article are less frequent. The present work focuses on the genre OSRA as the locale where the authorship analysis takes place. However, from the point of view of the analysis it has no intention of examining the rhetoric of the genre or its parts – IMRAD, i.e., Introduction, Methods, Results and Discussion -, as described by (Swales 1990: 137-74) except for the importance of the characteristics of the genre to discuss the results obtained in the analysis of the corpora.

2.3.2. Multi-authorship in OSRAs

Many different types of research outputs are considered to be of great value to demonstrate outstanding scientific activity in any given field and, consequently, contribute to fostering Information & Knowledge Societies. For example, in Portugal, the Foundation for Science and Technology (FCT) – the main national public scientific research funding body – provides evaluation guidelines for Research & Development Units (R&D units) on what types of scientific outcomes to report. The document of the latest FCT R&D units evaluation process, carried out in 2018, clearly requested reports on output in relation to:

> "contributions for knowledge advancement and/or application; publications; advanced training; initiation of undergraduate or Master students in research; organization of conferences, colloquia and/or seminars; patents, prototypes or products; knowledge and technology transfer; spin-offs; preservation, curation, and dissemination of R&D results and data, respecting the principles and practices of Open Science; promotion of scientific and technological culture (outreach); actions of special scientific, technological, cultural, artistic, social or economic relevance to society" (FCT 2018).

In the field of health, particular importance is given to the interface with society. In general, some of the most recent indicators of excellent research activity are patents of inventions resulting directly from research; the participation in international research networks to address global health problems like obesity, diabetes, or asthma; or the active involvement in the patients' associations activities that allow rapid understanding of health problems and feedback from health professionals and policymakers.

However, one of the standard indicators of the research quality still is the number of peer-reviewed publications. In the context of scientific output, original research articles (ORAs) have been historically regarded as having paramount importance for building a scientific career. This long-employed form of communication among peers is the primary source of knowledge about any given topic in science, technology, and humanities. For this reason, scientific publications such as reviews, short communications, case studies, letters or research articles (RAs), among others, published in scientific journals remain and will continue to be important. There is a common expression among academics that says that a researcher must "publish or perish," referring to the fact that if research activity is not communicated to one's peers, and most importantly, indicated as a reference by other scholars, then it is unknown, and it cannot be validated, which means that it does not exist in the scientific arena.

The first scientific papers published were usually single-authored, but the development of scientific methods and the advancement of knowledge has gradually favored the increase in the number of researchers who assume authorship of an article.

Though it is more frequent in areas like the Biomedical Sciences, Physics, or Medicine than it is the Humanities, it can be said that in general, authorship has become a collective condition.

Given their usual collective nature, scientific papers can technically have as many authors as the number of persons who contributed to the results being communicated. RAs with many authors are usually those reporting outcomes that resulted from very large collaborative projects. Some rather extreme examples are those described by Leung et al. (2015) on the genetics of the Fruit-fly, authored by one thousand investigators (Woolston 2015); the RA by Aad et al. (2012) reporting on the ATLAS Collaboration, published under the responsibility of two thousand nine hundred thirty two authors (King 2012); or the publication by Aad et al. (2015) reporting "a more precise estimate of the size of the Higgs boson" (Castelvecchi 2015) and that apparently broke "the record for the largest number of contributors to a single research article" (idem) with five thousand one hundred fifty-four authors.

Hyper-authorship does not seem to be such a recent tendency as the years of publication of the previously referred articles would suggest. Already in the 1990s, there were papers with as many as 182 authors being published on genome sequence (Cronin 2001). As exemplified above, the field of Physics is currently the one with the higher number of articles authored by more than one hundred authors (King 2012). However, papers with a high number of authors are also common in health sciences in areas like the autoimmune diseases, such as Langefeld et al. (2017) with 108 authors.

Still, RAs with lower numbers of authors reporting smaller-scale research results are the norm. In the health sciences, the average number of authors per article is between 5 and 10. Levsky et al. (2007: 371), in a study about the impact of "publications for promotion in academic medicine", refer an average of almost 6 authors per article in this area.

Overall, RAs are made up of text that has formally been produced by the number of people announced in the authors' list of the publication. Nonetheless, in practice, scientific authorship refers mainly to the scientific contribution of each author to the research topic and subsequent conclusions derived. The weight of contribution to textual authorship is usually lighter and placed in the background. In other words, authorship refers to the creation of the knowledge being communicated, rather than the text by which that knowledge is conveyed. Authorship carries a meaning related to knowledge creation. Textual authorship is assumed as a task that does not have to be carried by all listed authors at all moments of the article writing. Any journal's authorship policy will provide information on the eligibility criteria for authorship, which includes criteria in relation to the writing of the article. In the field of Health, the authorship policies of journals usually follow the International Committee of Medical Journals Editors recommendation (ICMJE), which has established the following criteria to help define whether the role of a participant in a research study is that of an author or a contributor:

- Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
- Drafting the work **or** [my emphasis] revising it critically for important intellectual content; AND
- Final approval of the version to be published; AND
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

[Defining the Role of Authors and Contributors]. Retrieved from URL http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html on 06/10/2019

Although the criteria are cumulative, it must be noticed that with regards to the textual production the second recommendation offers alternatives. As far as the writing of the article is concerned, both the person(s) who wrote and those who revised the article are considered to have fulfilled that criteria of authorship. Furthermore, the revision role is defined as that which is "critical for important intellectual content", but the interpretation of what is "critical" and "important" is left to the consideration of the authors. Consequently, there is space for a researcher who has fulfilled the first, third and fourth criteria, and did read the article, but only suggested minor corrections that do not alter the intellectual content, and most importantly the linguistic content, to still be an author.

This has linguistic implications as the text of an article could contain not only group but also and maybe even mainly, individual linguistic style-makers. No relevant empirical studies have been conducted on the most frequent writing strategies or practices followed by groups of authors to produce the text of an OSRA, especially in the field of health sciences (Ede 1990). Collaborative writing has been explored in the classroom environment (Brien and Fredericks 2020) and the professional context (López-Pellisa, Rotger, and Rodríguez-Gallego 2020) for writing assignments and technical writing, respectively; but researchers have not taken an interest in collaborative writing in the scientific research output context.

The practices vary according to the research community. Sometimes only the first and last author write the text of the article. The first author is usually the person who produces the first draft and is most probably the person directly working on the research problem, that is, the person carrying out the related experiments or research tasks. The last author is usually a senior researcher, most probably the Principal Investigator (PI), whose responsibility towards the publication is much more of scientific accountability and mentorship than of executive nature, supervising the research work proposed and carried out by other members of the team. In those cases, the other authors read and correct, or criticize. Whether that correction/revision is "critical for the intellectual content" of the article is difficult to assess from outside. In other cases, some sections are more frequently produced by the researcher who was directly involved in the implementation / operationalization of a given procedure or equipment. That is the case of the Methods & Materials section, which is written by whoever knows best or executed the procedure in question or understands the technicalities of equipment used.

During the text production of an OSRA, strategies of collective/group writing may vary from very participative approaches to practices that are centered on one person (the first author), meaning that in some cases, all authors will produce some writing – leaving some personal linguistic trace – and will also correct all parts of the OSRA text. In other cases, one researcher will be in charge of writing and will be leading the process of sending the manuscript back and forward, receiving feedback and contributions from the other authors.

In OSRAs, all four authorship functions described by Love (2002: 40-50), i.e., *precursory, executive, declarative,* and *revisionary authorship* may be present in the process of writing. Some functions, however, should be less represented. Since OSRAs introduce original research the 'amount' of precursory authorship should be significantly lower than what can be found, for example, in review articles which state they present a revision of earlier work on a given topic. Likewise, the declarative authorship should be less represented, since according to editors' recommendations, all authors must have contributed in a significant way. Not meeting this criterion may be interpreted as if the authorship is honorary or ghost, a practice not tolerated in scientific writing.

This leaves us with two main authorship options for OSRAs: executive and revisionary. Linguistic choices may, therefore, be influenced by two linguistic perspectives or levels of linguistic production: the individual level and the group level, in an iterative coexistence. In the context of this work, authorship profiling of linguistic background is regarded from the group or class perspective. Though many articles may have been written by only one person, the production process is assumed as one of collective creation.

In the health sciences, English is particularly relevant since most of the scientific dissemination media available to researchers use English as their main or only language of communication. The subfield of research of L1 influence belongs in the broader field of transfer studies within the research area of SLA. In that context, L1 influence had been addressed in several cross-linguistic and cross-cultural studies examining many different language pairs with regards to the use of rhetorical moves between L1 and non-L1 English writers.

For example, Xiao and Cao (2013) present a contrastive study of academic English of research articles authored by native and non-native (Chinese), using a multidimensional model to compare the frequency of 163 features organized in seven dimensions. The study concludes that there are significant differences between these groups concerning the involvement, commitment, and style, with native speakers using, for example, more discourse intensifiers than non-natives. Other studies have addressed the role of ESL or EFL, and particularly, ESP in the production of scientific discourse, comparing published writing produced by natives and non-natives English speakers to improve academic writing courses (ElMalik and Nesi 2008).

Nevertheless, studies approaching scientific writing as produced by L1 and non-LI English authors that focus on the linguistic idiosyncrasies of texts from the authorship profiling perspective for the detection of native language influence are far less frequent. Even rarer are studies addressing this topic within the field of health sciences. Gayle and Shimaoka (2017) addressed the lexico-grammatical differences found in scientific articles written in English by native and non-native authors in the field of pediatric oncology. The study analyzed a total of 22 859 abstracts with affiliations in 77 countries. Based on the differences exhibited by L1 and non-L1 users of English in the use of token sets ranging from one to four word phrases, the authors proposed a classification system named “Genuine Index” (GI) which can assess medical-scientific writing with regards to its compliance with standard English using the International English Language Testing System – IELTS’s aggregate scoring data for the skill writing as a reference; and classify the texts by native language of the authors. Reportedly, the results showed an extremely good overall performance of the GI model, with 93.3% of native language identifications being correct. However, the overall performance of the model in identifying abstracts whose authors are Japanese was rather poor, with only 26% being correctly identified as native Japanese authors. The study suggests that “editing and review processes [in reference to scientific journals] might partially obfuscate the L1 characteristics of non-native speaker Japanese authors”, i.e., native Japanese writing in English.

Works like that by Gayle and Shimaoka (2017) using similar computational approaches provide directions about the gaps to address in the NLI area. Unsuccessful results of native language identification in scientific texts do not necessarily mean that the non-native authors’ linguistic idiosyncrasies are obfuscated by editing, reviewing, or collective writing processes. It may also be the case that the variables used as style-markers of their written scientific writing have not been thoroughly examined. On the other hand, accounts of successful native language identification lack explanations about why certain variables work better than others in scientific writing.

This study focuses on NLID in scientific writing, specifically in the genre OSRA produced by European Portuguese and European Spanish native-speaking authors. The research design aims to balance quantitative and qualitative approaches to study linguistic variables as style markers in OSRA to test the hypothesis of a predominant language affecting a second language such as English used in a professional or academic context.

2.4. Chapter Summary

Thus far, I have discussed the different concepts and aspects that are relevant to authorship profiling in relation to language background and specifically concerning NLID applied to the OSRA in the field of health sciences.

In the first section, it was pointed out how authorship analysis has evolved in terms of the kind of texts it has been applied to, the methods used to carry the analyses, and the applications the results of this type of analysis can have. As can be seen, authorship analysis has been used in religious, literary, political, and forensic texts, and been applied to biblical works, novels, speeches, emails, police statements, short messages etc. Methodologically, authorship analysis has progressed from more qualitative approaches to quantitative, to a combination of both. Finally, applications of authorship analysis have expanded from the scholarly interests and objectives of authorship attribution of theological and literary works to forensic, marketing and cybersecurity applications. The development of authorship analysis as a field reflects the correspondence that it is possible to establish between linguistic and computational perspectives on its subfields and definitions. The development of authorship analysis as a field is also seen in the efforts of researchers with different scientific backgrounds working in authorship analysis and using mixed methods to attain better research results.

In the case of authorship profiling, the visibility of studies conducted by computer scientists suggests that research in this field has been more prolific than that conducted by linguists drawing on sociolinguistics and second language acquisition. However, some progress can be noticed in new names appearing in the field that may be seen as adapting a theoretical view based on linguistic theories and concepts from the field of SLA: NLID – native language influence detection and OLID – other language influence detection. Also, I described how sex and age are, so far, the most studied variables of authorship profiling and linguistic background is one of the least researched. Moreover, I argued that languages like Portuguese and Spanish are addressed less within NLI/NLID/OLID if compared, for example, to Russian or Chinese.

The second section of this chapter discussed language variation. I aimed at demonstrating variation at the level of the individual, i.e., idiolect, and at the group level, i.e., dialect, register, genre and style. A corpus linguistics approach to language variation was assumed as the most suited for the present research. In the second part of section two, several theoretical frameworks were examined that can assist the researcher in explaining variation, individual variation and variation in a context of language contact.

3. Methodology

The section describes the corpora collection process by addressing the corpora size and typology. There is an explanation of the criteria established for selecting the texts to be included in the corpora with details of the selection process, followed by a description of the corpora design. The pre-processing, preparation, and parsing of the texts are explained, followed by a discussion of the limitations to the corpora compilation. Finally, there is a general description of the five compiled corpora, with the name of the whole collection.

Section two details the study design and model of analysis, and describes the unified framework for investigating L1 Influence (Jarvis 2000, 2010), with an explanation of the reasons for choosing this method. Lastly, the section delineates the linguistic variables used for carrying out the comparisons for the determination of cross-linguistic influence in scientific text written in English by non-L1 authors; and the reasons for choosing those variables. The final section describes the operationalization of the research.

3.1. Corpora Design

L1 written discourse, in general, and academic texts, in particular, are well represented in corpora. Such a description is particularly evident in English. A query in the search engine Google using the following words and operators to look for corpora in English: ["academic" OR "scientific" AND "corpus" OR "corpora" AND "English"] returns 111,000,000 results. However, the same search for Spanish and Portuguese, i.e., ["academic" OR "scientific" AND "corpus" OR "corpora" AND "Spanish"/"academic" OR "scientific" AND "corpus" OR "corpora" AND "Portuguese"] returns 35,700,000 and 6,200,000 results, or 67.83% and 94.41%, respectively, less than English. Even if the same query is made using the keywords in Spanish and Portuguese, and the results – 7,090,000 and 2,700,000, respectively – are added to those obtained before using the keywords in English, without verifying possible overlapping, the decrease in comparison to English would be 61.50% and 92%, respectively[7].

[7] The query used was ["académico" OR "científico" AND "corpus" OR "corpora" OR "córpora" AND "español" OR "castellano"] and ["académico" OR “acadêmico” OR "científico" AND "corpus" OR "corpora" OR “córpora” AND "português"];

Despite the difference among the three languages, the results obtained, i.e., 111,000,000 for English, 42,790,000 for Spanish, and 8,900,000 for Portuguese, seem to provide plenty of research data. Likewise, there are also many corpora of L2 academic English. If a query is performed using the words and operators ["academic" OR "scientific" AND "corpus" OR "corpora" AND "English" AND "L2"], 3,680,000 results are obtained[8].

It is not possible to verify the characteristics of all the corpora in the results obtained from such queries. However, several of these corpora were selected for inspection to learn if they would serve the objective of this research. For that purpose, the first results of each query (up to ten per language and language variety) were scanned to check:

- if the corpora contained texts in the language varieties addressed in this study;
- if the corpora contained only OSRAs, or if there was some form of filtering for OSRAs;
- if the corpora texts were published after 2015 or if there was some form of filtering of that information;
- if access to the corpora was unrestricted;
- and if the corpora were annotated with PoS tags, parsing and metadata.

The corpora inspected have at least one characteristic that prevents their use in the context of this research. The most frequent characteristics that made these corpora unsuitable for this research were concerned with the type of text or genre, the disciplinary areas addressed, and the costs associated with access.

There were corpora containing research articles in health-related fields but restricted to parts of the articles. An example of this is the **GENIA** corpus (https://www.clarin.eu/resource-families/corpora-academic-texts) which contains texts from research papers in the field of Biomedicine but restricted to abstracts. Other corpora contain academic text but include many different genres and sometimes L1 authors at different levels of writing proficiency. This is the case of the **BAWE** – British Academic Written English corpus (https://www.sketchengine.eu/british-academic-written-english-corpus/) and the **CAEC**: Cambridge Academic English Corpus (https://www.sketchengine.eu/cambridge-academic-english-corpus/). The first contains academic work in Arts and Humanities, Social Sciences, Life Sciences, and Physical Sciences produced at universities in the UK; and the latter is a

[8] The queries were performed on September 23, 2020 in Google Chrome UI displayed in Portuguese (Portugal). The same queries may return different results over time and if the settings of the search engine are changed.

collection of both written and transcribed spoken academic texts containing many different genres (e.g., lectures, journals, essays), also produced by authors at different levels of writing proficiency, and from institutions in the UK and the USA.

A good choice for Portuguese would have been the **CoPEP**: Corpus of Portuguese from Academic Journals (https://www.sketchengine.eu/copep-corpus-of-portuguese-from-academic-journals/). The corpus contains research articles from the field of health sciences written in Brazilian and European Portuguese, and it can be consulted upon subscription to Sketch Engine. However, because it was built in 2018, the corpora for the present research were already completed by the time the CoPEP was made available to users and researchers. In the case of Spanish, no corpora were found that were accessible and contained OSRAs or parts of OSRAs. Some appealing projects are the **CELiST** (Corpus of English Life Sciences Texts), a sub-corpus of the Coruña Corpus of English Scientific Writing (CC), and the CC itself. However, these are both corpora under construction.

Therefore, the present research work was carried out using its own corpora. As Maia (1997: 3) explained, there are many reasons that justify the need for making our own corpus. In this study, the rationale concerned the need for corpora of a specific genre and the comparability of the texts. The need for creating our own corpora is justified by the lack of available and ready-to-use accessible and annotated corpora of the genre OSRA in the target language varieties (European Portuguese, European Spanish, British English, non-L1 English produced by L1 European Portuguese and non-L1 English produced by L1 European Spanish).

3.1.1. Corpora Type and Size

The proposed analyses were based on five small specialized corpora purposely built for this study as defined by Flowerdew (2004), Sinclair (2004), and Koester (2010) according to the following parameters:

- the corpora serve a specific purpose, i.e., to investigate a set of lexico-semantic and syntactic features to examine their potential to function as indicators of NLID;
- the corpora consist of only one textual genre, i.e., the OSRA;
- the corpora are of a specific discourse type. i.e., mainly argumentative;

- the corpora are of a specific variety of English (academic) and
- the corpora are of a specific study field, i.e., health sciences.

All the texts sorted and selected are OSRAs published in the field of health sciences.

The option of building small corpora was based on three main arguments.

- The first was the time frame available to complete the collection of the texts. Time frame refers to the actual amount of time available within the present doctoral program to collect the texts, i.e., approximately 6-8 hours a week from 2015 to 2018 or roughly one full-time equivalent year.
- The second argument refers to the pre-processing of the texts, which was extremely time-consuming due to the many steps required.
- The third argument refers to the fact that these are specialized texts of only one specific genre. Within the lexico-semantic dimension, specialized corpora, unlike general corpora, usually have a smaller number of different words, i.e., lower vocabulary diversity. This is because specialized texts feature specialized vocabulary that cannot be replaced with synonyms or equivalent expressions. Still, the frequency of function words, such as articles and pronouns, remains proportionally similar to what is found in general language corpora (Sinclair 2004). This allows the researcher to obtain relevant information from a smaller amount of total words because both specialized and general vocabulary are well represented in the frequency lists of any given specialized corpus (Weisser 2016: 31; Sinclair 2004). Other variables, such as morpho-syntactic or discursive variables, may be affected by the specialized corpora's homogeneity in the same way (Sinclair 2004).

The selection of all the OSRAs in the corpora was based on the criteria outlined according to recommendations from both health sciences researchers consulted in the work context and corpus linguistics scholars (Koester 2010; Sinclair 2004; Biber et al. 1998; Biber 1993; Flowerdew 2004).

3.1.2. Selection Criteria

Any selection process entails following some criteria to decide whether a given element "belongs" or not to the future collection. This working principle is not different when it comes to corpora (Sinclair 2004). As happens with almost everything concerning language, corpora can be of spoken language and written language. In the case of this study, only criteria that apply to written corpora are considered.

Sinclair (2004: 22) defines a corpus as a "collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research". Atkins, Clear, and Ostler (1992: 1), in turn, define a corpus as a "subset of an ETL (electronic text library), built according to explicit design criteria for a specific purpose". The present study aims to examine L1 influence in scientific writing produced in English by non-L1 authors. This research focusses on a specific time period and on five language varieties, three L1 and two non-L1. The corpora, therefore, meet criteria in relation to 1) the scientific genre; 2) time period; 3) copyright of the texts; 4) quality of the texts; and 5) authors' L1 and non-L1. The next subsections refer to the selection criteria used to build the five corpora.

3.1.2.1. OSRAs - Genre

The genre chosen here to represent scientific writing is the OSRA, restricted to health sciences.

An OSRA is a scientific research article disclosing new research, i.e., research carried out by the authors that adds to the existing body of knowledge in a given field. OSRAs present one or more tested hypotheses, and experimental research results of such testing are described and discussed. Therefore, OSRAs are considered to be primary sources of knowledge.

The vast majority of scientific journals provide clear definitions of what they accept for publication as an OSRA. Such definitions are usually found in journals' Authors Guidelines/Instructions under the section "article types". For example, the British Journal of Medicine and Medical Research defines OSRAs as "papers that include original empirical data that have not been published anywhere earlier (except as an abstract). Null/negative findings and replication/refutation findings are also welcome."[9]

Journals also usually establish certain limitations concerning the length of OSRAs in words, pages, or characters, the number of figures and/or tables allowed, the citation style to be used, and even the number of references, based on their pertinence for the topic. Such restraints are usually also described in the Authors' Guidelines/Instructions.

The OSRA can be considered a restricted or specialized register of scientific writing. This restriction can be described in terms of its association to research processes (Swales 1990: 179-201) as opposed to genres such as essays, lectures, or seminars that would generally be associated with learning processes within academia.

The choice of working with the OSRA responds to reasons of familiarity with the genre, gained in a professional context. My investigative interest in its linguistic layout is related to my need to gain more knowledge on this genre given its relevance to its community of practice, i.e., the researchers, whether they are professors, medical doctors, technicians, undergraduate and postgraduate students, or any other health or health-related professional. Swales (1990: 177) describes the research article as playing a central role in the research-processes from which many other genres are derived or with which they connect. Some of these are the abstract, oral and poster presentations, theses and dissertations, and grant proposals which precede, succeed, or coexist with the research article.

3.1.2.2. OSRAs - Time Period

The corpora texts were initially intended to cover a ten-year period, beginning from the most recent article accessible until the target number of OSRAs per corpus, i.e., 65 OSRAs, was achieved. However, because of constraints that arose during the collection process described in 3.1.4, the time period was extended to twelve years. Therefore, the collection of OSRAs for these corpora covers articles from 2006 to 2018. The time period per corpus is shown in Table 5 below.

[9] https://www.sciencedomain.org/journal/12/authors-instruction

CORPUS	Years	**Time Period** Corpora average = 10,20 years
L1 Portuguese Corpus (**PT-EU**)	2007-2017	11 years
L1 Spanish Corpus (**ES-EU**)	2007-2017	11 years
L1 English Corpus (**EN-GB**)	2011-2018	8 years
Non-L1 English by L1 PT^{EU} Corpus (**EN-PT^{EU}**)	2006-2017	12 years
Non-L1 English by L1 ES^{EU} Corpus (**EN-ES^{EU}**)	2009-2017	9 years

Table 5 – Time period covered per corpus

The corpora are deemed synchronic and closed (Atkins, Clear, and Ostler 1992: 6) for the purpose of this research. However, the corpora can be extended by adding more OSRAs from previous and/or subsequent years, diversifying the kind of research articles to include, for example, clinical cases, meta-analyses, reviews, short papers, or other; or even broadening the scientific genres to incorporate dissertations, thesis, reports, etc.

3.1.2.3. OSRAs - Access Type

An aspect that must be considered when building corpora refers to the copyright of the texts included in the collection (Atkins, Clear, and Ostler 1992: 4; Weisser 2016: 32-33). The legislation in relation to copyright differs from country to country, but in the European Union in general, the standard for written work that has already been published is that the copyright lasts up to seventy years after the author's death, unless the copyright is inherited by relatives or others (Weisser 2016: 33).

However, most of the countries with copyright legislation, including those within the EU, also recognize a concept called 'fair use', which grants the use of parts of copyrighted work for research, educational or other non-commercial/non-profit purposes without the expressed permission of the copyright holder (Davies 2002). In the context of corpora compilation, "the copyright law that matters is the law of the country from which the corpus materials are distributed, NOT the country where the original texts were created OR the country from which end users access the material" Davies (2002: - online "Legal aspects of corpora compiling"). Although the corpora compiled for this research are not intended for general distribution or access, it can be said that at least a limited number of researchers may eventually get access to it. For that reason, it is the Portuguese law that is taken into consideration for this matter.

In Portugal, copyright is established by the Decree-Law number 63/85 – *Código do Direito de Autor e dos Direitos Conexos*. Chapter II of that decree regulates what in Portuguese is called "*utilização livre e permitida*"[10] or its equivalent "fair use", which in its articles f) and o) foresees the use of copyrighted material for educational and research purposes to the extent that no direct or indirect economic or commercial advantage is obtained[11]. Nonetheless, to avoid any copyright infringement issues or any further legal constraints as to the usage of the texts in corpora for research purposes, it was decided that mostly open access OSRAs would be selected for the corpora. Also, the intention in choosing open access material was to avoid expense in purchasing articles from scientific journals and support the open access policy encouraged by the European Union and by the Portuguese national policy in relation to the dissemination of scientific investigation in all areas.

Open access makes available scholarly content like scientific research articles, dissertations, conference proceedings etc. made available as online, with no cost for readers, free of restrictions concerning copyright and licensing, and free of impediments associated with access , i.e., not needing, for example, any specific software or user profile to access the content (Johnson, Watkinson, and Mabe 2018: 97).

Open access research articles are usually published under creative commons licenses that allow using the material under certain terms (Johnson, Watkinson, and Mabe 2018: 100). A standard license of open access, as described in the webpage of creative common licenses (https://creativecommons.org/licenses/), is one known as Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0). According to the description in the license https://creativecommons.org/licenses/by-nc-nd/4.0/, the first acronym (BY) regards attribution, which establishes that whoever uses the material acknowledges the authorship,

[10] It can be translated as 'free and permitted use' or 'open and admissible use'

[11] Article f) and o) of the Decree-Law 63/85, in Portuguese "f) A reprodução, distribuição e disponibilização pública, para fins de ensino e educação, de partes de uma obra publicada, contanto que se destinem exclusivamente aos objectivos do ensino nesses estabelecimentos e não tenham por objectivo a obtenção de uma vantagem económica ou comercial, directa ou indirecta"; "o) A comunicação ou colocação à disposição do público, para efeitos de investigação ou estudos pessoais, a membros individuais do público por terminais destinados para o efeito nas instalações de bibliotecas, museus, arquivos públicos e escolas, de obras protegidas não sujeitas a condições de compra ou licenciamento, e que integrem as suas colecções ou acervos de bens;"

provides a link to the license, declares if there were changes made to the content or the form, and complies with this term to leave it clear that the licensor is not endorsing his/her use. The second term (NC) refers to noncommercial use, which says that whatever use is given to the material under this license, the user must not profit/have commercial purposes. Lastly, the third acronym (ND) refers to noderivatives, which says that if the user "remixes, transforms, or builds upon the material", he/she is not allowed to distribute the modified material (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Most of the OSRAs collected for the corpora of this study are published in open access. Others were collected based on free access, such as "editor's choice" of the journals. A small number of the corpora OSRAs were accessed by institutional subscription of the University of Porto and used under the Portuguese legal rights mentioned above. Sections 3.1.3, 3.1.4, and 3.1.5 provide further explanation of the reasons for choosing OSRAs that are not published in open access.

3.1.2.4. OSRAs - Quality

During the planning stage of the corpora compilation process, several health sciences and applied linguistics researchers were consulted on the characteristics of the OSRAs to be included in the collection.

One aspect most of the researchers identified as being very important concern the quality of the texts in terms of language and writing. The researchers recommended that a certain level of language quality was ensured because it could decrease the possibility of significant results of cross-linguistic variation arising from basic editing and writing mistakes or errors contained in the texts.

Whereas there may be different criteria that can be used to define quality, in the health sciences, the quality of OSRAs is frequently associated with the rank of the journals according to peer-reviewed scientific journal indexes such as the SCImago (https://www.scimagojr.com/index.php) which considered criteria like impact factor to classify the journals. Therefore, the OSRAs included in the collection were chosen from journals of recognized quality. The quality of the texts was ensured by choosing indexed peer-reviewed journals originally edited in the respective language varieties addressed in this study, i.e., EN-GB/ES-EU/PT-EU.

3.1.2.5. OSRAs - Native Language of Authors

In a study comparing OSRAs produced by native (L1) and non-native (non-L1) authors, the ideal procedures would have been to survey the authors about their native and non-native languages or apply language tests to learn their language levels. Both procedures would have implied contacting at least 325 authors, i.e., one corresponding author per OSRA in the corpora. Moreover, if information concerning the other authors' participation in the OSRA writing process was confirmed, more surveys would also have to be requested to learn their native and non-native languages. This would have increased the number of surveys or language tests to 2520, which is the total amount of authors represented in the corpora (372 in the PT-EU OSRAs; 463 in the ES-EU OSRAs; 513 in the EN-GB OSRAs; 598 in the EN-PT^{EU} OSRAs; and 574 in the EN-ES^{EU} OSRAs).

Given the difficulties that such actions would have involved, learning first-hand the L1/non-L1 languages of the corpora authors was considered impractical. Therefore, the "nativeness" of the authors concerning each language variety addressed here was decided based on compliance with several premises designed to minimize as much as possible the likelihood of assuming incorrect information. As such, the "nativeness" of the OSRA texts contained in the corpora was ascertained by the presumed origin of the authors, which, in turn, was delimited by the following criteria:

- At least the first two authors' and the last author' names should match Portuguese, Spanish, and English typical names (e.g., *Rui Ferreira; Rafael Sáez-Jiménez; Malcolm J. Jackson*);
- Authors' affiliation(s) should refer to addresses in Portugal (PT), Spain (ES), or The United Kingdom (GB), according to the corpus to compile (e.g., *Faculdade de Medicina da Universidade do Porto, Portugal*; *Centro de Salud Presentación Sabio C/Alonso Cano, 8 C.P. 28933. Móstoles – Madrid, España; Oxford Stone Group, Department of Urology, Nuffield Department of Surgical Sciences, The Churchill Hospital, Oxford*);
- If funding sources exist, they should preferably be national funding sources (e.g., *Alto Comissariado para a Saúde; Instituto de Salud Carlos III; UK Biotechnology and Biological Sciences Research Council*).

All three criteria must be met for an OSRA to be considered as authored by a native speaker of the respective language variety. This method is a stricter version of the one proposed and used by Wood (2001: 78-79) to decide on the native language of authors of research articles of the scientific journals Nature and Science. The restriction concerns the inclusion of the third criterion for funding institutions which is not considered in that publication. Similar versions of selection criteria have also been used by authors such as Yakhontova (2006), Pan, Reppen, and Biber (2016) Noorizadeh-Honami and Chalak (2018), and Kafes (2018) to classify authors of research articles as native or non-native users of English.

The presumed non-L1 authors of the corpora's OSRAs are assumed to be either EFL or ESL users. In other words, non-L1 authors were considered to belong to the expanding circle and, possibly, to the outer or extended circle (Kachru 1985). The inclusion in the expanding circle seems obvious since the authors are presumed to be native Portuguese and native Spanish. Portugal and Spain are two countries that do not have a history of colonial ties with the United Kingdom, a condition described by Kachru (1985: 13) as characteristic of the expanding circle. The inclusion in the outer or extended circle is understood in this research as the ESL situation based on learning/practicing interactions, such as training programs, internships, scientific meetings, etc., involving native speakers, i.e., scientists from the inner circle.

Taking Kachru’s definitions of English strata as a basis, the assumption is made that the authors of the OSRAs in these research corpora belong in different stages of the interlanguage continuum, having learned English in presumably different contexts and also using the language for different sociolinguistic and functional purposes. L1 users are exposed to the language from birth and use it extensively and in a wide variety of situations. Non-L1 users learn the language later in life and use it in more restricted contexts (Pérez-Llantada 2012). However, it is also assumed that they are all located on the same side of the continuum, i.e., native users at one extreme of the continuum as native speakers of the target language the non-L1 users have learned, and non-L1 users at some point close to the extreme as advanced users of the target language.

3.1.3. Pre-processing, preparing, and parsing the texts

The OSRAs selected for each corpus were pre-processed, prepared, and parsed. Each OSRA was downloaded from the online version of the journal where it was published and saved in PDF format. When available, texts were copied directly from the HTML version of the article. Otherwise, the texts were manually extracted from the PDF version into a Word document and corrected to maintain its original format.

Each OSRA was processed to include only the following parts of the texts in the corpora: title, introduction, results, discussion, and conclusions sections. The "abstract" section was not considered because several other researchers have already examined this part of the research article (RA) whether it is of the original type of RA, the review type, or others (Hu and Cao 2011; Salager-Meyer 1992; McKnight and Srinivasan 2003; Anderson and Maclean 1997; Busch-Lauer 1995). Moreover, the "abstract" section has been considered a separate genre that is "easy to recognize" and one that "distills" the content of an article (Swales 1990: 179) and whose "purpose, rhetorical construction and persuasive intent are all distinct from the article itself" (Hyland 2004: 64).

Also, the "Methods and Materials" section was disregarded because this section would be less likely to contain original linguistic style-markers produced by the author(s). Usually, authors of this research area are advised to produce a "Methods and Materials" section that is descriptive of the steps taken to conduct the research. It is frequent to find recommendations to organize the text in clearly separated sections and the specific use of the past tense, as well as advice in relation to the presentation of the research protocol, the names of equipment brands, software versions, name of animal models used in experimental studies, name of services or product suppliers, and even references to national and international legislation researchers are required to abide by (Michel and Ceelen 2007; Kallet 2004).

After selecting the OSRA sections to be included in the corpora, each article was revised to find and remove from the text symbols (e.g., β; α; ∑; ±; ®), numeric and bibliographic references, abbreviations (e.g., Fig.; e.g.; i.e.; vs.; al.; etc.) and equipment names or product trademarks that could interfere with the format of the files and the subsequent linguistic analysis. After that, each text was copied and saved as a text file in Unicode codification for English and UTF-8 for Portuguese and Spanish to be used later with other processing software.

To avoid the use of the extended original titles of the articles, each OSRA was coded with the abbreviation of language name according to the ISO 639-1 Code, plus the indication of the variety, plus the abbreviation of the article type, plus a consecutive number from 1 to 65. For example, for the first OSRA in the corpus containing texts written in European Portuguese by L1 authors, the code is PT-EU_OSRA_001. Figure 1 shows the workflow of the pre-processing and preparation of the files for parsing.

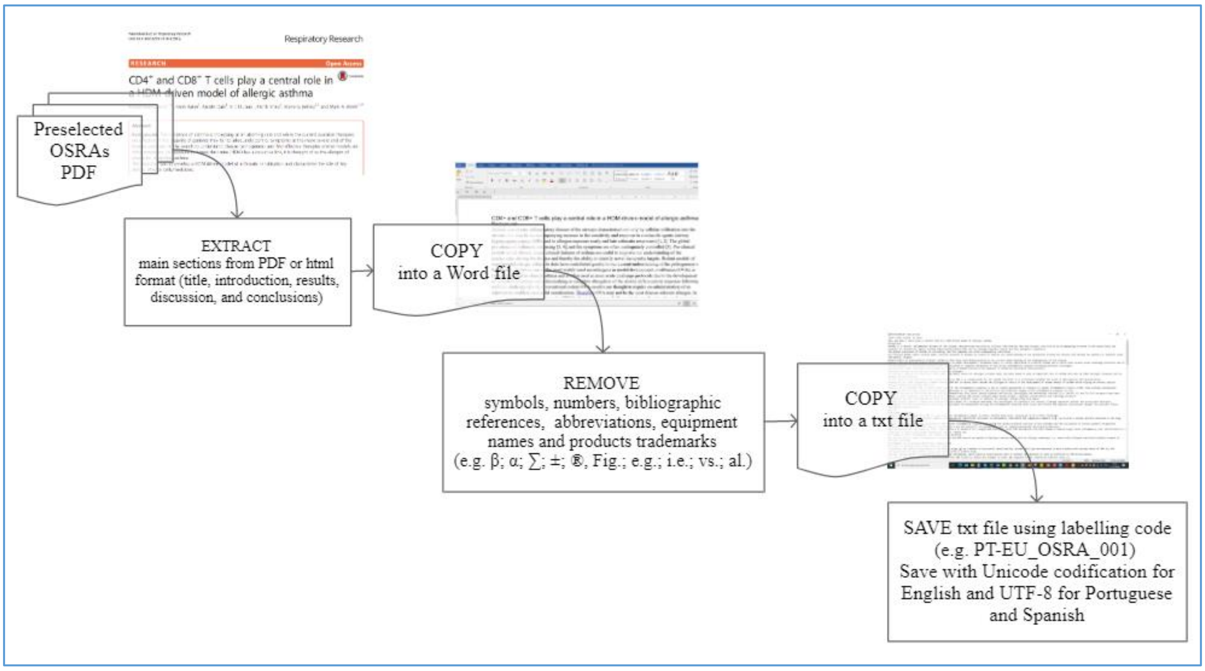

Respiratory Research
RESEARCH
Open Access
CD4+ and CD8+ T cells play a central role in a HDM-driven model of allergic asthma
Preselected OSRAs PDF
EXTRACT
main sections from PDF or html format (title, introduction, results, discussion, and conclusions)
COPY
into a Word file
REMOVE
symbols, numbers, bibliographic references, abbreviations, equipment names and products trademarks
(e.g. β; α; ∑; ±; ®, Fig.; e.g.; i.e.; vs.; al.)
COPY
into a txt file
SAVE txt file using labelling code
(e.g. PT-EU_OSRA_001)
Save with Unicode codification for English and UTF-8 for Portuguese and Spanish

Figure 1 – Corpora pre-processing and preparation stages before VISL parsing

The files were parsed using the Internet-based software Visual Interactive Syntax Learning – VISL [https://visl.sdu.dk/]. All texts written in Portuguese were parsed with the multi-level Constraint Grammar parser PALAVRAS (Bick 2000, 2014). The Spanish texts were parsed using the system HIS-PALAVRAS (Bick 2006), and the English texts were parsed with the Constraint Grammar system EngGram (Bick 2012, 2010). To generate the computer analysis each .txt file was uploaded in the interface of the corresponding language choosing the options Language > Sentence analysis > Machine Analysis > Upload interface of the webpage and filling in the mandatory information requested by the system, i.e., name, email, copyright status of the text, and chosen parser (full analysis). Each file obtained was saved with the same name of the .txt file that originated the parsed file adding the label ‘parsed’, i.e., ‘EN-GB_OSRA_001’ generated ‘parsed_EN-GB_OSRAs_001’.

The program for lexical analysis Wordsmith (Scott 2018b) was used to produce wordlists, statistics, and concordances according to the different variables of the study. All statistical analyses were performed using the Statistical Package for Social Sciences IBM SPSS Statistics Version 25 for Windows under the license of the University of Porto (UPorto) available at https://atlas.up.pt/Software/UPORTO/SPSS/, using the author’s student UPorto login.

The first corpus built was the PT-EU corpus, followed by the ES-EU corpus. Later, the EN-GB corpus was compiled, and the EN-PT^{EU} and the EN-ES^{EU} corpora were built last. The order of the compilations was decided by the availability of OSRAs originally written in PT-EU and ES-EU. Innumerable open access OSRAs were available in English as this is a very common language of publication in the health sciences field. However, this was not the case for Portuguese and Spanish since many of the indexed peer-review scientific journals edited in Portugal and Spain are not always published in open access, or they do not publish the whole article in their native languages (only the abstracts).

The PT-EU compilation began at the index of Portuguese Medical Journals – Indexrmp (http://www.indexrmp.com/), a collection of 175 journals edited in Portugal. Only peer-reviewed and indexed journals from this index were explored to obtain the OSRAs necessary for the study. The ES-EU corpus started with the Spanish Elsevier indexed journals and progressed from there to any available OSRA that fulfilled the selection criteria. The EN-GB corpus started with the Wiley Online Library and developed according to the OSRAs found that had all the characteristics of the criteria established.

Three different corpora were built that contain 65 OSRAs each. According to the corpora design, the first corpus is written in European Portuguese (PT-EU) by presumably L1 Portuguese authors. The second is written in European Spanish by presumably L1 Spanish authors, and the third is written in British English by presumably L1 English authors. All the OSRAs are published in peer-reviewed scientific journals indexed in different databases, such as Elsevier, SciELO, and Wiley Online.

The next step was to compile the second set of texts, the non-L1 corpora. As previously indicated, these corpora include OSRAs written in English by L1 Portuguese and L1 Spanish authors. The same selection criteria (See section 3.1.2) were followed for choosing the texts.

For these collections, more publishers and databases were available. The process began with the journal BJU International of the Wiley Online Library, and every attempt was made to associate OSRAs with the same or similar topics and keywords to those found in the L1 corpora. After the compilation was finished, the L2 corpora were made of two sets of 65 OSRAs: one containing OSRAs written in English by presumably L1 European Portuguese authors; the other containing OSRAs written in English by presumably L1 European Spanish authors. After all the OSRAs had been pre-processed, prepared, and converted into .txt, and parsed, the files were uploaded to WordSmith Tools 7.0 to obtain lists of general statistics. The resulting corpora and the total number of tokens per corpus are shown in Table 6.

CORPUS	TWC	TWCaE
PT-EU	194 705	143 786
ES-EU	238 198	162 731
EN-GB	246 166	171 170
EN-PT^{EU}	264 439	163 437
EN-ES^{EU}	299 082	184 279
TOTAL	**1 242 590**	**825 403**

Table 6 – Final Corpora Compiled (TWC – Total Word Count; TWCaE – Total Word Count after Edition)

3.1.4. Limitations to the compilation of the corpora

Many studies in discourse analysis of scientific text genres have focused on clinical or medical discourse (ElMalik and Nesi 2008; Galve 1998; Salager-Meyer 1994, 1990; Williams 1996). For this reason, the initial idea of this work was to build corpora containing life and health sciences OSRAs that would include texts presenting basic experimental research rather than clinical results or studies. However, this turned out to be difficult to undertake for the PT-EU corpus since basic research produced in Portugal in this area is mostly published directly in English in international peer-reviewed journals. The same problem arose with the ES-EU corpus. The competitiveness for career advancement and national and international funding is probably the reason for this situation. Publishing basic experimental research in Portuguese/Spanish would not provide the same visibility and exposure as in English.

Since the decision was taken to begin the corpora compilation with the Portuguese OSRAs (L1), the compilation process was inevitably drawn to a small number of indexed, peer-reviewed journals. Although the journals referred to publish in Portuguese, the articles have a more clinical/medical nature, which ultimately influenced the characteristics of the texts chosen to be part of the PT-EU corpus. As a consequence of the criteria used for corpora design, the first corpus that was built influenced all the other corpora, which are more of a clinical nature than the basic research type of article initially proposed for the research project.

3.1.5. General Description of the Corpora

In total, five corpora were compiled, parsed, and organized in two sets. The collection is called Comparative Corpora of Research Articles – CoRA and it is available in the open-access repository Zenodo upon request to the author (Sosa-Napolskij 2021, March 26). Set 1 refers to the L1 corpora. It contains OSRAs written in European Portuguese (PT-EU), European Spanish (ES-EU), and British English (EN-GB) by L1 authors of those language varieties. Set 2 refers to the non-L1 groups. It contains OSRAs written in English by L1 Portuguese and L1 Spanish authors.

Additionally, the corpus EN-GB by L1 authors, which is part of set 1, is also considered within set 2 since the EN-GB corpus is both an L1 corpus and a corpus written in English. In operational terms, it is part of both sets. For this reason, the L1 EN-GB corpus is called a pivot corpus. The use of the pivot language concept was inspired by that used in Machine Translation (Kay 1997; Cohn and Lapata 2007; Utiyama and Isahara 2007; Wu and Wang 2009), and the term pivot corpus was incorporated into the proposed design to express the central role played by the L1 EN-GB corpus used as a standard reference in the comparisons performed. Figure 2 below illustrates the organization of the corpora.

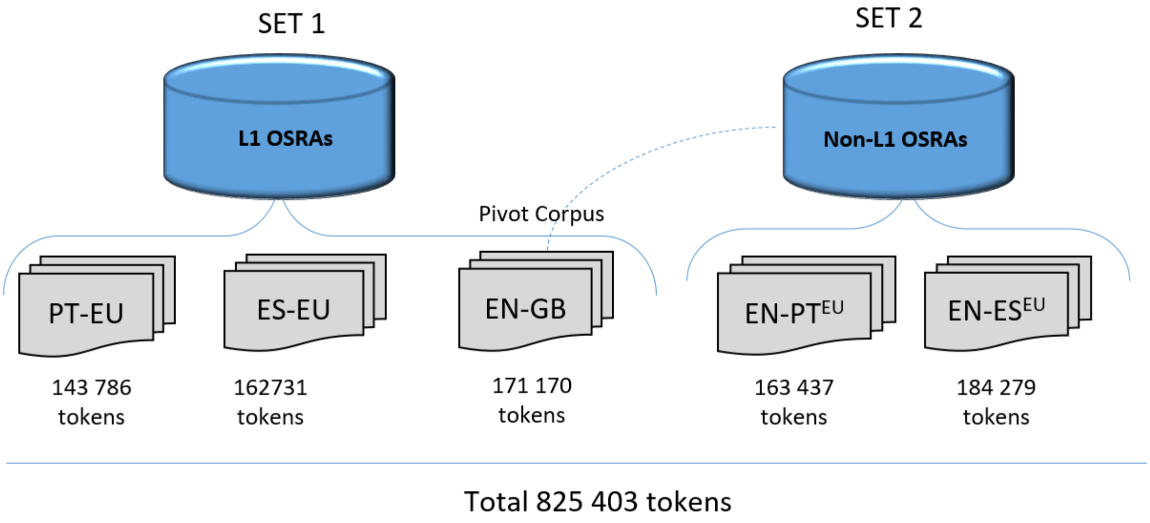


Figure 2 – Diagram of the corpora designed for the study – CoRA

Initially, only open-access OSRAs were chosen to be included in the CoRA. However, the open-access criterion introduced limitations in relation to the OSRAs available on a given topic. So, eventually, other types of publications were considered, i.e., freely available articles (e.g., Editor's choice) and OSRAs to which access was granted through the University of Porto's institutional subscriptions to the corresponding journals. As shown in Figure 3*,* of the 325 OSRAs included in the corpora, 52.62% (171) are published in open access; 29.54% (96) are freely available in the corresponding issue of the journal but are not declared as 'open access'; 17.23% (56) were obtained via the institutional subscription of the University of Porto, and less than 1% (2) corresponded to the Editor's choice.

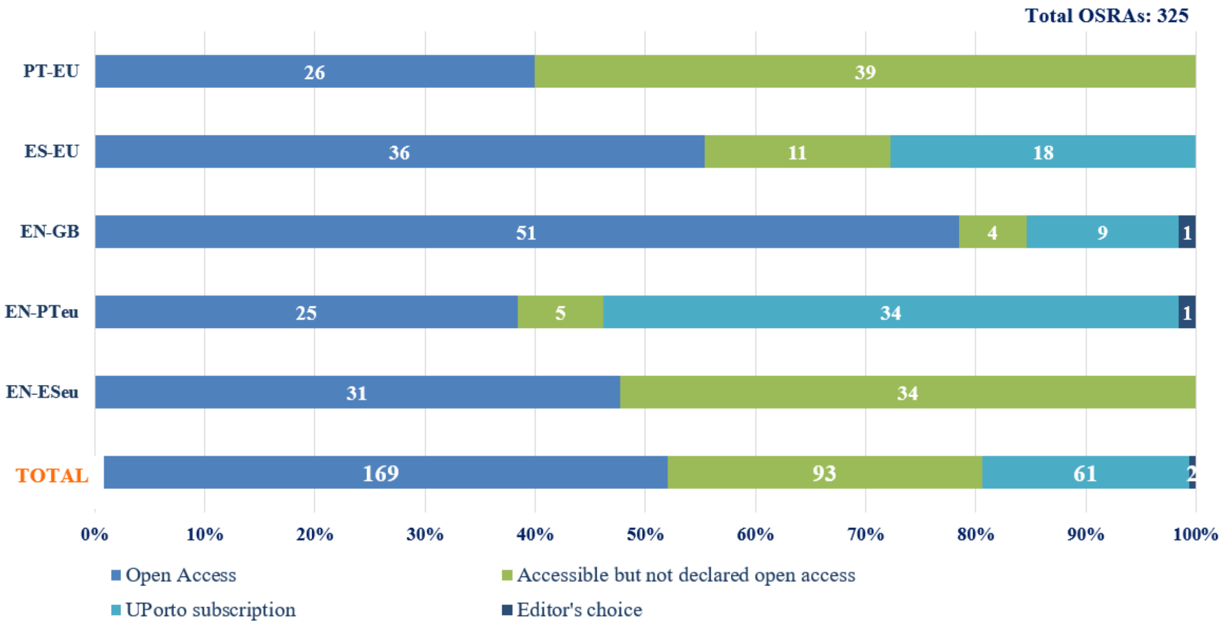
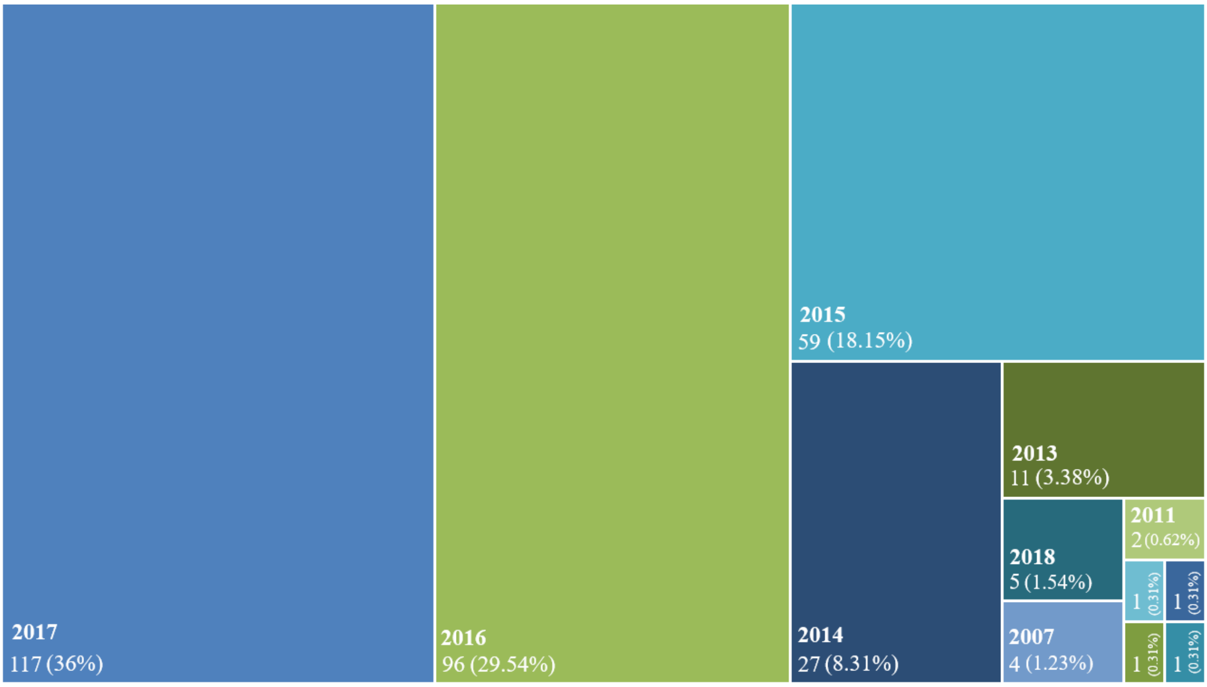


Figure 3 – Per corpus and total distribution of OSRAs according to access type

The aim of finding OSRAs published in recent years resulted in all of the articles being published within a global time frame of 13 years, from 2006 to 2018.

However, most of the articles chosen for the corpora (96.62%) were published within a smaller time frame of five years, as follows: 2013 (15), 2014 (27), 2015 (58), 2016 (101), and 2017 (113). The other seven years refer to the 3.38% of the OSRAs in the corpora, i.e., 2006 (1), 2007 (4), 2009 (1), 2011 (3), 2012 (1), and 2018 (1). Figure 4 below shows the distribution of the years of publications in the CoRA. For the detailed distribution of the years of publications by corpus, see Annex I.



CORPUS	2017	2016	2015	2014	2013	2018	2007	2011	2012	2010	2009	2006	Total
PT-EU	24	19	10	5	3	0	3	0	0	1	0	0	65
ES-EU	27	25	6	5	1	0	1	0	0	0	0	0	65
EN-GB	22	19	10	3	4	5	0	1	1	0	0	0	65
EN-PT^{EU}	15	17	21	10	1	0	0	0	0	0	0	1	65
EN-ES^{EU}	29	16	12	4	2	0	0	1	0	0	1	0	65
Total	117	96	59	27	11	5	4	2	1	1	1	1	325

Figure 4 –OSRAs in the CoRA, presented per year with a table of distribution per corpus

The average number of authors (NoAs) in the CoRA is 15 (Figure 5). OSRAs with the lowest NoAs were written by one author (n=3; 2 in the PT-EU corpus and 1 in the ES-EU corpus). OSRAs with the highest NoAs were written by more than 30 authors (n=3). That is, two OSRAs with 31 and 37 authors are part of the EN-ES^{EU} corpus, and one OSRA with 32 authors is in the ES-EU corpus. Despite these extreme numbers, the average and median values of NoAs within each corpus and between the corpora are fairly similar. However, as can be verified in Figure 5 below, the NoAs of the OSRAs within the English corpora is higher than the NoAs of the OSRAs in the Portuguese and the Spanish corpora. The PT-EU corpus has the smallest average NoAs, while the EN-ES^{EU} corpus has the largest average NoAs in the CoRA.

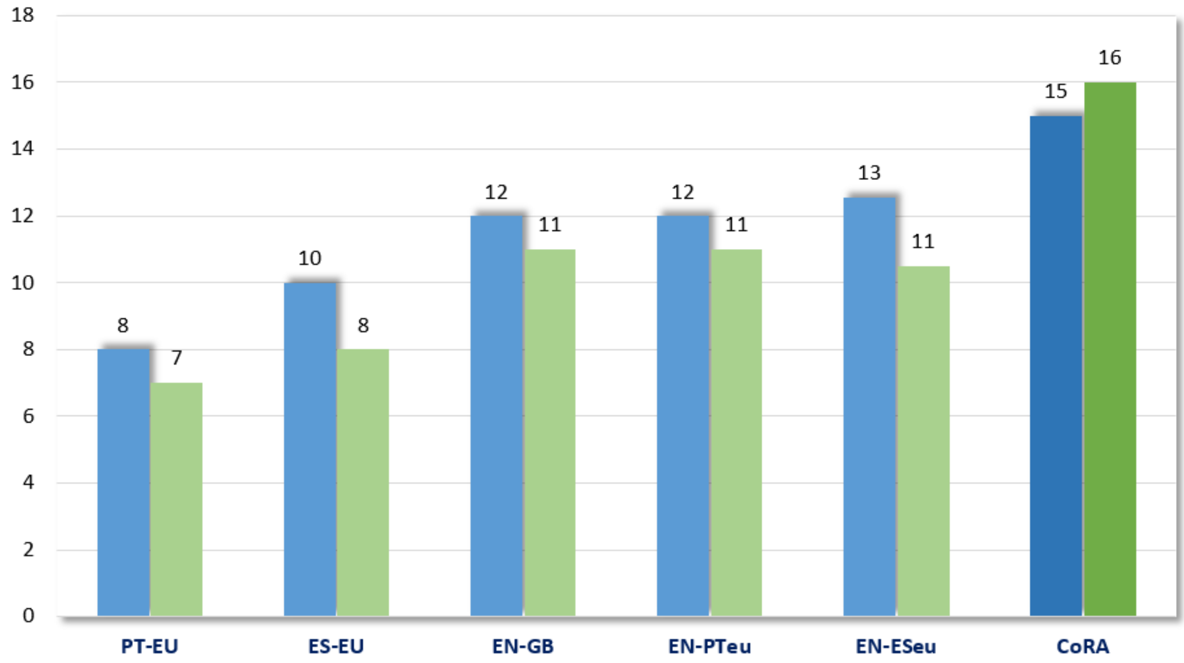


Figure 5 – Average and median values of NoAs per corpus and in total within CoRA

Overall, the texts used in the corpora were extracted from 135 different peer-review indexed scientific journals, of which 15 provided texts for the PT-EU corpus, 16 for the ES-EU corpus, 24 for the EN-GB corpus, 46 for the EN-PT^{EU} corpus, and 34 for the EN-ES^{EU}. Annex II and III present the complete list of the journals from which the corpora texts were extracted and a word cloud representing the main topics in the CoRA according to the OSRAs keywords. The lists of the OSRAs included in each corpus within the CoRA are provided in Appendixes 1 to 5.

3.2. Study Design and Model of Analysis

This study is based on cross-linguistic analyses of class variables found in the CoRA. In this context, class refers to variables "identified in the language or dialect of groups of writers" – i.e., the five corpora compiled– and not to features "observed and described in the idiolect of a single writer" (McMenamin 2002: 130).

The cross-linguistic analyses are performed sequentially. First, quantitative techniques are applied. Then, the results obtained from quantitative analyses are used to inform qualitative approaches to examine recurrent patterns (McMenamin 2002) and distributions that may distinguish the writing style of the OSRAs authors in relation to their respective L1 .

The analyses are based on within-group and between-group text-based corpus techniques that examine a group of selected relevant linguistic and non-linguistic variables. Moreover, the quantitative analyses used the comparison-based approach to investigating L1 influence developed by Jarvis (2000, 2010).

3.2.1. Comparison-Based Approach to Investigating L1 Influence

Two of the main approaches used to investigate L1 influence are the detection-based approach and the comparison-based approach to transfer research (Jarvis 2010).

The first aims at recognizing the language-background (usually L1) of a given language user based on the linguistic patterns he displays when using another language (usually a non-L1) (Jarvis 2000: 171). The detection-based approach is a response to finding alternative forms for investigating L1 influence (Jarvis 2010). This approach is closely related to authorship attribution as addressed in computational linguistics, drawing on the premise of accuracy to determine if there is L1 influence in a given text by a given author (Jarvis 2012b: 20). This approach uses computer software to automatically detect if a certain piece of writing in a given language reflects its author(s) influences from another language. This detection is done by providing the software with linguistic information of the languages involved so that the software can identify patterns and "predict which category [language] a particular text belongs to" (Jarvis 2010: 184) .

In turn, the comparison-based approach resorts to comparisons within and between languages to determine L1 influence (Jarvis 2010, 2000). It has been used in work by SLA scholars who would explicitly or implicitly compare language learners of English to native speakers to try to explain non-native errors when producing the language (Jarvis 2010: 172). From the need to explain language acquisition errors by non-native learners, many theoretical and methodological approaches were proposed to compare native and non-native language users to establish if there was L1 influence. Some of those theories are examined in section 2.2.4.2.

By using the comparison-based approach to study L1 influence in the scientific text, I seek to determine if there are linguistic features that may indicate that a certain text written in English by non-L1 authors contains elements from the authors' L1 and, if so, which features are significant. This approach is chosen because it allows one not only to determine if there is L1 influence but also to support the explanation of its nature, mechanisms, and context (Jarvis 2010: 182).

Jarvis (2000) revised the work developed by scholars working in transfer studies from approximately 1960 until 2000 and subsequent adjustments derived from theoretical and empirical research (Jarvis 2010; Jarvis and Pavlenko 2008; Jarvis 2012b). The first methodological framework covering at least the most important approaches to investigating language transfer using comparisons resulted from this examination. The name of the framework is Unified Framework for Investigating L1 Influence, and it is a model specially designed for examining language transfer, specifically L1 influence (Jarvis 2000, 2010; Jarvis and Pavlenko 2008).

Jarvis's (2000) initial investigation showed inconsistencies among results presented by researchers in relation to three elements. The first element referred to the concept of L1 influence, its definition, and extension. Some studies reported very low percentages of errors produced due to L1 influence, whereas others considered L1 influence the main source of "L1-induced errors". This difference showed that the definition of L1 influence differed from study to study (Jarvis 2000: 246). The second element considered how proficiency in a non-L1 affects L1 influence. In this regard, some studies found a direct correlation between L1 influence and proficiency in a non-L1. Other studies demonstrated the exact opposite, i.e., that the relation was inversely proportional, and the more proficient a learner was, the less

L1 influence was observed. Lastly, some studies found no relation between proficiency levels in a non-L1 and L1 influence (Jarvis 2000: 247). Finally, the third element referred to how differences and similarities between an L1 and a non-L1 of an individual affected L1 influence. The assumption in this regards had fluctuated from a position where it was believed that the more different an L1 and a non-L1 of an individual were, the more likely it was that L1 influence would occur; to a position where L1 influence was associated with similarities between an L1 and a non-L1; to an assumption that both differences and similarities between an L1 and a non-L1 of an individual explained L1 influence (Jarvis 2000: 248).

This presentation of "conflicting claims about the nature of L1 influence and its interaction with other factors" (Jarvis 2000: 248-49) was associated by this author with theoretical and methodological issues, which led to a proposal aimed at consolidating the approaches used in transfer studies based on the following three components:

- A definition of L1 influence that could accommodate the different transfer theories, i.e., a definition as impartial as possible (p. 249);
- A description of the "types of evidence" to take into account when studying L1 influence (p. 249);
- An account of the variables to be controlled to carry out a "rigorous investigation of transfer (p. 249).

With regards to the first component, Jarvis (2000: 252) offers a definition that is based on Odlin (1989) and Selinker (1972), and describes L1 influence as:

> "any instance of ***learner*** data where a statistically significant correlation (or probability-based relation) is shown to exist between some feature of ***learners' IL performance*** and their L1 background." (emphasis added)

As pointed out by the author, this definition is clear in informing the conditions under which L1 influence can be said to exist. That is, an evidence showing that there is L1 influence in a given case of non-L1 performance can only be accepted if there is a relation between the L1 background and the non-L1 performance that is statistically significant or linguistically relevant, or both.

This study assumes such a definition, but three aspects are adapted to the characteristics of the data obtained from the CoRA and the research questions. The first aspect refers to extending evidence of L1 influence from only 'statistically significant' data to also qualitative relevant results. Any variable found to be statistically significant will also be examined the linguistic patterns that can be associated with this significance. Therefore, this study seeks to combine both in order to report more comprehensive results and conclusions.

The second aspect refers to the individuals writing in a non-native language, i.e., English, whom we call non-L1 users or non-L1 authors, instead of learners because the authors of the CoRA are assumed to be advanced users of scientific English and not English learners in the strict sense of the word.

Lastly, the third aspect refers to using the term interlanguage (IL) to refer to the totality of the non-L1 linguistic output of non-L1 users. Because the authors in the CoRA are assumed to be advanced users, not all the output they produce in the non-L1 (English) is interlanguage. Some will reflect native-like competence. Therefore, interlanguage is not used to indicate non-L1 performance. Only the variables that can be statistically and linguistically shown to function as markers of the relation between the CoRA authors' L1 background and their non-L1 performance will be called interlanguage (IL). This position concerning the definition of L1 influence is in agreement with that proposed later by Jarvis (2010: 170), which is "the relationship between source-language [i.e., L1] group membership and target-language behavior [i.e., non-L1]".

Concerning the second component, i.e., the pieces of evidence or premises that must be considered when studying L1 influence, Jarvis (2000: 253-59; 2010: 170;84) proposes four. Table 7 below shows what these pieces of evidence are after adaptation to the present study.

Premise Type	Type of Evidence	Type of Comparison
group-based	I) Intra-L1-group homogeneity in English performance by non-L1 authors; II) Inter-L1-group heterogeneity in English performance by non-L1 authors;	I) Within-group II) Between-group
source-language-based	III) Cross-language congruity between non-L1 authors' L1 and their performance in English; IV) Intralingual contrast between English produced by L1 and non-L1 users.	III) Between-language IV) Within language

Table 7 – Types of evidence to demonstrate L1 influence (Jarvis 2000, 2010).

As mentioned in the L1 influence concept described above, all four types of evidence must be examined from a quantitative or a qualitative perspective. In this study, the first is used to describe the data and reduce the variables to those that entail significant differences between the groups. The qualitative examination seeks to explain such differences, evaluate them and understand the implications of the findings.

The quantitative examination considers the frequencies of the variables in analysis to compare the groups, and the qualitative examination looks at linguistic patterns (especially syntactic) associated with whatever variable is informed by the quantitative examination and contrasts the groups.

Quantitatively speaking, the first type of evidence (intra-L1-group homogeneity) refers to finding uniformity in the frequency a given variable is used by a group of authors who are all L1 users of the same language when writing in a non-L1 like English. From the qualitative point of view, intra-L1-group homogeneity refers to the consistency the users of an L1 (Portuguese, Spanish, English) exhibit in using linguistic patterns related to a given variable when writing in a non-L1 like English.

Quantitatively, the inter-L1-group heterogeneity refers to detecting statistically significant differences in the frequency of use of a given variable between EN non-L1 authors who are L1 users of different languages. Qualitatively, it refers to finding differences in how English is written by the non-L1 authors (Portuguese and Spanish).

The third type of evidence refers to not finding statistically significant differences in the frequency of use of a given variable between non-L1 authors when writing in English and when writing in their L1. Qualitatively, the same evidence seeks to identify consistency in how the study variables are used by non-L1 authors when writing in English and when writing in their L1.

Finally, the fourth type of evidence refers to finding statistically significant differences between the frequencies of the variables in texts produced by L1 and non-L1 authors when writing in English. Qualitatively, this evidence seeks to find a contrast between the linguistic patterns produced by these authors.

Taking as a reference the CoRA, this means that L1 Portuguese authors would be expected to have the same behavior in relation to a given variable when writing in English. Likewise, L1 Spanish authors would be expected to behave similarly in relation to a given variable writing in English. However, L1 Portuguese and L1 Spanish authors would be expected to differ from each other in the way they behave in relation to a given variable when writing in English. Moreover, L1 Portuguese/Spanish authors' behavior when writing in English would be consistent with what they would do in their respective L1s in relation to the same variable. Concurrently, these authors' behavior when writing in English should differ from what L1 English authors would be.

Reportedly, in an ideal situation, all four pieces of evidence must be found at statistically significant levels in order to claim L1 influence. Jarvis (2000: 255; 59) admits that, if after studying all four, two effects are found, L1 influence can be argued. However, he (2010: 181) highlights that the four types of evidence can be combined in as many as six pairs (i.e., 1-2; 1-3; 1-4; 2-3; 2-4; 3-4) and while not all combinations have to be examined:

> "one should be skeptical of any argument for the presence or absence of cross-linguistic influence that does not use these types of evidence in combination with one another in a way that, at the minimum, establishes (either quantitatively or qualitatively, or both) whether the target-language behavior in question is a group-based phenomenon and whether it is *also* a source-language-based phenomenon" (emphasis added). (Jarvis 2010: 181)

Thus, in the light of this reasoning, the minimum number of comparisons required to establish group-based and source-language-based L1 influence can be said to be three. If only comparisons 1-2 are carried out, then only group-based results will be obtained. If only the comparisons 3-4 are carried out, then only source-language-based results will be obtained. In both cases, only one type of influence can be claimed, either group-based or source-language-based.

If only the comparisons 1-3 and 2-3 are performed, then the results obtained will only account for similarities and/or differences within and between the L1 of non-L1-authors and their performance in English. These combinations would leave the comparison with native speakers of English out of the equation. In this case, it will not be possible to verify if a given feature observed in non-L1 English authors when writing in English and when writing in their L1 is also common in texts produced by L1 English authors. This will make it difficult to affirm that a certain variation in the English produced by non-L1s is infrequent in L1 English authors and so that this variation is due to L1 influence.

Likewise, if only the combinations 1-4 and 2-4 are carried out, the results obtained will only account for similarities and/or differences within and between the English produced by L1 and non-L1 authors of more than one L1 group. These combinations would leave the comparisons with the L1 of the non-L1 English authors outside of the equation, and in such cases, it would not be possible to verify if the features found in the English produced by non-L1 authors are similar to what these authors would do in their L1, and so claim L1 influence (Jarvis 2010: 182).

In summary, besides any of the combinations mentioned above, at least one more type of comparison is needed to discern L1 influence. In this thesis, all four types of comparisons are examined to obtain the strongest evidence possible to support L1 influence in OSRAs. If at least one group-based (1 or 2) and one source-language-based (3 or 4) L1 effects are found, L1 influence can be inferred.

Finally, Jarvis (2000: 260-61) refers to the factors that must be controlled for research on L1 influence to be as unbiased as possible, assuring impartiality to the greatest extent possible. He suggests nine conditions that should be controlled in the context of any L1 influence investigation:

1. age,
2. personality, motivation, and language aptitude,
3. social, educational, and cultural background,

4. language background (all previous L1s and L2s),
5. type and amount of target language exposure,
6. target language proficiency,
7. language distance between the L1 and target language,
8. task type and area of language use, and
9. prototypicality and markedness of the linguistic feature.

Again, the author recommends that all conditions are verified to guarantee neutrality and accuracy in the study of L1 influence. However, he acknowledges the difficulties in compiling linguistic data that is consistent in terms of, for example, the "personality, motivation, and language aptitude" (p.260-61) of the users. Information concerning personality would demand the participation of psychologists, and information concerning motivation and language aptitude requires the preparation and implementation of surveys.

In this thesis, the conditions 1, 2, and 5 were not verified beyond what can be inferred based on the concept of genre, community discourse and register. Instead of chronological age, the notion of scientific maturity can be considered, admitting that authoring a publication requires the ability to draft and a certain resilience to go through the editorial process. In this respect, it can be said that all authors in the CoRA have some basic scientific maturity because they have all published at least one article, i.e., the one compiled in the CoRA. Moreover, while every OSRA author has his/her own set of personal and professional motivations to learn a language and a certain language aptitude, all OSRA authors share certain characteristics. Both L1 and non-L1 English authors will have learned scientific English in specific situations and under specific conditions associated with the scientific register. However, L1 and non-L1 English authors will differ concerning the knowledge of general English they had when they started learning, how this register was taught to them, how much was taught, and the type of teaching they received. Similarly, the type and amount of target language exposure will be different for L1 and non-L1 English scientists, but a certain homogeneity can still be expected in each group.

This study complies with conditions 3, 4, 6, 7, 8, and 9. Overall, the balance of the linguistic data analyzed is guaranteed by means of the criteria applied to compile the CoRA.

Social, educational, and cultural background (condition 3) is verified through the concept of community discourse. That is, it can be argued that all authors are scientific researchers seeking to communicate their claims and findings using the genre OSRA and "specific lexis" they all share and know (Swales 1990: 24-26).

The language background (condition 4) is verified to the extent that is pertinent for this study, i.e., one L1 and one non-L1 that is English. The condition was also verified by following the criteria established in subsection 3.1.2.5 to decide on the L1/non-L1 of the OSRAs authors.

The proficiency in English (condition 6) is assumed to be that of a native (L1) for three of the groups and advanced non-native (non-L1) for two of the groups in the CoRA. The advanced level of proficiency of the non-L1 authors is assumed based on the linguistic outcome contained in the OSRAs, which has been proofread during the editorial correction process before being published.

The distance between the CoRA languages (condition 7) is assured by the homogeneity within each corpus. Three of the corpora contain OSRAs written in one specific L1 variety, i.e., European Portuguese, European Spanish, and British English; and the other two corpora contain English as produced by L1 speakers of the varieties European Portuguese and European Spanish. The distance among the L1 varieties and the non-L1 varieties should be similar among all OSRAs in the corpora.

The task type and area of language use (condition 8) is guaranteed in choosing one specific genre, i.e., the OSRA, from the same scientific field, i.e., health sciences. The prototypicality and markedness of the linguistic features (condition 9) are also controlled because all the texts in the CoRA belong in scientific writing. Thus, the linguistic features in all OSRAs should be consistent with what is likely to be found in that register. Likewise, the markedness of any linguistic element should be considered in the light of the prototypical characteristic of the register and the language variety in which the OSRA is written.

3.2.2. Studied Variables

The selection of the variables to be considered for comparison followed linguistic and computational approaches to authorship profiling, specifically to NLID. From the computational perspective, the variables are divided into two main groups. These are (a) content-independent; and (b) content-dependent variables (Weren et al. 2014: 267). Other terms used to nominate these groups are, correspondingly, style-based/content-based (e.g.

Argamon et al. (2009)); and non-linguistic/linguistic features (e.g. Kurdi (2019)); content-agnostic features and, by opposition, features that consider content (Sousa-Silva et al. 2010).

In the computational tradition, content-independent refers to features that do not convey linguistic meaning but rather express relations among words or other information about the text (Weren et al. 2014). Some examples of this type of variable are the mean frequency of function words, the mean number of words in a sentence, or the mean number of sentences in a paragraph.

In contrast, content-dependent refers to features that convey meaning, which may or may not vary according to the text domain (Kurdi 2019). In the context of this research, domain refers to the register of the text (scientific exposition) as proposed by Biber (1989) and to genre (research article) as described by (Swales 1990). It does not refer to the field of knowledge to which the CoRA texts belong.

The variables chosen were those considered to be relevant for the study. As a type of authorship profiling, NLID examines variables that are shared by groups of language users. While other forms of authorship profiling investigate the variation of language according to gender or age, NLID investigates the characteristics shared by groups of language users according to their L1. In this context, the relevance of a variable is defined first and foremost in terms of its representation of the class based on McMenamin's (2002: 130) definition of class style-marker.

Therefore, the class variable is any variable that is present and can be measured in no less than 95% of all the OSRAs in each corpus. The 95% threshold is used to guarantee that the observations reflect the tendency of the group, while leaving the possibility of some OSRAs not containing certain variables.

For example, within the variable 'punctuation marks' (See Table 9 for all variables), the sub-variables 'comma', 'semicolon', and 'colon' were selected for comparison. However, after obtaining their frequencies of distribution, it was observed that the semicolon and the colon were present in less than 95% of the OSRAs in each corpus. Table 8 shows the percentages at which these variables were found in the five corpora.

Variables	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}
Semicolon	76.92%	93.84%	93.84%	84.61%	93.84%
Colon	16.92%	24.61%	89.23%	89.23%	81.53%

Table 8 – Percentages of OSRAs containing the variables semicolon and colon

Similarly, within the variable 'number of words according to length', the results of the WordSmith tool included words from one letter-length up to 50 letter-length. After closer observation, it was verified that only words up to 15 letters were present in at least 95% of all the OSRAs in each corpus. Therefore, words with 16 letters or more were not examined.

The variables to be tested and analyzed within the content-independent group refer to those that have been shown to be most useful for authorship profiling – as the parent field of NLID; or that have been described as relevant for academic and scientific discourse (Argamon et al. 2009; Biber and Conrad 2009). For example, interjections are not considered since these are not expected to be found in research articles, although they are known to be found and play an important role for example in twitter texts (Silva et al. 2011).

The variables selected within the content-dependent group are informed by the quantitative results obtained from the analysis of the content-independent variables. For example, if the comparisons concerning variable 1 (V1), i.e., the average frequency of the punctuation mark comma show significant differences between any of the non-L1 groups and the L1 group, linguistic analysis is carried out to explain such differences. In this case, the explanation seeks to understand the link, if any, between the results obtained and the use of the punctuation mark in the academic and the general discourse of the authors' L1.

In the case of variables like V2: Average word length and V7: Nº of words according to length (See V7, V8, V9 in Table 9), the finding of significant differences may lead to the analysis of word-formation processes that contribute to extending the length of words. Some processes like nominalizations are more relevant than others to academic and scientific writing because they have been described as being "extremely common" (Biber and Conrad 2009: 116), and so, in this case, specific nominalizations would be content-dependent variables to be examined. Table 9 below presents the variables studied.

Variable Type	Categories	Variables

Content-independent	Punctuation Marks Distribution
	Sentence Length
	Paragraphs Distribution
		V1: number of commas;
		V2: average sentence length in words;
		V3: number of paragraphs;

	Lexical density	**V4:** standardized type/token ratio

	Words Length Distribution	**V5** number of 1 to 5-letter words;
		V6: number of 6 to 10-letter words;
		V7: number of 11 to 15-letter words

	Function Words distribution	**V8**: number of definite articles; **V9**: number of indefinite articles; **V10**: number of coordinating conjunctions; **V11**: number of subordinating conjunctions; **V12**: number of prepositions; **V13**: number of demonstrative pronouns; **V14**: number of relative pronouns; **V15:** number of personal pronouns.

	Part-of-speech distribution	**V16**: number of adjectives;
		V17: number of adverbs;
		V18: number of nouns;
		V19: number of verbs.

Content-dependent	Informed by the results obtained from the quantitative analysis of content-independent variables	i.e., Analysis of demonstrative pronouns Analysis of adjectives Analysis of adverbs

Table 9 – The variables studied

In relation to V4 – STTR, it is worth noting that the type/token ratio (TTR) is the quotient of the number of running words by the number of different words in a given text. A standardized TTR calculates this ratio every n words in a given document (Scott 2018a: 355). WordSmith uses 1000 words as a default setting to calculate the STTR (p. 355). Even though WordSmith allows for the user to change the number of words to be considered in the standardization of the TTR (p. 355), no particular advantages were recognized in changing the default settings for this study. Therefore, the STTR was calculated using the default settings, i.e., every 1000 words.

The analyses seek to examine first if there are significant differences among the groups in relation to frequencies of occurrences of the variables. The variables whose distribution are shown to be statistically significantly different between the groups are analyzed linguistically. Also, POS are analyzed linguistically regardless of statistical results for frequencies. The linguistic analyses seek to examine differences/similarities of recurrent patterns between the CoRA groups. Differences/similarities are analyzed in terms of:

a) deviation from the norm; and
b) variation within the norm authors are influenced by (McMenamin 2002).

The norm comprises general language and, in a more restricted sense, standard L1 academic language, whether English, Portuguese or Spanish. **Deviation** refers to language mistakes or errors (e.g., *'focus in' instead of 'focus on'), and **variation** refers to linguistic choices that are correct in both general language and the academic norms, or accepted mainly within the academic norm, even when not so frequent or accepted in the general language (McMenamin 2002: 135). However, language mistakes or errors are not expected since all the OSRAs in the CoRA have been peer-reviewed and proof-read. Therefore, the linguistic examination is expected to be described in terms of variation.

All variables are analyzed as ratios of the absolute value and the total number of tokens of the OSRA to which it relates in the corresponding corpus. Using ratios ensures the proportionality of the data since all proposed variables may increase their frequency as the number of words in an OSRA increases. The exceptions to the ratio rule are the following:

- The measurement of the average sentence length in words (ASLiW) because it is expressed as a mean value, and it does not depend directly on the total extension of the OSRA to which it relates;
- The measurement of the standardized type/token ratio (STTR) because it is obtained from WordSmith;
- The number of paragraphs is expressed in absolute values since this variable does not depend directly on the total number of words in an OSRA.

The frequencies of 18 variables were obtained from the parsed .txt files of the OSRAs in each corpus using WordSmith. 13 variables (V1; and V8 to V19) were extracted using queries written with the specific syntaxes of those variables in the parsed files. The syntaxes of the queries used to extract the frequencies of those 13 variables per corpus are presented below in Table 10.

Variable	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}
V1: number of commas	/,SOURCE:/	/,/	/ [,] PU @PU/		
V8: number of definite articles	/<artd>/		/ART/		
V9: number of indefinite articles	DET				
V10: number of coordinating conjunctions	/KC/ and /<kc>/		KC		
V11: number of subordinating conjunctions	/KS/ and /<ks>/		KS		
V12: number of prepositions	/PRP/ and /<prp>/		/PRP @/ and /<prp-/		
V13: number of demonstrative pronouns	/<dem>/				
V14: number of relative pronouns	/<rel>/				
V15: number of personal pronouns	/<PERS>/				
V16: number of adjectives	/ADJ/				
V17: number of adverbs	/ADV/				
V18: number of nouns	[N F P]; [N F S]; [N M P]; [N M S]; [N M S/P]; [N M/F P] and [N M/F S]		[N P] and [N S]		
V19: number of verbs	[V]				

Table 10 – Query syntaxes for extracting variables of analysis

Variables 2, 4, 5, 6, and 7 were obtained differently. Variables 2 and 4 (average sentence length in words per OSRA (ASLiW) and the type/token ratio) within each corpus were obtained automatically from WordSmith Tools upon requesting the corpora texts statistics. Variables 5, 6, 7, i.e., the number of 1-to-5, 6-to-10, and 11-to-15-letter words, were calculated in MS Excel by adding up the numbers of the word list columns corresponding to the number of letters (from 1-to-*n*-letters) per OSRA within each corpus obtained automatically using WordSmith Tools.

Only variable 3, i.e., total paragraph count, was counted manually because, despite WordSmith Tools calculating the number of paragraphs automatically, the .txt files used in parsing were not prepared to respect the paragraph division within each OSRA since initially that variable was not going to be analyzed.

3.2.3. Operationalization of the Research Questions

1. **Are there variables associated with L1 influence in OSRAs written in English by PT-EU and ES-EU L1 authors in the field of health sciences? If so,**
    1.1 what are those variables?

As mentioned in the introduction, this is the core question of the present empirical research, which uses descriptive and inferential statistics in relation to the variables in Table 9 to describe all the corpora in the CoRA and compare the groups to understand their differences. The examination is both quantitative and qualitative since linguistic analysis is also implemented.

I implemented Jarvis' (2000, 2010) framework for Investigating L1 Influence, and compared the groups with each other. The comparisons are designed to account for L1 vs. non-L1 differences in scientific writing in English and are operationalized through statistical tests, and then through the linguistic analysis of the variables.

The combination of approaches (quantitative and qualitative) attempts to discriminate style-markers of the linguistic choices and preferences of non-L1 English authors who are L1 speakers of PT-EU/ES-EU. The comparisons address the linguistic patterns (variables) as observed in scientific writing.

According to the proposed methodological framework of analysis, the corpora that need to be considered in the assessment of L1 influence in OSRAs written in English by L1 PT^{EU} authors are the PT-EU, EN-GB, EN-PT^{EU}, and EN-ES^{EU}. Likewise, the corpora that need to be considered in the assessment of L1 influence in OSRAs written in English by L1 ES^{EU} authors are the ES-EU, EN-GB, EN-PT^{EU}, and EN-ES^{EU}. For all variables, I posed the following questions for both EN-PT^{EU} and EN-ES^{EU} OSRAs (Table 11).

Effect of L1 Influence	L1 influence EN-PT^{EU} questions	L1 influence EN-ES^{EU} questions
I) Intragroup homogeneity	Are the EN-PT^{EU} / EN-ES^{EU} OSRAs uniformly distributed?	
II) Intergroup heterogeneity	Are the EN-PT^{EU} and EN-ES^{EU} OSRAs statistically significantly different?	
III) Cross-language congruity	Are the EN-PT^{EU} and PT-EU OSRAs statistically similar?	Are the EN-ES^{EU} and ES-EU OSRAs statistically similar?
IV) Intralingual contrast	Are the EN-PT^{EU} and EN-GB OSRAs statistically significantly different?	Are the EN-ES^{EU} and EN-GB OSRAs statistically significantly different?

Table 11 – L1 Influence questions to test in the compiled corpora according to the framework for Investigating L1 Influence (Jarvis 2010, 2000).

As described above, at least two out of four effects have to be found to claim L1 influence in any of the groups.

Questions in relation to L1 effects II to IV are answered using the independent-samples *t*-test, which assesses the difference between the means of two independent groups in relation to a given dependent variable that has been measured on a continuous scale. Independence refers to the fact that the two groups are not related and have been measured only once in terms of the dependent variable being analyzed, and 'continuous scale' refers to the variables being presented as intervals or ratios (Eddington 2016: 53-64).

As a parametric statistical test, the independent samples *t*-test requires that the samples meet another three assumptions besides the samples' independence and continuous scale of measurement. If these assumptions are not met, the results of the test cannot be considered valid. These assumptions also refer to the variables in the analysis. These are 1) absence of outliers, 2) normality of the distribution of the variable, and 3) homogeneity of variance of the samples.

Therefore, before statistical analysis, all variables are examined to detect outliers, test normality, and verify the homogeneity of variance.

The homogeneity of the samples' variances is determined by Levene's test, which "asks whether the variances between [two samples of a given variable] are significantly different from each other" and is a test performed as part of the independent *t*-test (Eddington 2016: 56). Question I was examined by contrasting the results of Levene's test of homogeneity of variances with the results of the independent samples *t*-tests performed to prove inter-L1 heterogeneity (groups EN-PT^{EU} vs. EN-ES^{EU}).

As stated in previous studies (Paquot 2013; Jarvis 2000), from the quantitative perspective, a practical form of proving intra-L1 homogeneity is by verifying if the variance within each group in relation to a given variable is smaller than the difference between the groups. That is, if Levene's test shows that the variances of two samples of a given variable are not significantly different, i.e., they are fairly similar, and the independent samples *t*-test shows that the mean values of the variable frequencies of the same samples are significantly different, then the samples are uniform but from different populations.

Outliers are identified with SPSS using descriptive statistics and double-checked using the labeling rule proposed by Hoaglin and Iglewicz (1987), according to which the upper boundary is calculated with the formula Q3+(2.2*(Q3-Q1)), and the lower boundary is calculated with the formula Q1-(2.2*(Q3-Q1)). The letter Q stands for quartile which is any of "the scores which cut off the bottom 25%, 50% and 75% of scores in a sequence of scores ordered from the smallest to the largest […], known as the first, second and third quartiles." (Cramer and Howitt 2004: 133).

Based on the Central Limit Theorem, all variables' samples were assumed to be approximately normally distributed. The Central Limit Theorem states that "the sampling distribution of the mean for any population, given an adequate sample size, will approximate a standard normal distribution" (Aberson et al. 2000: 289). In statistics, for a sample to be considered as having an adequate size it must be of at least 30 observations (Anderson 2010). Since all samples in this study contain 65 observations, and the chosen statistical test works with mean values, the sample sizes are considered to be large enough to assume their approximate normal distribution.

In total, 95 independent samples *t*-tests were performed to answer questions II to IV for each variable per group (EN-PT^{EU} and EN-ES^{EU}). Nineteen of Levene's tests performed as part of the independent samples *t*-tests were used to answer question I. The level of significance used in this research is $p < .05$ for questions II and IV. For questions I and III, since uniformity and congruity are expected, respectively, a result of p > .05 is associated with an existing L1 effect.

Finally, for statistical results that indicated the possibility of the existence of an L1 influence effect, Cohen's *d* was calculated with the following formula (Eddington 2016: 54):

$$Cohen's\ d = \frac{2(t)}{\sqrt{df}}$$

Cohen's *d* tells us about the magnitude or size of the effect of the independent variable (i.e., the language variety) on the dependent variable (i.e., those in Table 9), which can be small (from - 0.2 to 0.2), medium (from - 0.5 to 0.5) or large (from - 0.8 to 0.8) (Eddington 2016: 54).

The second part of the analyses consists of an examination of POS variables based on a method that follows five main steps. Variables 8 to 19 are further analyzed in terms of the specific POS that may function as NLID markers using this method.

In step (1) an *ad hoc* threshold is used to cut across the corpora to obtain group characterizing variables. This analysis entails the extraction of the most frequent POS in each corpus present in at least 50% of the OSRAs of each corpus. Because NLID works with variables that characterize the group in relation to the L1, a criterion of group distribution must be applied to guarantee that whatever POS is associated with NLID is as representative as possible of the group. At the stage of selection of the variables, a 95% threshold was used to guarantee that the observations reflected the tendency of the group while leaving the possibility of some OSRAs not containing certain variables. However, at this stage, the same criterion turned out to be too restrictive as only a few POS could be obtained per corpus using a 95% threshold. Moreover, an important part (about 50% to 75%) of those 95% are present in all corpora at very similar frequencies, leaving little material for NLID analysis. Therefore, after several trials using different cutoff points, I verified that to obtain a number of POS large enough to allow for linguistic analysis and still ensure class representation, the threshold had to be lowered to 50%. The threshold responds to particular research conditions. The threshold used to extract NLID characterizing variables may vary according to the size of the corpora, the genre, the register and any other element that may interfere with the text. Hence, the *ad hoc* in the naming of this step.

In step (2) an analysis of the most frequent words obtained in step 1 in each corpus are examined by looking at:

- Words ranked higher in the EN-GB corpus than in the non-L1 English corpora (EN-PT^{EU}/EN-ES^{EU}); and whose equivalent(s) in the L1 Portuguese/Spanish corpora (PT-EU/ES-EU) ranked lower than in the L1 English corpus (EN-GB) corpus;
- Words ranked much higher in one or both non-L1 English corpora (EN-PT^{EU}/EN-ES^{EU}) than in the L1 English corpus (EN-GB); and whose equivalent(s) in the corresponding L1 corpora (PT-EU/ES-EU) also ranked higher than the EN-GB corpus.

In step (3), the potential of these words to function as NLID markers is also first assessed quantitatively following the Unified Framework for Investigating L1 Influence described in chapter 3 (Jarvis 2010; Jarvis and Pavlenko 2008; Jarvis 2000). The intragroup homogeneity is also assessed by Levene's, and the intergroup heterogeneity, the cross-language congruity, and the intralingual contrast were evaluated by the Mann-Whitney test, the alternative to the independent samples *t*-test that is used when the samples lack normal distribution, have many outliers, or the variable is measured on an ordinal scale (Eddington 2016: 58). In this case, although the variables are all measured on a continuous scale, the POS samples deemed to have NLID marker potential were not normally distributed and, in many cases, had some outliers.

In steps (4) and (5) the words considered to have the potential to function as NLID markers, based on quantitative analysis, are analyzed linguistically, specifically when taking their syntactic tagging from VISL into consideration.

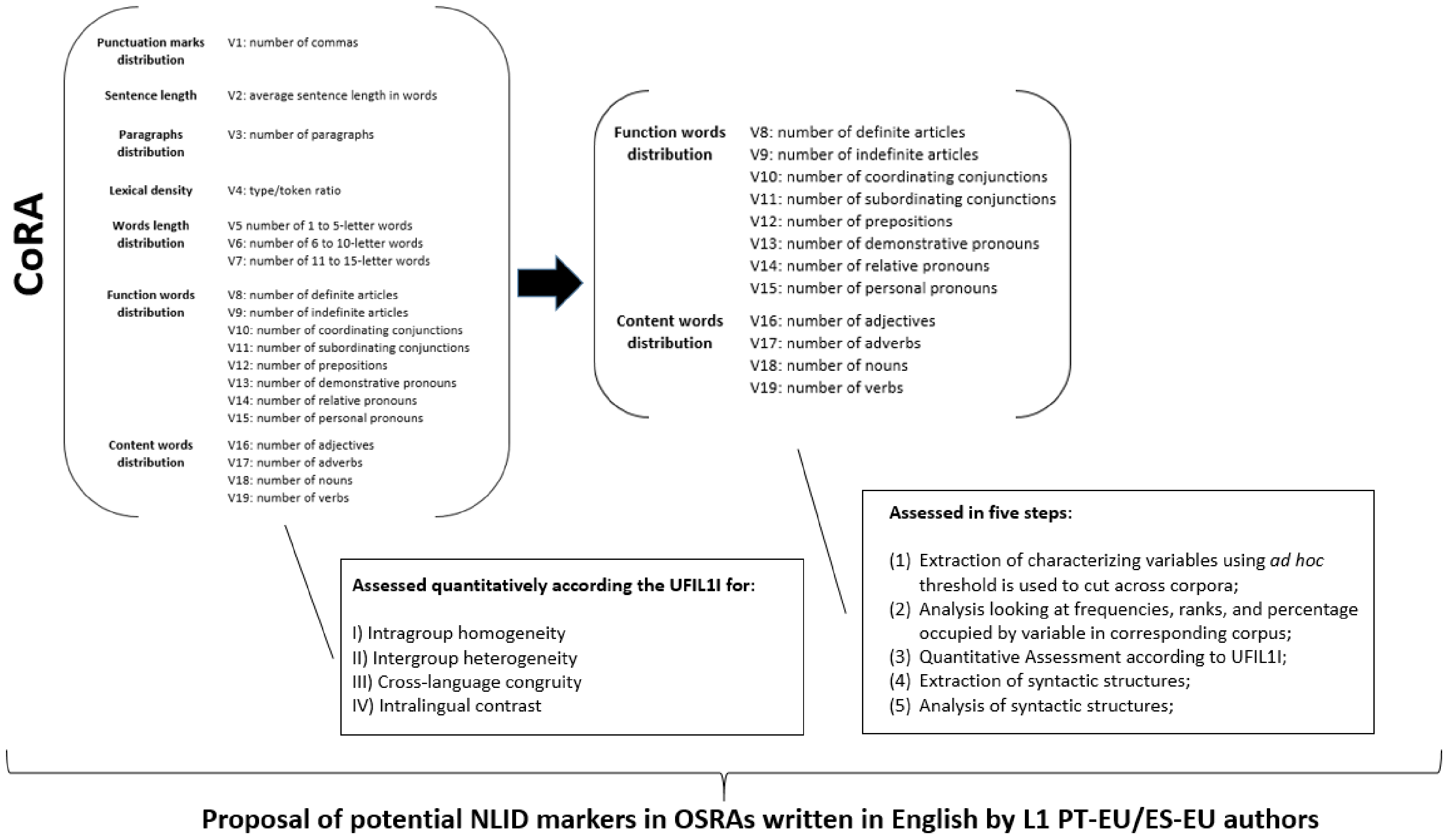
Figure 6, in the end of this section, shows a diagram representing the operationalization of the research question 1 and its derivative 1.1.

2. Is it possible to explain the absence/presence of L1 influence variables in OSRAs written in English by L1 authors of PT-EU and ES-EU L1?

The second research question is addressed by examining and discussing the results obtained from the comparisons carried out to answer question 1 and derivative 1.1. The objective is to interpret the results either in the light of linguistic theories from the fields of linguistics, language transfer, and sociolinguistics reviewed in the theoretical part of this research or by contrast with previous studies reporting comparable or opposite results.

3. **Are there implications associated with the absence/presence of L1 influence variables in OSRAs written in English by L1 authors of PT-EU and ES-EU L1?**

This question is addressed in the final discussion of the study. It is assumed that any results obtained will have some implication. Therefore, the very completion of the study contributes to the characterization of written scientific English, specifically of that realized by L1 PT-EU and ES-EU researchers in the health sciences. I aim at reflecting on other possible meanings of the findings in other areas. The first area refers to the teaching of scientific English to health sciences students. The second is related to direct professional significance for proofreaders and editors of OSRAs in the health sciences in Portugal. Similarly, I reflect on implications for translators working with Portuguese/Spanish and English in the health sciences.



CoRA

Punctuation marks distribution — V1: number of commas

Sentence length — V2: average sentence length in words

Paragraphs distribution — V3: number of paragraphs

Lexical density — V4: type/token ratio

Words length distribution — V5 number of 1 to 5-letter words; V6: number of 6 to 10-letter words; V7: number of 11 to 15-letter words

Function words distribution — V8: number of definite articles; V9: number of indefinite articles; V10: number of coordinating conjunctions; V11: number of subordinating conjunctions; V12: number of prepositions; V13: number of demonstrative pronouns; V14: number of relative pronouns; V15: number of personal pronouns

Content words distribution — V16: number of adjectives; V17: number of adverbs; V18: number of nouns; V19: number of verbs

Assessed quantitatively according the UFIL1I for:

I) Intragroup homogeneity
II) Intergroup heterogeneity
III) Cross-language congruity
IV) Intralingual contrast

Function words distribution — V8: number of definite articles; V9: number of indefinite articles; V10: number of coordinating conjunctions; V11: number of subordinating conjunctions; V12: number of prepositions; V13: number of demonstrative pronouns; V14: number of relative pronouns; V15: number of personal pronouns

Content words distribution — V16: number of adjectives; V17: number of adverbs; V18: number of nouns; V19: number of verbs

Assessed in five steps:

(1) Extraction of characterizing variables using *ad hoc* threshold is used to cut across corpora;
(2) Analysis looking at frequencies, ranks, and percentage occupied by variable in corresponding corpus;
(3) Quantitative Assessment according to UFIL1I;
(4) Extraction of syntactic structures;
(5) Analysis of syntactic structures;

Proposal of potential NLID markers in OSRAs written in English by L1 PT-EU/ES-EU authors

Figure 6 – Research question 1 operationalization diagram.

3.3. Chapter Summary

Chapter 3 presented the details of the methodology followed in this research. In the first part, I described the corpora design. I explained that even though there are corpora of scientific discourse produced by L1 English, Portuguese, and Spanish authors in their respective L1 and by L1 Portuguese and Spanish authors writing in English, these are not suitable for this study. Most of the existing corpora of academic text were not adequate due to matters of text typology, genre, the disciplinary areas addressed, and the costs associated with access. Some corpora contain research articles in health-related fields but restricted to parts of the articles such as the abstract. Other corpora contain academic text but include many different genres (e.g., lectures, journals, essays) and sometimes L1 authors at different levels of writing proficiency. For these reasons, this research was carried out using its own corpora. The rationale for using one's own corpora is the need for corpora of a specific genre and the comparability of the texts since, to the best of my knowledge, there are no ready-to-use, accessible and annotated corpora of the genre OSRA in the target language varieties (European Portuguese, European Spanish, British English, non-L1 English produced by L1 European Portuguese and non-L1 English produced by L1 European Spanish).

The chapter continues with the description of the corpora type and design, the selection criteria, the pre-processing, preparing, and parsing of the texts, and the limitations to the corpora compilation. At the end of that section, a general description of the corpora is presented and the name of the collection (Comparative Corpora of Research Articles - CoRA) is provided. CoRA is a synchronic, personally compiled, and parsed collection of 325 OSRAs and eight hundred twenty-five thousand four hundred and three tokens from the health sciences published in peer-reviewed indexed journals from 2006 to 2018. CoRA contains 5 corpora: three of OSRAs written in European Portuguese, European Spanish, and British English by L1 authors of those language varieties; and two of OSRAs written in English by non-L1 authors whose native language is either European Portuguese or European Spanish. Most of the OSRAs in the CoRA are published in open access and the parts of the OSRAs included in the collection were the introduction, results, discussion, and conclusions.

The second part of the chapter outlines the study design and model of analysis. First, there is a description of the approach adopted to investigating the influence of the L1 on scientific writing in English. An explanation is provided of the comparison-based approach called Unified Framework for investigating L1 influence (Jarvis 2000, 2010) by demonstrating the four types of evidence –Intra-L1-group homogeneity, Inter-L1-group heterogeneity, Cross-language congruity, and Intralingual contrast – that can prove L1 influence by following group-based and source-language-based premises and by carrying out within and between group, and between and within language comparisons.

This is followed by a description of the variables considered for comparison. The variables to be tested and analyzed refer to those that have been shown to be most useful for authorship profiling, the parent field of NLID, or that have been described as relevant for academic and scientific discourse (Argamon et al. 2009; Biber and Conrad 2009). Since the detection of L1 influence implies examining variables that are shared by groups of language users, the variables analyzed in this study are first those that are found and can be measured in no less than 95% of all the OSRAs in each corpus, and then those informed by the results obtained in the first analyses. Overall, 19 variables from 7 categories were chosen for analysis. The analyses are thought to examine first the differences among the groups in relation to frequencies of occurrences of the variables, and then those variables with significantly different distributions between the groups are analyzed linguistically. Also, the POS are analyzed linguistically regardless of the statistical results obtained.

Finally, the operationalization of the research questions was explained. To answer the first questions I formed the hypotheses that result from the implementation of Jarvis' (2000, 2010) framework for Investigating L1 Influence, explained how the groups are compared, and decided on the number of effects needed to claim L1 influence in any of the groups (at least two out of four effects). There is a description of how the L1 effect of type I is found by Levene's test, and how effects II to IV are found using the parametric test independent-samples *t*-test. Also, there is an explanation of all the assumptions that must be met to carry out independent-samples *t*-tests and the Cohen's *d* in order to learn the magnitude or size of the effect. This is followed by a description of the method followed to examined variables 8 to 19 in terms of the specific POS that may function as NLID markers. The second research question is expected to be answered by interpreting the results of the analyses developed to answer question 1 and derivative 1.1 according to linguistic theories from the fields of linguistics, language transfer, and sociolinguistics. The third research question is addressed in the final discussion of the study, on the understanding that any results obtained will have some implications.

4. Investigating L1 Influence in OSRAs

This chapter presents the findings obtained from the examination of the CoRA and examines nineteen variables. The chapter is divided into six sections. The first section presents the statistical and general linguistic characterization of the CoRA. Sections 4.2 to 4.5 present the findings of the comparisons performed to assess L1 influence in the two groups of OSRAs described in chapter 3, i.e., OSRAs written in English by L1 PT^{EU} authors (EN-PT^{EU}); and OSRAs written in English by L1 ES^{EU} authors (EN-ES^{EU}). The findings are presented per variable according to the four types of evidence proposed by the methodological framework described in chapter 3. The variables are grouped in each section according to the results obtained. The statistical results are shown for all variables. The findings obtained from the linguistic analysis are presented for parts-of-speech. The last section discusses the implications of the findings.

4.1. Corpora Characterization

As shown in Table 12 below, the five corpora are similar in relation to the total number of tokens, types, standardized type/token ratio, length of sentences in words, and length of words in characters.

Corpus	Average of tokens (running words)	Average of types (distinct words)	Average of standardized type/token ratio (STTR)	Average of sentence length (in words)	Average of word length (in characters)
PT-EU	2192,55	687,71	32,67	76,91	5,38
ES-EU	2484,66	711,89	30,02	79,14	5,19
EN-GB	2742,17	739,54	27,74	101,31	5,23
EN-PT^{EU}	2513,75	677,45	28,21	95,11	5,29
EN-ES^{EU}	2834,28	714,91	26,17	102,63	5,28

Table 12 – Mean values of tokens, types, STTR, sentence, and word length per corpus in the CoRA

Similarly, each of the nineteen variables has a fairly similar distribution in terms of frequencies across the five corpora. For practical reasons, the descriptive statistics per variable are shown in the corresponding subsection of the results of each variable.

After analysis of the assumptions of the independent-samples *t*-tests, 26 outliers were detected in the corpora samples and corrected following deletion or value replacement procedures. All cases corresponded to actual lowest or highest sample values. Two cases, however, corresponded to exceptional situations. One was found in the distribution of commas in the ES-EU corpus, and another four were detected in the type/token ratios of the PT-EU, the ES-EU, and the EN-PT^{EU} corpora.

In the first case, the data extracted from the corpora indicated that OSRA 52 of the ES-EU corpus had zero commas. After careful reading of the OSRA in its PDF format, it was determined that no mistake or typing error had been made when inserting the values of the number of commas in the database. The authors of article 52 of the ES-EU corpus simply did not use any grammatical comma in any of the OSRA parts included in the compilation. Writing a whole OSRA without using one comma is very rare. A close reading of the OSRA showed that the absence of commas corresponded to actual grammatical errors or mistakes. Two examples of such errors in Spanish are the absence of commas *(a.)* after an adverbial subordinate clause that provides information about the location of the information explained thereafter, and *(b.)* in an enumeration.

a. *Como puede observarse en la Figura 1**[,]** estos microorganismos van aumentando su concentración desde la boca hasta el recto**[,]**siendo máxima en el colon con aproximadamente* 10^{12} *UFC (Unidades Formadoras de Colonias) por gramo de contenido intestinal.*

b. *El colon está habitado por unas cuatrocientas especies bacterianas y se divide en colon ascendente o proximal**[,]** colon transversal**[,]** colon descendente o distal y colon sigmoidal.*

In general, the complete lack of commas, whether due to grammatical errors or due to the OSRA authors' style, is not distinctive of the category 'OSRAs written in ES-EU'. In fact, in this case, since the errors are found throughout the whole OSRA, it seems that there was some problem with the edition of the document or that it was intentional. Therefore, to account for this, it was decided to delete the observation. As a result, the ES-EU corpus was left with 64 cases, with one missing case corresponding to 1.5% of the comma sample.

In the second group of cases, the analysis detected four outliers in the PT-EU, the ES-EU, and the EN-PT^{EU} corpora of variable 3 (standardized type/token ratio) samples. After close examination of the data, it was verified that the values of V4 for those observations were zero. This is because the standardized type/token ratio (STTR) was computed by WordSmith every 1000 tokens (as set by default in the software and maintained for reasons explained above in section 3.2.2), and texts shorter than 1000 get an STTR of 0. The detected outliers belong to OSRAs with 916, 763, 765, and 947 tokens. The outliers were replaced with the next lowest value in their respective distribution, resulting in a 6.15% winsorized[12] total sample. Table 13 summarizes the rest of the outliers found and the action taken to correct them statistically.

Variable	Corpus	OSRA	Type of extreme value	Action taken to correct outliers	% of samples winsorization
V1: number of commas	ES-EU	52	Lowest	deleted	Does not apply
V2: average sentence length in words	PT-EU	45	Highest	replaced with the next highest value in the respective distribution	4.61%
	ES-EU	2			
	EN-ES^{EU}	58			
V3: number of paragraphs	ES-EU	54	Highest	replaced with the next highest value in the respective distribution	9.22%
	ES-EU	62			
	EN-GB	18			
	EN-GB	26			
	EN-GB	38			
	EN-ES^{EU}	49			
V4: type/token ratio	PT-EU	35	Lowest	replaced with the next lowest value in the respective distribution	6.15%
	ES-EU	64			
	EN-PT^{EU}	14			
	EN-PT^{EU}	20			

[12] It refers to a form of treatment of a "genuine outlier", i.e., an extreme value that does not result from measurement, transcription, interpretation, sampling or other errors; it is an authentic extreme value to which the researchers "assign lesser weight or modify […] so it is closer to the other sample values" (Ghosh and Vogt 2012: 3455)

V8: number of definite articles	ES-EU	66	Lowest	replaced with the next lowest value in the respective distribution	4.61%
		67			
	EN-GB	8	Highest	replaced with the next highest value in the respective distribution	
V9: number of indefinite articles	EN-GB	8	Highest	replaced with the next highest value in the respective distribution	1,54%
V12: number of prepositions	EN-GB	8	Highest	replaced with the next highest value in the respective distribution	3.08%
	EN-ES^{EU}	38	Lowest	replaced with the next lowest value in the respective distribution	
V15: number of personal pronouns	EN-GB	9	Highest	replaced with the next highest value in the respective distribution	1.54%
V18: number of nouns	EN-GB	8	Highest	replaced with the next highest value in the respective distribution	3.08%
	EN-ES^{EU}	38	Lowest	replaced with the next lowest value in the respective distribution	
V19: number of verbs	EN-GB	8	Highest	replaced with the next highest value in the respective distribution	4.61%
		9			
	EN-ES^{EU}	38	Lowest	replaced with the next lowest value in the respective distribution	

Table 13 – Summary of outliers per variable and corpus.

No outliers were detected in the remaining samples of the variables.

The assumption of homogeneity of variances indicated by Levene's test was met in 72 cases of comparisons (75.79%) described in Table 14 below.

Corpora compared	Variable	*F*	*p* > 0.05
EN-PT^{EU} – EN-ES^{EU}	number of commas	3,587	0,060
	number of paragraph	2,098	0,150
	standardized type/token ratio	2,697	0,103
	number of 1 to 5-letter words	0,021	0,884
	number of 6 to 10-letter words	0,035	0,852
	number of 11 to 15-letter words	0,045	0,833
	number of definite articles	0,657	0,419
	number of indefinite articles	0,013	0,908
	number of coordinating conjunctions	0,506	0,478
	number of subordinating conjunctions	2,778	0,098
	number of prepositions	0,033	0,855
	number of demonstrative pronouns	0,762	0,384
	number of relative pronouns	0,043	0,836
	number of personal pronouns	0,033	0,856

	number of adjectives	2,816	0,096
	number of nouns	0,929	0,337
	number of verbs	0,272	0,603
EN-PT^{EU} – PT-EU	number of commas	3,010	0,085
	average sentence length in words	2,053	0,154
	standardized type/token ratio	1,680	0,197
	number of 1 to 5-letter words	0,094	0,760
	number of 6 to 10-letter words	0,878	0,351
	number of 11 to 15-letter words	0,431	0,513
	number of definite articles	3,875	0,051
	number of indefinite articles	0,199	0,657
	number of coordinating conjunctions	1,681	0,197
	number of subordinating conjunctions	0,246	0,621
	number of prepositions	0,232	0,631
	number of personal pronouns	0,647	0,423
	number of adjectives	1,144	0,287
	number of adverbs	0,042	0,837
EN-PT^{EU} – EN-GB	number of commas in OSRA	0,329	0,567
	standardized type/token ratio	0,724	0,396
	number of 1 to 5-letter words	3,328	0,070
	number of 6 to 10-letter words	0,929	0,337
	number of 11 to 15-letter words	0,465	0,497
	number of definite articles	3,390	0,068
	number of indefinite articles	0,134	0,715
	number of coordinating conjunctions	1,067	0,304
	number of subordinating conjunctions	0,693	0,407
	number of demonstrative pronouns	1,611	0,207
	number of relative pronouns	0,069	0,793
	number of adjectives	0,007	0,931
	number of adverbs	0,567	0,453
	number of nouns	0,390	0,533
	number of verbs	3,280	0,072
EN-ES^{EU} – ES-EU	number of commas in OSRA	0,167	0,683
	number of paragraph	2,062	0,153
	standardized type/token ratio	3,764	0,055
	number of 1 to 5-letter words	1,374	0,243
	number of 6 to 10-letter words	3,516	0,063
	number of 11 to 15-letter words	0,001	0,971
	number of definite articles	0,002	0,967
	number of indefinite articles	0,149	0,700
	number of coordinating conjunctions	0,130	0,719
	number of subordinating conjunctions	3,776	0,054
	number of prepositions	2,026	0,157
	number of personal pronouns	1,130	0,290
	number of adjectives	0,078	0,780
	number of adverbs	1,189	0,278

EN-ES^{EU} – EN-GB	average sentence length in words	0,000	0,999
	number of paragraph	3,319	0,071
	standardized type/token ratio	0,303	0,583
	number of 1 to 5-letter words	2,961	0,088
	number of 6 to 10-letter words	1,256	0,265
	number of 11 to 15-letter words	0,210	0,647
	number of indefinite articles	0,064	0,800
	number of coordinating conjunctions	0,055	0,815
	number of subordinating conjunctions	0,485	0,487
	number of relative pronouns	0,004	0,948
	number of adjectives	2,749	0,100
	number of nouns	0,110	0,741

Table 14 – Levene's tests results for samples that met the homogeneity of variance assumption

However, the assumption of homogeneity of variances indicated by Levene's test was not met in 23 of the cases compared (24.21%) and described below in Table 15.

Corpora compared	**Variable**	***F***	***p* < 0.05**
EN-PT^{EU} – EN-ES^{EU}	average sentence length in words	7,027	0,009
	number of adverbs	4,755	0,031
EN-PT^{EU} – PT-EU	number of paragraphs	5,070	0,026
	number of demonstrative pronouns	4,377	0,038
	number of relative pronouns	4,784	0,031
	number of nouns	5,445	0,021
	number of verbs	16,293	0,000
EN-PT^{EU} – EN-GB	average sentence length in words	8,428	0,004
	number of paragraph	8,326	0,005
	number of prepositions	5,950	0,016
	number of personal pronouns	10,215	0,002
EN-ES^{EU} – ES-EU	average sentence length in words	4,615	0,034
	number of demonstrative pronouns	56,306	0,000
	number of relative pronouns	5,573	0,020
	number of nouns	5,243	0,024
	number of verbs	6,633	0,011
EN-ES^{EU} – EN-GB	number of commas	5,707	0,018
	number of definite articles	7,486	0,007
	number of prepositions	4,964	0,028
	number of demonstrative pronouns	4,461	0,037
	number of personal pronouns	10,111	0,002
	number of adverbs	6,845	0,010
	number of verbs	4,355	0,039

Table 15 – Levene's tests results for samples that did not meet the homogeneity of variance assumption

Overall, eight variable samples met the homogeneity of variance across all the corpora. These are standardized type/token ratio, number of 1 to 5-letter words, number of 6 to 10-letter words, number of 11 to 15-letter words, number of indefinite articles, number of coordinating conjunctions, number of subordinating conjunctions, and number of adjectives.

Six variable samples (number of nouns, number of paragraphs, number of personal pronouns, number of prepositions, number of relative pronouns, and number of adverbs) did not meet the assumption of homogeneity of variance in two corpora pairs; three variable samples (average sentence length in words, number of verbs, number of demonstrative pronouns) did not meet the assumption of homogeneity of variance in three corpora pairs; and two samples (number of commas and number of definite articles) did not meet the assumption of homogeneity of variance in one corpora pair.

For the 23 cases in which the assumption of homogeneity of variances was not met, the output of Welch's *t*-test was used to interpret the results of the comparisons (Eddington 2016: 56). Welch's *t*-test is run automatically when the independent samples *t*-test is run and can be examined immediately after interpreting Levene's results.

As expected, the most frequent words in all the corpora are articles and certain prepositions and conjunctions. As shown in Figure 7, in the English corpora, whether containing OSRAs authored by L1 or non-L1 users of the language, the most frequent words are the function words '*the', 'of', 'in', 'and',* and '*to'*. In the PT-EU corpus, the most frequent words are the function words '*de', 'a', 'e', 'o',* and *'que'*; and in the ES-EU corpus, the most frequent words are the function words *'de'*, *'la'*, *'en'*, *'el'* and *'y'*. The least frequent words in each corpus are the verbal forms '*abandonada'* (PT-EU), '*abandonado*' (ES-EU), *'ablated'* (EN-ES^{EU}), and '*abbreviated'* (EN-GB); and the adverb *'alarmingly'* (EN-PT^{EU}).

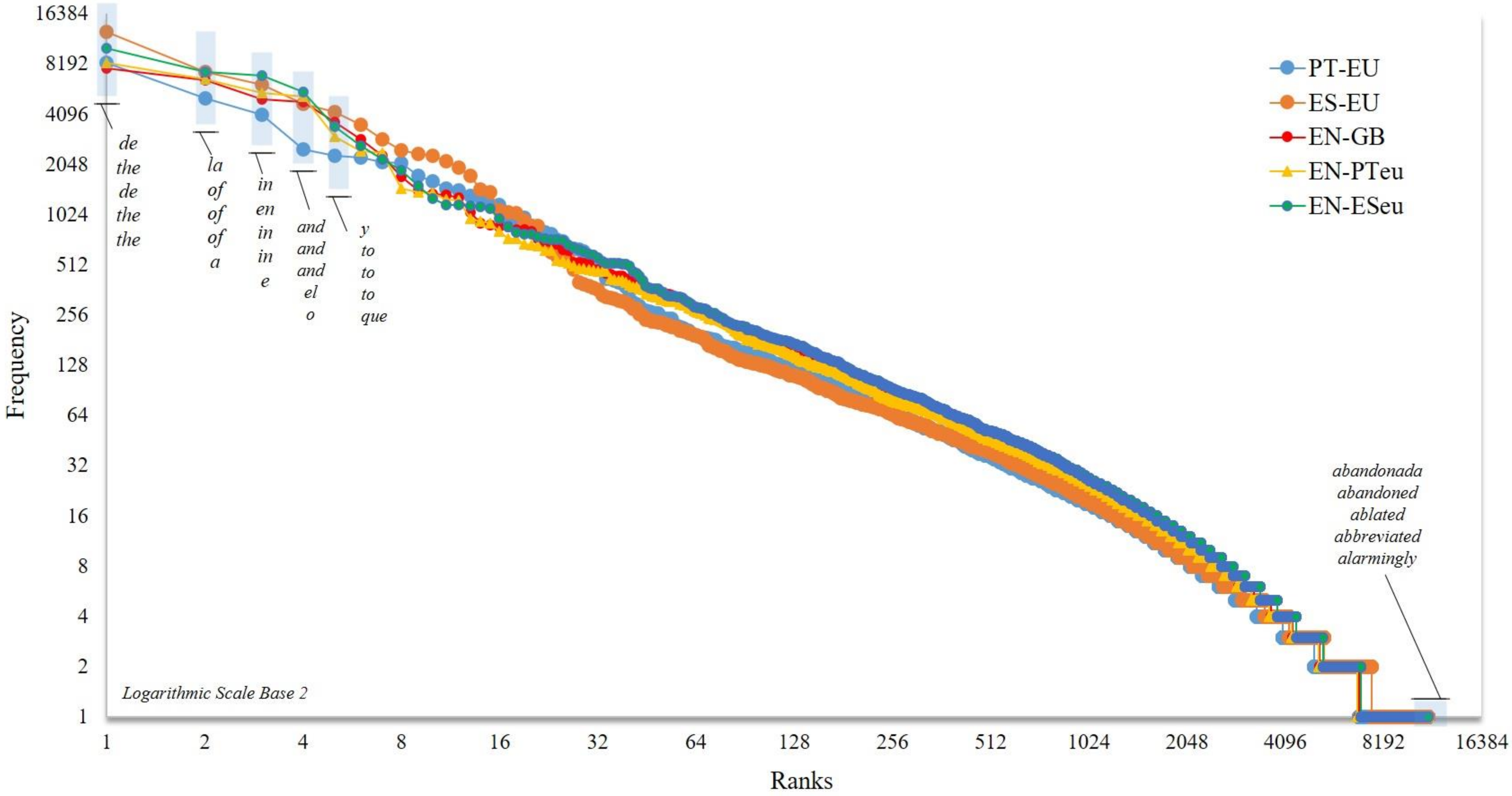


Figure 7 – Most and least frequent words in the CoRA

The most frequent nouns in the CoRA are health-related terms, which confirms the general topic of the texts, i.e., clinical health research. Below, in Table 16, are the ten most frequent nouns in each corpus, sorted by frequency. These fifty nouns correspond to 53% of all nouns in the CoRA.

Noun	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Total
[patient]			998	1416	1311	3725
[study]			728	820	951	2499
[cell]				653	1012	1665
[level]				566	812	1378
[group]			460	466	432	1358
[effect]			325	310	548	1183
[result]				399	552	951
[estudo]	931					931
[estudio]		873				873
[caso]	492	363				855
[grupo]	407	346				753
[gene]			379		364	743
[disease]			347	372		719
[mouse]					579	579
[expression]					521	521
[year]			510			510
[ano]	498					498
[datum]			465			465
[case]			444			444
[risco]	413					413
[año]		406				406
[resultado]	399					399
[valor]	347					347
[analysis]			346			346
[age]				298		298
[edad]		296				296
[table]				289		289
[população]	288					288
[idade]	285					285
[fator]	254					254
[diferencia]		252				252
[casos]		248				248
[dato]		241				241
[factor]		222				222
[nivel]		219				219

Table 16 – Most frequent nouns in the CoRA

The analyses of the data and the texts were carried out upon characterization of the corpora and are described in the following sub-sections.

4.2. Variables with no associated effects of L1 influence

The results of the comparisons run to verify effects of L1 influence in OSRAs written in English by the EN-PT^{EU} and the EN-ES^{EU} OSRA authors are presented below per variable. This section, in particular, is dedicated to variables whose quantitative results do not suggest L1 influence.

4.2.1. V2: average sentence length in words

After examining the samples of V2, the following descriptive statistics are obtained:

V2: average sentence length in words (ASLiW)							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	29,12	0,53	28,05	30,19	28,58	18,57	4,31
ES-EU	31,67	0,65	30,36	32,97	31,57	27,88	5,28
EN-GB	27,90	0,43	27,05	28,75	27,94	11,81	3,44
EN-PT^{EU}	26,75	0,59	25,57	27,93	25,49	22,55	4,75
EN-ES^{EU}	27,84	0,46	26,91	28,76	27,52	13,89	3,73

The lowest median value of the average sentence length in words is in the EN-PT^{EU} corpus and the highest in the ES-EU corpus.

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V2 (ASLiW) of the EN-PT^{EU} and EN-ES^{EU} groups are significantly different (Levene's test**,** *F* = 7.027, *p* = .009);

II. There are no statistically significant differences in the average sentence length in words between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(121.161) = 1.450, *p* = .150.
Both groups have similar average sentence length values expressed in words (*MD* = 1.09; *SED* = .749; 95% CI = 2.57 to -.397 words);

III. There are statistically significant differences in the average sentence length in words between the EN-PT^{EU} and PT-EU groups, *t*(128) = 2.979, *p* = .003.
The PT-EU OSRAs contain significantly lengthier sentences than the EN-PT^{EU} OSRAs (*MD* = 2.37 words; *SED* = .795; 95% CI = 3.943 to .795 words);

There are statistically significant differences in the average sentence length in words between the EN-ES^{EU} and ES-EU groups, *t*(115.094) = 4,777, *p* = .001.
The ES-EU group produces significantly longer sentences than the EN-ES^{EU} groups (*MD* = 3.83; *SED* = .802; 95% CI = 5.42 to 2.24 words);

IV. There are no statistically significant differences in the frequency of commas between the EN-PT^{EU} and EN-GB groups, *t*(117) = 1.579, *p* = .117.
Both groups have about the same average sentence length in words (*MD* = 1.15; *SED* = .727; 95% CI = 2.59 to -.292 words).
there are no statistically significant differences in the average sentence length in words between the EN-ES^{EU} and EN-GB groups, *t*(128) = .099, *p* = .921.
Both groups have almost identical values of average sentence length (*MD* = .062; *SED* = .629; 95% CI = 1.31 to -1.18 words).

For V2 in OSRAs written in English by the Portuguese L1 authors, no effects of L1 influence are found. Likewise, no effects of L1 influence are found for V2 in OSRAs written in English by the Spanish L1 authors.

The results indicate that the variance of the average sentence lengths in words of the OSRAs within the EN-PT^{EU} and the EN-ES^{EU} groups are significantly different. Furthermore, the average sentence lengths of the two groups are not significantly different, with the EN-PT^{EU} group writing roughly 27 words per sentence and the EN-ES^{EU} group 28 words per sentence, on average.

The L1 PT-EU and L1 ES-EU authors seem to adjust the length of sentences when producing OSRAs in English since the comparisons of V2 of the EN-PT^{EU}/ EN-ES^{EU} groups and the L1 English group (EN-GB) show that there are no significant differences between the groups, with all three having very similar means of V2, i.e., 27,90 for EN-GB; 26,75 for EN-PT^{EU}; 27,84 for EN-ES^{EU}.

When the EN-PT^{EU} and the EN-ES^{EU} groups are compared to each of their corresponding L1 counterpart groups, i.e., PT-EU and ES-EU, significant differences are observed in relation to V2, with the L1 groups producing significantly longer sentences.

Additionally, the PT-EU and ES-EU groups are also compared, and it was verified that these language groups are also significantly different with regards to V2 (*t*(128) = 3.012, *p* =

.003) and that the L1 Spanish group is the one writing the longest sentences of the two, and in general of the five groups.

In the CoRA, the average sentence length in words (V2) is fairly uniform in all three groups writing OSRAs in English and significantly different between the three language groups analyzed in this study, i.e., Portuguese, Spanish, and English. Therefore, no effects of L1 influence can be argued for the EN-PT^{EU} or the EN-ES^{EU} groups in relation to the average sentence length in words.

4.2.2. V12: number of prepositions

After examining the samples of V12, the following descriptive statistics are obtained:

V12: frequency of prepositions (all values in words/1000 tokens)							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	197,30	1,65	194,01	200,58	196,30	176,09	13,27
ES-EU	188,58	1,80	184,98	192,18	190,30	211,41	14,54
EN-GB	139,28	2,07	135,15	143,41	139,20	277,89	16,67
EN-PT^{EU}	140,34	1,49	137,36	143,33	139,50	145,20	12,05
EN-ES^{EU}	140,30	1,56	137,19	143,40	141,00	157,25	12,54

The lowest median values are in the corpora written in English, while the highest median values are found in the L1 Portuguese and Spanish corpora.

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V12 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test**,** *F* = .033, *p* = .855);

II. There are no statistically significant differences in the frequency of prepositions between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = .021, *p* = .983.
Both groups have almost identical means of prepositions per thousand words (*MD* = .046; *SED* = 2.16; 95% CI = 4.31 to -4.22 prepositions per thousand words);

III. There are statistically significant differences in frequency of prepositions between the EN-PT^{EU} and PT-EU groups, *t*(128) = 25.622, *p* = .001.

The PT-EU OSRAs contain significantly more prepositions per thousand words than the EN-PT^{EU} OSRAs (*MD* = 56.95; *SED* = 2.22; 95% CI = 61.35 to 52.56 prepositions per thousand words);

There also are statistically significant differences in the frequency of prepositions between the EN-ES^{EU} and ES-EU groups, *t*(128) = 20.277, *p* = .001.
The ES-EU OSRAs contain significantly more prepositions per thousand words than the EN-ES^{EU} OSRAs (*MD* = 48.28; *SED* = 2.38, 95% CI = 53.00 to 43.57 prepositions per thousand words);

IV. There are no statistically significant differences in the frequency of prepositions between the EN-PT^{EU} and EN-GB groups, *t*(116.539) = .417, *p* = .677.
Both groups have similar means of prepositions per thousand words (*MD* = 1.06; *SED* = 2.55; 95% CI = 6.11 to -3.98 prepositions per thousand words).

There are also no statistically significant differences in the frequency of prepositions between the EN-ES^{EU} and EN-GB groups, *t*(118.887) = .394, *p* = .695.
Both groups have very similar means of prepositions per thousand words (*MD* = 1.02; *SED* = 2.59; 95% CI = 6.14 to -4.10 prepositions per thousand words).

No differences are found in relation to the number of prepositions between any of the groups writing OSRAs in English, either the L1 or non-L1 authors.

Also, the Portuguese authors writing OSRAs in English differ significantly from Portuguese authors writing in their L1 (PT-EU), with the latter using significantly more prepositions than the former. Likewise, the Spanish authors writing OSRAs in English (EN-ES^{EU}) also differ from their Spanish counterpart writing in their L1 (ES-EU) as the ES-EU authors, like the PT-EU authors, also use significantly more prepositions per thousand words.

Additionally, a comparison between the PT-EU and ES-EU groups is run, and it showed that there also are significant differences between these two language groups (*t*(128) = 3.570, *p* = .001).

In the CoRA, V12 is a variable that behaves uniformly in all groups writing in English and significantly different in the three language groups analyzed in this study, i.e., Portuguese, Spanish, and English. Therefore, no effects of L1 influence can be argued either for the EN-PT^{EU} or the EN-ES^{EU} group in relation to the frequency of prepositions'.

The groups are linguistically examined to understand if this lack of differences reflects the use of the prepositions and if there are specific prepositions that can function as L1 influence markers.

The tag PRP, which stands for the part of speech ‘preposition’ or ‘with prepositional syntactic function’ (as tagged by the VISL system https://visl.sdu.dk/), is used to refer to all words (single and multi-word expressions) within this category.

First, the PRPs found in up to 50% of all OSRAs within each corpus are extracted together with each PRP’s total frequency, as shown below in Table 17.

Rank	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Threshold
1	**[de]** n=13439	**[de]** n=13938	**[of]** n=6467	**[of]** n=6497	**[of]** n=7175	95%
2	**[em]** n=5377	**[en]** n=5849	**[in]** n=4925	**[in]** n=5079	**[in]** n=6334	
3	**[com]** n=2233	**[a]** n=2601	**[with]** n=2306	**[with]** n=2366	**[with]** n=2144	
4	**[a]** n=2183	**[con]** n=2159	**[for]** n=1600	**[to]** n=1338	**[to]** n=1545	90%
5	**[por]** n=1190	**[por]** n=1164	**[to]** n=1525	**[for]** n=1188	**[for]** n=1198	
6	**[para]** n=953	**[para]** n=1035	**[by]** n=871	**[by]** n=909	**[by]** n=1098	
7	**[entre]** n=594	**[entre]** n=576	**[from]** n=806	**[from]** n=662	**[on]** n=654	85%
8	**[como]** n=462	**[como]** n=526	**[on]** n=719	**[on]** n=563	**[from]** n=614	80%
9	**[após]** n=173	**[durante]** n=224	**[at]** n=593	**[between]** n=479	**[at]** n=551	75-70%
10	**[sobre]** n=132	**[según]** n=218	**[between]** n=497	**[at]** n=446	**[between]** n=502	65%
11	**[sem]** n=126	**[sobre]** n=196	**[as]** n=418	**[as]** n=342	**[as]** n=432	
12	**[relativamente=a]** n=124[13]	**[sin]** n=167	**[than]** n=300	**[after]** n=271	**[after]** n=371	
13	**[durante]** n=113	**[mediante]** n=162	**[within]** n=183	**[than]** n=236	**[than]** n=329	60%
14	**[apesar=de]** n=96[14]	**[aunque]** n=149	**[after]** n=165	**[due=to]** n=167	**[during]** n=309	
15		**[tras]** n=139	**[including]** n=154	**[such=as]** n=141	**[such=as]** n=204	55%
16		**[de]** n=103	**[over]** n=138	**[without]** n=139	**[due=to]** n=128	
17		**[desde]** n=95	**[during]** n=130	**[during]** n=111	**[among]** n=120	50%
18		**[hasta]** n=94	**[without]** n=127	**[into]** n=108	**[including]** n=109	
19		**[en-cuanto-a]** n=93	**[such=as]** n=120	**[regarding]** n=102	**[through]** n=107	
20		**[respecto-a]** n=92	**[due=to]** n=104	**[through]** n=88	**[without]** n=97	
21			**[across]** n=99	**[including]** n=73	**[into]** n=96	
22			**[through]** n=99		**[according=to]** n=76	
23			**[into]** n=84			
24			**[despite]** n=72			

Table 17 – Most frequent PRPs in the CoRA (present in 50% or higher of the OSRAs) ranked by the number of occurrences and with total frequency in the corresponding corpus

[13] The tag corresponds to the main POS of the expression, i.e., the PRP ‘a’.

[14] Idem, i.e., the PRP ‘de’.

Looking at the three corpora of OSRAs written in English, it can be seen that the L1 English authors are those with the largest number of prepositions distributed in 50% of the corpus (EN-GB). They are followed by the Spanish authors writing in English (EN-ES^{EU}) and then by the Portuguese authors writing in English (EN-PT^{EU}). Similarly, in the L1 corpora, the English authors also stand out for having the largest number of prepositions distributed in 50% of the corpus, and as in the English corpora, they are followed by the Spanish authors and then by the Portuguese authors both writing in their respective L1.

After the general observations are obtained, the PRPs in the three English corpora are analyzed in terms of distribution, looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. Table 18 shows nineteen of the PRPs considered to be unlikely to function as NLID markers in OSRAs written in English by the Portuguese/Spanish authors since the percentage they occupy within their corresponding corpus and the rank are similar among the groups, despite the numbers of occurrences being higher in either the EN-GB corpus or in one or both of the non-L1 English corpora.

N[15]	PRP	Corpus	Occurrences in Corpus	Rank	% in Corpus
1	as	EN-GB EN-PT^{EU} EN-ES^{EU}	418 342 432	11 11 11	0.24 0.21 0.23
2	including	EN-GB EN-PT^{EU} EN-ES^{EU}	154 73 109	15 18 21	0.09 0.04 0.06
3	among	EN-GB EN-PT^{EU} EN-ES^{EU}	136 100 120	18 21 17	0.08 0.06 0.07
4	into	EN-GB EN-PT^{EU} EN-ES^{EU}	84 96 108	23 18 21	0,05 0,05 0,07
5	of	EN-GB EN-PT^{EU} EN-ES^{EU}	6467 6497 7175	1 1 1	3,78 3,98 3,89
6	through	EN-GB EN-PT^{EU} EN-ES^{EU}	99 88 107	22 20 19	0,06 0,05 0,06
7	between	EN-GB EN-PT^{EU} EN-ES^{EU}	497 479 502	10 9 10	0,29 0,29 0,27

[15] The numeration in this and subsequent tables does not follow any particular criterion.

8	than	EN-GB	300	12	0,18
		EN-PT^{EU}	236	13	0,14
		EN-ES^{EU}	329	13	0,18
9	to	EN-GB	1525	5	0,89
		EN-PT^{EU}	1338	4	0,82
		EN-ES^{EU}	1545	4	0,84
10	despite	EN-GB	72	24	0,04
		EN-PT^{EU}	72	28	0,04
		EN-ES^{EU}	42	32	0,02
11	with	EN-GB	2306	3	1,35
		EN-PT^{EU}	2366	3	1,45
		EN-ES^{EU}	2144	3	1,16
12	without	EN-GB	127	18	0,07
		EN-PT^{EU}	97	16	0,05
		EN-ES^{EU}	139	20	0,09
13	due=to	EN-GB	104	23	0,06
		EN-PT^{EU}	167	14	0,10
		EN-ES^{EU}	128	17	0,07
14	such=as / as	EN-GB	538	11	0,31
		EN-PT^{EU}	483	11	0,30
		EN-ES^{EU}	636	11	0,35
15	at	EN-GB	593	11	0.35
		EN-PT^{EU}	446	12	0.27
		EN-ES^{EU}	551	11	0.30
16	from	EN-GB	806	7	0.47
		EN-PT^{EU}	614	7	0.33
		EN-ES^{EU}	662	8	0.41
17	on	EN-GB	719	8	0.42
		EN-PT^{EU}	563	8	0.34
		EN-ES^{EU}	654	7	0.35
18	in	EN-GB	4925	2	2,88
		EN-PT^{EU}	5079	2	3,11
		EN-ES^{EU}	6334	2	3,44
19	by	EN-GB	871	6	0,51
		EN-PT^{EU}	909	6	0,56
		EN-ES^{EU}	1098	6	0,60

Table 18 – PRPs unlikely to function as NLID markers in OSRAs written in English by Portuguese/Spanish authors, given their similar distribution in each corpus.

After this group is excluded, two groups of PRPs are analyzed. One is examined to verify strategies of avoidance by the non-L1 English authors. The other group is examined to verify if there are PRPs that could function as NLID markers in OSRAs written in English by the Portuguese/Spanish authors.

The first group comprises four PRPs with higher numbers of (a) occurrences, (b) ranks, and (c) percentages in the EN-GB corpus and whose analysis is important to exclude potential strategies of avoidance of use. These PRPs are shown below in Table 19.

N	Preposition	Corpus	Occurrences in Corpus	Rank	% in Corpus
1	for	EN-GB EN-PT^{EU} EN-ES^{EU}	1600 1188 1198	6 7 7	0.93 0.73 0.65
2	across	EN-GB EN-PT^{EU} EN-ES^{EU}	99 16 15	21 53 51	0.06 0.01 0.01
3	over	EN-GB EN-PT^{EU} EN-ES^{EU}	138 42 38	16 31 35	0.08 0.03 0.02
4	within	EN-GB EN-PT^{EU} EN-ES^{EU}	183 79 43	15 23 31	0.11 0.05 0.02

Table 19 – PRPs with a higher number of occurrences, higher or similar ranks, and higher percentages in the EN-GB corpus.

The frequencies of the PRPs in Table 19 are compared to see if there are significant differences between the groups. For all PRPs the fourth L1 effect of the unified framework (Jarvis 2010, 2000) is tested for both the EN-PT^{EU} and EN-ES^{EU} OSRAs, stated as follows:

Effect of L1 Influence	L1 influence EN-PT^{EU} question	L1 influence EN-ES^{EU} question
IV) Intralingual contrast	Are the frequencies of the PRP [for]/[across]/[over]/[within] in the EN-PT^{EU}/ EN-ES^{EU} and the EN-GB corpora statistically significantly different?	

Given that the data is not normally distributed and has some outliers, the Mann-Whitney test is used to assess for a mean difference between the groups. The level of significance used is p < .05. Eight tests are carried out. The number of occurrences of all prepositions is normalized by 100. Table 20 shows the results obtained.

	IV - Intralingual contrast (Mann-Whitney Test)	Effect IV of L1 influence found for EN-PT^{EU}?	IV - Intralingual contrast (Mann-Whitney Test)	Effect IV of L1 influence found for EN-ES^{EU}?
	corpora examined			
Preposition	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. EN-GB	
	***p* reference value < .05**			
for	*Z* = -3.369 *p* = .001 *M* rank EN-GB = 76.62 *M* rank EN-PT^{EU} = 54.38	*yes*	*Z* = -3.070 *p* = .002 *M* rank EN-GB= 75.64 *M* rank EN-ES^{EU}= 55.36	*yes*
across	*Z* = -1.610 *p* = .107 *M* rank EN-GB= 26.12 *M* rank EN-PT^{EU}= 19.05	*no*	*Z* = -2.006 *p* = .045 *M* rank EN-GB= 26.50 *M* rank EN-ES^{EU}= 17.77	*yes*
over	*Z* = -1.876 *p* = .061 *M* rank EN-GB= 36.06 *M* rank EN-PT^{EU}= 27.41	*no*	*Z* = -2.956 *p* = .003 *M* rank EN-GB= 40.27 *M* rank EN-ES^{EU}= 28.80	*yes*
within	*Z* = -2.180 *p* = .029 *M* rank EN-GB= 47.59 *M* rank EN-PT^{EU}= 36.12	*yes*	*Z* = -3.704 *p* = .001 *M* rank EN-GB= 48.04 *M* rank EN-ES^{EU}= 29.03	*yes*

Table 20 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups in relation to the 4 PRPs with potential to mark strategies of avoidance.

The results of the Mann-Whitney's tests indicate statistically significant differences in the ranked number of occurrences between the EN-GB and the EN-ES^{EU} OSRAs, with the EN-GB OSRAs having a greater ranked number of occurrences of the PRPs "for", "across", "over" and "within" than the EN-ES^{EU} OSRAs. As for the EN-PT^{EU} OSRAs, the results of the Mann-Whitney's tests show statistically significant differences in the ranked number of occurrences of the PRPs "for" and "within", with the EN-GB OSRAs having a greater ranked number of occurrences of those prepositions. However, no statistically significant differences are found between the mean ranks of the number of occurrences of the prepositions "across" and "over" between the EN-GB and the EN-PT^{EU} OSRAs.

Based on the significance of the results, the PRPs "for" and "within" are further analyzed for both groups (the EN-PT^{EU} and the EN-ES^{EU}), while the PRPs "across" and "over" are analyzed only for the EN-ES^{EU} group. The analyses are based on the concordances of the parsed files. The concordances were obtained with WordSmith 7.0 (Scott 2018b), from which the syntactic structures containing the PRPs are extracted.

The significant differences in the number of occurrences of the PRP [for] between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups are obvious in the syntactic structures shown in Table 21, which are more frequent in the EN-GB than in the other two groups.

PRP	Syntactic structure of word following the PRP "for"	Example	EN-GB	EN-PT^{EU}	EN-ES^{EU}
for	N P NOM	for patients/children/adults	191	73	89
	ADJ POS @>N	for long-term/serious/fast	84	44	53
	ART S @>N	for a [cohort]/[restriction]	70	36	27
	NUM P @>N	for two/four/2030	59	25	19
	DET S @>N	for this/that/each/any	59	29	76
	DET P @>N	for these/all/both/some	78	49	43
	ADV @FOC>	for both	25	6	7
	INDP P @P	for these/those	22	4	5
	V PCP1 @ICL-P	for testing/treating/defining	29	13	33
	ADJ POS	for acute/high/persistent	121	109	99
	INDP S @P	for this/that/each	15	5	4

Table 21 – Syntactic structures that support the significant differences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups in relation to the preposition "for"

The significant differences in the number of occurrences of the PRP [within] between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups are evident in the syntactic structures shown in Table 22.

PRP	Syntactic structure of word following the PRP "within"	Example	EN-GB	EN-PT^{EU}	EN-ES^{EU}
within	ART S/P @>N	within the body/routine	63	39	21
	NUM P @>N	within 1420 [kpb]/18 [days]	22	2	4
	DET S @>N	within each stratum/this stem	20	4	1
	ART S @>N	within a year/transcription	14	6	4
	NUM P @P	within 104 hours/5560 min	11	0	0
	N S NOM	within Europe/Bacteroidetes	7	0	1

Table 22 – Syntactic structures that support the significant differences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups in relation to the preposition "within"

The significant differences in the number of occurrences of the PRP [across] between the EN-GB and the EN-ES^{EU} groups are justified by most of the syntactic structures associated with that PRP, shown in Table 23 below.

PRP	Syntactic structure of word following the PRP "across"	Example	EN-GB	EN-ES^{EU}
across	ART S/P @>N	across the [study/groups/scheme]	42	4
	DET P @>N	across all ages/these groups	16	0
	ADJ POS @>N	across various/diverse [N]	10	1
	NUM P @>N	across 12 months/four categories	7	0
	N S NOM @P	across Europe/Wales	6	0
	N S NOM @>N	across treatment/fracture	5	0
	DET P @PN	across these/all	2	0
	V PCP1 @ICL-P	across waking hours/ageing British	2	0
	ADV @FOC>	across both studies	1	0
	V PCP2 STA	across repeated [N]	1	0

Table 23 – Syntactic structures that support the significant differences between the EN-GB and the EN-ES^{EU} groups in relation to the PRP "across"

As shown in Table 24, the significant differences in the number of occurrences of the PRP [over] between the EN-GB and the EN-ES^{EU} groups are justified by all the syntactic structures associated with that PRP.

PRP	Syntactic structure of word following the PRP "over"	Example	EN-GB	EN-ES^{EU}
over	ART S/P @>N	over the course/year/time	50	19
	N S NOM @P	over time/count/Ukraine	20	8
	ART S @>N	over a [period/one month]	14	5
	ADJ POS @>N	over long/different/recent [N]	12	2
	NUM P @>N	over 50 years/13420 million	13	1
	DET P @>N	over other/a=number=of/all	4	2
	N S NOM @>N	over CVD	2	1
	KC @CO	over and above	6	0
	KS @SUB	over whether	4	0
	ADJ POS @P	over made	4	0
	DET S @>N	over this age/period/year	3	0
	INDP S/P @P	over half/which	3	0
	N P NOM @P	over models	1	0
	NUM S @>N	over 29 years/26 hours	1	0

Table 24 - Syntactic structures that support the significant differences between the EN-GB and the EN-ES^{EU} groups in relation to the PRP "over"

The second group of PRPs analyzed comprises PRPs whose numbers of (a) occurrences, (b) ranks, and (c) percentages of occurrences are higher in both or one of the non-L1 groups (EN-PT^{EU} and EN-ES^{EU}) than in the L1 (EN-GB) and therefore, could work as NLID markers of language transfer if similar occurrences, ranks, and percentages are verified in the L1 PT-EU and ES-EU corpora. Table 25 shows such PRPs.

N	PRP	Corpus	Occurrences in corpus	Rank	% in corpus
1	after	EN-GB	165	16	0,10
		EN-PT^{EU}	271	14	0,17
		EN-ES^{EU}	371	14	0,20
2	during	EN-GB	130	17	0,08
		EN-PT^{EU}	111	17	0,07
		EN-ES^{EU}	309	16	0,17
3	regarding	EN-GB	29	41	0,02
		EN-PT^{EU}	102	19	0,06
		EN-ES^{EU}	52	29	0,03
4	according to	EN-GB	45	32	0,03
		EN-PT^{EU}	74	25	0,05
		EN-ES^{EU}	76	22	0,04

Table 25 – PRPs that could function as NLID markers of language transfer

Since these PRPs are chosen as possible markers of language transfer, their equivalents in the L1 Portuguese/Spanish corpora are extracted, and their frequencies are compared. Table 26 shows two of those PRPs and their equivalents in Portuguese/Spanish as found in the corresponding L1 corpora.

N	PRP	Corpus	Occurrences in corpus	Rank	% in corp us
1	após	PT-EU	173	9	0,12
	trás	ES-EU	139	16	0,09
	after	EN-GB	165	16	0,10
		EN-PT^{EU}	271	14	0,17
		EN-ES^{EU}	371	14	0,20
2	durante	PT-EU	136	13	0,09
	durante	ES-EU	224	9	0,14
	during	EN-GB	130	17	0,08
		EN-PT^{EU}	111	17	0,07
		EN-ES^{EU}	309	16	0,17

Table 26 – Prepositions/prepositional expressions that may function as NLID markers in OSRAs written in English by the L1 Portuguese/Spanish authors

The other two PRPs, i.e., “regarding” and “according to”, are analyzed as groups of expressions since their translation into Portuguese/Spanish may have more than one equivalent, and in fact, more than one equivalent is found in the PT-EU and ES-EU L1 corpora. Table 27 shows the groups of equivalent expressions with a prepositional function within the “regarding” and “according=to” groups found in the CoRA.

N	PRP	Corpus	Occurrences in corpus	Rank	% in corpus
3	sobre / relativamente=a / em=relação=a / quanto=a / acerca=de / em=torno=de /	PT-EU	446	10	0,31
	sobre /en=cuanto=a / respecto=a / en=relación=com / con=respecto=a / acerca=de / respecto=de / con=relación=a / en=lo=referente=a / en=relación=con	ES-EU	460	11	0,28
	about / regarding / relative=to / in=relation=to / as=to / with=respect=to / concerning / as=for	EN-GB	157	28	0,09
	regarding / about / concerning / relative=to / with=respect=to / with=regard=to / in=relation=to / in=regard=to	EN-PT^{EU}	191	19	0,12
	regarding / about / with=respect=to / concerning / in=relation=to / relative=to / as=for / in=regard=to	EN-ES^{EU}	164	29	0,09
4	segundo / consoante / em=função=de / conforme	PT-EU	221	21	0,15
	[según / en=función=de / de=acuerdo=con / conforme=a / conforme=con	ES-EU	293	10	0,18
	according=to	EN-GB	45	32	0,03
	according=to / in=accordance=with / in=line=with	EN-PT^{EU}	109	25	0,07
	according=to / in=line=with / in=accordance=with	EN-ES^{EU}	101	22	0,05

Table 27 – PRPs analyzed as groups since their translation into Portuguese/Spanish may have more than one equivalent in PT-EU/ES-EU corpora

The frequencies of these PRPs, i.e., “after”, “during”, “regarding”, and “according=to” and equivalents, are compared to examine significant differences between the groups. For all PRPs, the following questions are asked for both the EN-PT^{EU} and the EN-ES^{EU} corpora:

Effect of L1 Influence	**L1 influence** **EN-PT^{EU} questions**	**L1 influence** **EN-ES^{EU} questions**
I) Intragroup homogeneity	Are the frequencies of the PRP in the EN-PT^{EU} / EN-ES^{EU} OSRAs uniformly distributed?	
II) Intergroup heterogeneity	Are the frequencies of the PRP in the EN-PT^{EU} and EN-ES^{EU} OSRAs statistically significantly different?	
III) Cross-language congruity	Are the frequencies of the PRP in the EN-PT^{EU} and PT-EU OSRAs statistically similar?	Are the frequencies of the PRP in the EN-ES^{EU} and ES-EU OSRAs statistically similar?
IV) Intralingual contrast	Are the frequencies of the PRP in the EN-PT^{EU} and EN-GB OSRAs statistically significantly different?	Are the frequencies of the PRP in the EN-ES^{EU} and EN-GB OSRAs statistically significantly different?

The Mann-Whitney test is used to assess for a mean difference between the groups given that the data is not normally distributed and has some outliers. As previously explained (chapter 3), questions I and II are answered together. The level of significance used is $p < .05$ for questions II and IV. Because questions I and III look for uniformity and congruity, respectively, a result of $p > .05$ is associated with an L1 effect.

Table 28 below shows the results and mean ranks of all comparisons. As can be seen in Table 28, only one PRP, i.e., “durante”/“during”/“during”, may mark L1 influence in OSRAs written in English by the Spanish authors. However, none of the PRPs are found to mark L1 influence in the Portuguese authors writing in English.

	L1 Influence Effects								
	I) Intra-L1 homogeneity (Levene's test)	II) Inter-L1 heterogeneity (Mann-Whitney Test)	EN-PT^{EU} and EN-ES^{EU} similar in variance but different in means?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for EN-PT^{EU}?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for EN-ES^{EU}?
	Corpora Examined			Corpora Examined			Corpora Examined		
PRP	EN-PT^{EU} vs. EN-ES^{EU}	EN-PT^{EU} vs. EN-ES^{EU}		EN-PT^{EU} vs. PT-EU	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. ES-EU	EN-ES^{EU} vs. EN-GB	
	Reference *p* values			Reference *p* values			Reference *p* values		
	p > .05 **AND**	*p* < .05?		*p* > .05?	*p* < .05?		*p* > .05?	*p* < .05?	
após / tras / after	*F* = .839 *p* = .362	*Z* = -1.799 *p* = .072 Mean ranks: EN-PT^{EU}: 49.97 EN-ES^{EU}: 60.83	*no*	*Z* = -1.268 *p* = .205 Mean ranks: EN-PT^{EU}: 53.83 PT-EU: 46.59	*Z* = -1.071 *p* = .284 Mean ranks: EN-PT^{EU}: 54.95 EN-GB: 48.74	*no*	*Z* = -4.008 *p* = .001 Mean ranks: EN-ES^{EU}: 58.01 ES-EU: 35.19	*Z* = -3.113 *p* = .002 Mean ranks: EN-ES^{EU}: 61.58 EN-GB: 43.19	*no*
durante / durante / during	*F* = 23.416 *p* = .001	*Z* = -3.352 *p* = .001 Mean ranks: EN-PT^{EU}: 33.82 EN-ES^{EU}: 51.90	*no*	*Z* = -1.396 *p* = .163 Mean ranks: EN-PT^{EU}: 34.62 PT-EU: 41.47	*Z* = -.328 *p* = .743 Mean ranks: EN-PT^{EU}: 39.63 EN-GB: 41.29	*no*	*Z* = -1.426 *p* = .154 Mean ranks: EN-ES^{EU}: 53.56 ES-EU: 45.44	*Z* = -3.209 *p* = .001 Mean ranks: EN-ES^{EU}: 54.13 EN-GB: 36.51	*yes*
regarding and equivalents with prepositional function extracted from all the corpora (see Table 27)	*F* = 1.941 *p* = .167	*Z* = -.383 *p* = .702 Mean ranks: EN-PT^{EU}: 54.60 EN-ES^{EU}: 52.36	*no*	*Z* = -4.067 *p* = .001 Mean ranks: EN-PT^{EU}: 43.99 PT-EU: 68.91	*Z* = -.011 *p* = .991 Mean ranks: EN-PT^{EU}: 48.53 EN-GB: 48.46	*no*	*Z* = -5.288 *p* = .001 Mean ranks: EN-ES^{EU}: 39.75 ES-EU: 72.39	*Z* = -.321 *p* = .748 Mean ranks: EN-ES^{EU}: 46.70 EN-GB: 48.49	*no*
according=to and equivalents with prepositional function extracted from all the corpora (see Table 27)	*F* = .664 *p* = .417	*Z* = -1.749 *p* = .080 Mean ranks: EN-PT^{EU}: 50.29 EN-ES^{EU}: 41.31	*no*	*Z* = -2.184 *p* = .029 Mean ranks: EN-PT^{EU}: 43.12 PT-EU: 32.79	*Z* = -1.289 *p* = .198 Mean ranks: EN-PT^{EU}: 34.55 EN-GB: 28.59	*no*	*Z* = -5.710 *p* = .001 Mean ranks: EN-ES^{EU}: 33.90 ES-EU: 66.49	*Z* = -.141 *p* = .888 Mean ranks: EN-ES^{EU}: 35.29 EN-GB: 35.95	*no*

Table 28 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups in relation to the two groups of expressions used as PRPs with the potential to function as NLID marker

The PRPs [during]/[durante]/[durante] are similarly distributed across and within the PT-EU, the EN-GB, and EN-PT^{EU} corpora and significantly differently distributed in the ES-EU, the EN-GB, and the EN-ES^{EU} (see Figure 8). That is, between 57% and 65% of the OSRAs in the PT-EU, the EN-GB, and EN-PT^{EU} corpora contain "during"/"durante". Additionally, the number of occurrences is not significantly different between those corpora. However, the PRPs "during"/"durante" are found in 75% of the ES-EU and the EN-ES^{EU} corpora but only in 65% of the EN-GB corpus. Likewise, the number of occurrences is similar between the ES-EU and the EN-ES^{EU} corpora but significantly different between the EN-GB and the EN-ES^{EU} corpora.

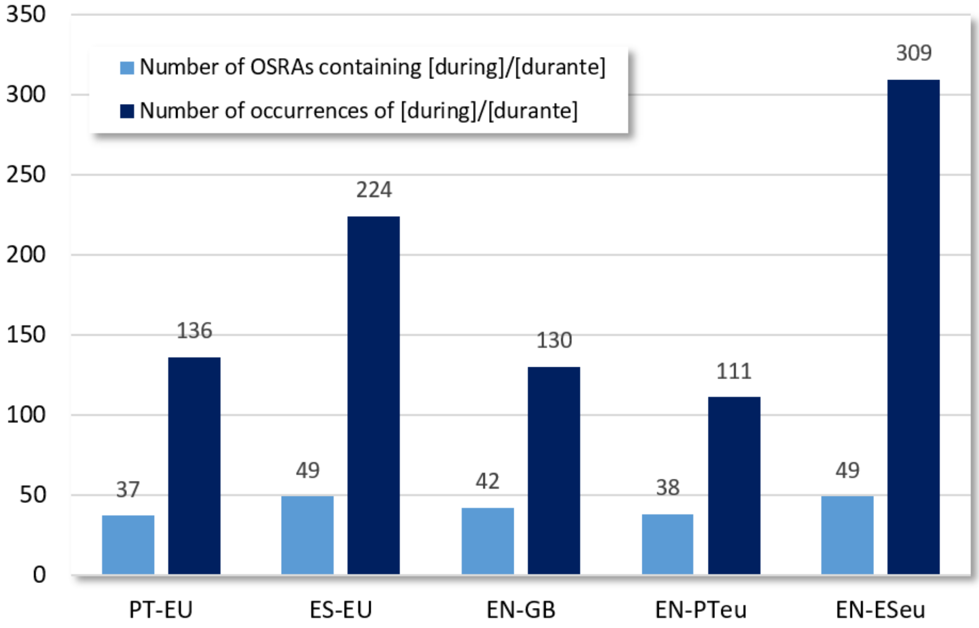


Figure 8 – Distribution of the PRPs "during"/"durante"/"durante"

Based on the significance of the results, the PRP "during" and its Spanish equivalent "durante" are further analyzed.

The significant differences found between the EN-GB and the EN-ES^{EU} groups are most evident in the syntactic structures shown in Table 29.

PRP	Syntactic structure of word following the PRP "during"	Example	EN-GB	EN-ES^{EU}
during	ART S/P @>N	during the scan/study/survey	46	111
	N S NOM @P	during follow-up/treatment	21	57
	ADJ POS @>N	during normal/early/extensive [N]	7	57
	N S NOM @>N	during DNA biding/embryogenesis	16	41

Table 29 – Syntactic structures with the number of occurrences in the EN-GB and the EN-ES^{EU} corpora that support the significant differences between those groups in relation to the PRP "during"

The syntactic structures following the PRP "during" with the largest differences between the EN-GB and the EN-ES^{EU} groups are [ART S/P @>N] and [ADJ POS @>N].

The equivalent of the structure [ART S/P @>N] in Spanish has the same syntactic order and is equally very frequently found in the ES-EU corpus of the CoRA, i.e., [durante + DET + N] with gender and number variations.

The Spanish equivalent of the second structure usually follows the order [durante + DET + N + ADJ] also with gender and number variations, e.g., "durante la inyección intracoronaria"/"durante el horizonte temporal"/"durante las fases iniciales"/"durante un tiempo máximo". A few other cases maintain the adjective (ADJ) before the noun (N), as in "durante los primeros años de vida". However, as can be seen in the previous examples, either structure will usually be accompanied by a determiner (DET).

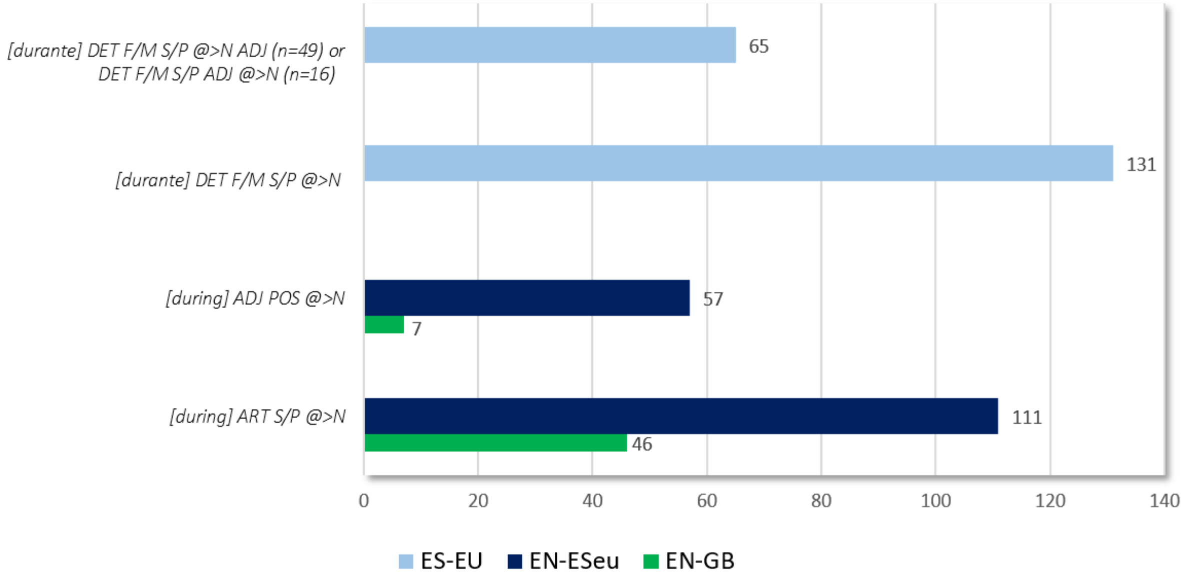


Figure 9 – Syntactic structures of "during"/"durante"/"durante"

The exception to using a determiner (DET) in syntactic structures like those shown above, i.e., (durante + DET + N + ADJ) are cases, like (a) and (b) taken from the CoRA, that can drop the determiner and maintain grammaticality.

a) A pesar de que los 3 pedúnculos cerebelosos convergen en las paredes laterales y el techo del IV ventrículo, la colindancia directa de los pedúnculos cerebelosos superiores e inferiores con el interior de la cavidad del IV ventrículo les confieren mayor riesgo de lesionarse ***durante abordajes quirúrgicos*** *en esta región. [ES-EU_OSRA_048]*

b) Se seleccionaron aquellas situaciones que presumiblemente por motivo de consulta, situación clínica u orientación diagnóstica podrían llevar al menos una observación del paciente ***durante periodos superiores*** *a 12 horas para la monitorización de tratamientos y seguimiento de la enfermedad, y/o ingreso hospitalario. [ES-EU_OSRA_035]*

The other syntactic structures found in the EN-GB and the EN-ES^{EU} corpora (n=17) do not differ greatly with regards to the number of occurrences in either corpus. However, it is worth noting that some structures are only found in OSRAs produced by the L1 English authors (i.e., 7 and 9-12 in Table 30) and other structures are present only in OSRAs authored by the Spanish authors writing in English (i.e., 8 and 13-17, Table 30).

N	PRP	Syntactic structure of word following the PRP "during"	Example	EN-GB	EN-ES^{EU}
1	"during"	V PCP1 @ICL-P	during ageing/labelling	5	10
2		NUM P @>N	during 2010/three months	2	9
3		DET S @>N	during this process/period	7	3
4		ART S @>N	during a setup/an acute	6	3
5		KC @CO	during and [after]	4	2
6		N P NOM @P	during times/periods	2	3
7		PERS 3P GEN @>	during their	4	0
8		PERS NEU 3S GEN	during its	0	3
9		NUM P @P	during 1992–1995	2	0
10		NUM S S @P	during 2010	2	0
11		NUM S S S @P	during 2003	1	0
12		PERS 3P GEN @SUBJ>	during their	1	0
13		DET P @>N	during these [periods]	0	1
14		INDP P @P	during which	0	1
15		INDP S/P @P	during which	0	1
16		NUM @>N	during first [exposure]	0	1
17		PREF @>N	during ex-vivo	0	1

Table 30 – Syntactic structures in the EN-GB and the EN-ES^{EU} corpora with similar distributions in relation to the number of occurrences

As has been shown, there are no significant differences between the groups in the CoRA in relation to the frequency of use of PRPs in general. The EN-GB group has the highest number of PRPs distributed in at least 50% of the OSRAs in the CoRA. When the frequencies of the PRP distributed across 50% or more of the OSRAs in the CoRA are analyzed, significant differences are found in the frequency of use of the PRPs "for" and "within" between the British authors and both the Portuguese and the Spanish authors writing in English, with the British using these prepositions significantly more frequently. Upon comparison, the British authors are also found to use the prepositions "across" and "over" significantly more frequently than the Spanish authors writing in English, while the Portuguese authors use "across" and "over" as frequently as British authors. These results could indicate that the Spanish authors avoid using the prepositions "for", "within", "across", and "over", and the Portuguese authors avoid using the prepositions "for" and "within". Since the significant differences found between the L1 and the non-L1 English authors are observed under conditions of genre and register constraints and discourse community standards, the PRPs "for", "within", "across", and "over" may be useful in detecting non-nativeness in scientific writing in English, especially when used in phrases with the syntactic structures described above.

Finally, the results show that the preposition "during" is used significantly more frequently by the Spanish authors writing in English than by the British authors and the Portuguese authors writing in English. Furthermore, the Spanish authors use the preposition "durante" when writing in their L1 as frequently as the Spanish authors use "during" when writing in English, which could indicate language transfer associated with "during" and its equivalent "durante".

4.2.3. V14: number of relative pronouns

After examining the samples of V14, the following descriptive statistics are obtained:

V14: frequency of relative pronouns (all values in words/1000 tokens)							
CORPORA	Mean	Std. Error	95% Confidence Interval for Mean		Median	Variance	Std. Deviation
			Lower Bound	Upper Bound			
PT-EU	15,12	0,52	14,07	16,16	15,40	17,72	4,21
ES-EU	13,80	0,55	12,69	14,90	13,20	19,89	4,46
EN-GB	8,42	0,38	7,66	9,19	8,50	9,49	3,08
EN-PT^{EU}	7,65	0,36	6,93	8,37	7,30	8,47	2,91
EN-ES^{EU}	7,74	0,37	7,00	8,47	7,70	8,82	2,97

The frequency of relative pronouns is lower in all the English corpora, either L1 or non-L1 than in the Portuguese and the Spanish corpora. Portuguese writing OSRAs in their L1 are those who use relative pronouns more frequently, but when they write OSRAs in English, they become the group with the least number of relative pronouns.

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V14 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test**,** *F* = .043, *p* = .836);

II. There are no statistically significant differences in the frequency of relative pronouns between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = .164, *p* = .870.
Both groups have similar means of relative pronouns per thousand words (*MD* = .085; *SED* = .516; 95% CI = 1.11 to -.936 relative pronouns per thousand words);

III. There are statistically significant differences in the frequency of relative pronouns between the EN-PT^{EU} and PT-EU groups, *t*(113.859) = 11.758, *p* = .001.
The PT-EU OSRAs contain significantly more relative pronouns per thousand words than the EN-PT^{EU} OSRAs (*MD* = 7.46; *SED* = .635; 95% CI = 8.72 to 6.21 relative pronouns per thousand words);

There also are statistically significant differences in the frequency of relative pronouns between the EN-ES^{EU} and ES-EU groups, *t*(111.420) = 9.123, *p* = .001.

The ES-EU OSRAs contain significantly more relative pronouns per thousand words than the EN-ES^{EU} OSRAs (*MD* = 6.06; *SED* = .664; 95% CI = 7.38 to 4.74 relative pronouns per thousand words);

IV. There are no statistically significant differences in the frequency of relative pronouns between the EN-PT^{EU} and EN-GB groups, *t*(128) = 1.471, *p* = .144. Both groups of OSRAs have similar means of relative pronouns per thousand words (*MD* = .773; *SED* = .526; 95% CI = 1.81 to -.267 relative pronouns per thousand words).

There are also no statistically significant differences in the frequency of relative pronouns between the EN-ES^{EU} and EN-GB groups, *t*(128) = 1.299, *p* = .196. Both groups have similar means of relative pronouns per thousand words (*MD* = .689; *SE* = .531; 95% CI = 1.74 to -.361 relative pronouns per thousand words).

For V14 no effects of L1 influence are found in the EN-PT^{EU} or the EN-ES^{EU} groups concerning the frequency of relative pronouns. That is, no differences are found between the L1 PT-EU and L1 ES-EU groups writing OSRAs in English and between these and the EN-GB group writing in English. Furthermore, the EN-PT^{EU} and the EN-ES^{EU} groups differ significantly from their L1 PT-EU and L1 ES-EU counterparts writing in their respective L1 who use significantly more relative pronouns per thousand words.

The frequency of relative pronouns is not significantly different between the PT-EU and ES-EU groups when writing in their respective L1s. Nevertheless, the frequency of relative pronouns is uniform between all groups writing in English and is significantly different in the PT-EU and the ES-EU groups writing in English and the PT-EU and ES-EU groups writing in their respective L1s. Hence, no effects of L1 influence are argued either for the EN-PT^{EU} or the EN-ES^{EU} group in relation to the frequency of relative pronouns.

The groups are linguistically examined to understand if such a lack of differences reflects the use of relative pronouns and if there are specific relative pronouns that can function as L1 influence markers. For that, the relative pronouns found in at least 50% of all OSRAs within each corpus are extracted together with the total frequency of each relative pronoun, as shown below in Table 31.

Rank	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Threshold
1	[que] n=1070	[que] n=1154	[which] n=431	[which] n=407	[which] n=457	95%
2	[como] n=526	[como] n=244	[that] n=343	[that] n=356	[that] n=436	
3	[o=que] n=140	[el=que] n=207	[who] n=281	[when] n=216	[when] n=185	90-85%
4	[quando] n=106	[lo=que] n=153	[when] n=143	[as] n=140	[as] n=181	80%
5	[o=qual] n=82	[cuando] n=122	[as] n=116	[who] n=109	[who] n=101	75%
6	[bem=como] n=61	[según] n=109	[where] n=115		[where] n=64	70-60%
7		[así=como] n=86				50%

Table 31 – Most frequent relative pronouns in the CoRA (present in 50% or more of the OSRAs) ranked by the number of occurrences and with total frequency in the corresponding corpus

After examination of the groups, it can be seen that all the corpora have a similar number of relative pronouns present in at least 50% of the OSRAs. Additionally, the relative pronouns in the English groups are all the same, except for [where] that is below the 50% threshold in the EN-PT^{EU} group.

After the general observations are made, the relative pronouns in the three English corpora are analyzed in terms of distribution, looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. After such an analysis, all relative pronouns are considered to be unlikely to function as NLID markers in OSRAs written in English by the Portuguese/Spanish authors since the percentage they occupy within each of their corresponding corpus and their ranks are similar, even though some numbers of occurrences are higher in either the EN-GB corpus or in one or both of the non-L1 English corpora.

N	Relative Pronoun	Corpus	Occurrences in corpus	Rank	% in corpus
1	which	EN-GB EN-PT^{EU} EN-ES^{EU}	431 407 457	1 1 1	0,25 0,25 0,25
2	that	EN-GB EN-PT^{EU} EN-ES^{EU}	343 356 436	2 2 2	0,20 0,22 0,24
3	who	EN-GB EN-PT^{EU} EN-ES^{EU}	281 109 101	3 5 5	0,16 0,07 0,05
4	when	EN-GB EN-PT^{EU} EN-ES^{EU}	143 216 185	4 3 3	0,08 0,13 0,10
5	where	EN-GB EN-PT^{EU} EN-ES^{EU}	115 43 64	6 8 6	0,07 0,03 0,03
6	as	EN-GB EN-PT^{EU} EN-ES^{EU}	116 140 181	5 4 4	0,07 0,09 0,10

Table 32 – Relative pronouns considered to be unlikely to function as NLID markers in OSRAs written in English by the Portuguese/Spanish authors given their similar ranks and percentage in the corresponding corpus

No further analyses are carried out. Relative pronouns do not appear to have the potential to function as NLID markers in relation to their overall frequency in the CoRA or distribution within the corresponding corpus.

4.2.4. V15: number of personal pronouns

After examining the samples of V15, the following descriptive statistics are obtained:

V15: frequency of personal pronouns (all values in words/1000 tokens)							
CORPORA	Mean	Std. Error	95% Confidence Interval for Mean		Median	Variance	Std. Deviation
			Lower Bound	Upper Bound			
PT-EU	8,18	0,55	7,07	9,28	7,73	19,80	4,45
ES-EU	18,90	0,58	17,73	20,07	18,64	22,18	4,71
EN-GB	12,46	0,92	10,62	14,29	10,96	54,61	7,39
EN-PT^{EU}	12,69	0,52	11,65	13,72	13,14	17,39	4,17
EN-ES^{EU}	13,19	0,55	12,09	14,28	12,71	19,62	4,43

The frequency of personal pronouns is very similar in all the English groups. The group with the lowest mean of personal pronouns is the PT-EU, while the ES-EU is the group that most frequently uses personal pronouns.

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V15 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test**,** *F* = .033, *p* = .856);

II. There are no statistically significant differences in the frequency of personal pronouns between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = .663, *p* = .509.
Both groups have almost identical means of personal pronouns per thousand words (*MD* = .50; *SED* = .755; 95% CI = 1.99 to -.993 personal pronouns per thousand words);

III. There are statistically significant differences in the frequency of personal pronouns between the EN-PT^{EU} and PT-EU groups, *t*(128) = 5.965, *p* = .001.
The EN-PT^{EU} OSRAs contain significantly more personal pronouns per thousand words than the PT-EU OSRAs (*MD* = 4.51; *SED* = .756; 95% CI = 6.00 to 3.01 personal pronouns per thousand words);

There also are statistically significant differences in the frequency of personal pronouns between the EN-ES^{EU} and ES-EU groups, *t*(128) = 7.124, *p* = .001.

The ES-EU OSRAs contain significantly more personal pronouns per thousand words than the EN-ES^{EU} OSRAs (*MD* = 5.71; *SED* = .802; 95% CI = 7.30 to 4.13 personal pronouns per thousand words);

IV. There are no statistically significant differences in the frequency of personal pronouns between the EN-PT^{EU} and EN-GB groups, *t*(100.984) = .219, *p* = .827. Both samples have similar means of personal pronouns per thousand words (*MD* = .231; *SED* = 1.05; 95% CI = 2.32 to -1.86 personal pronouns per thousand words).

There are also no statistically significant differences in the frequency of personal pronouns between the EN-ES^{EU} and EN-GB groups, *t*(104.738) = .684, *p* = .496. Both samples have similar means of personal pronouns per thousand words (*MD* = .731; *SED* = 1.07; 95% CI = 2.85 to -1.39 personal pronouns per thousand words).

No effects of L1 influence are found in OSRAs written in English by the L1 PT-EU or L1 ES-EU authors in relation to the frequency of personal pronouns.

On the one hand, all OSRAs written in English, either by L1 or non-L1 authors, have similar frequencies of personal pronouns, i.e., about 13 personal pronouns per thousand words. On the other hand, the two non-L1 English groups writing OSRAs in that language (EN-PT^{EU} and EN-ES^{EU}) differ from their respective L1 counterparts writing in their respective L1s (PT-EU and ES-EU). The L1 Portuguese authors writing OSRAs in Portuguese use significantly fewer personal pronouns per thousand words than native Portuguese writing OSRAs in English, and L1 Spanish authors use significantly more personal pronouns per thousand words when writing OSRAs in their native language than when writing OSRAs in English. Additionally, a comparison is run between the PT-EU and the ES-EU groups showing that there also are significant differences between these two language groups (*t*(128) = 13.348, *p* = .001).

In the CoRA, V15 is a variable that behaves uniformly in all groups writing in English and significantly different in the three language groups analyzed in this study, i.e., Portuguese, Spanish, and English. Therefore, no effects of L1 influence can be argued either for the EN-PT^{EU} or the EN-ES^{EU} group in relation to frequency of personal pronouns.

However, to understand if specific personal pronouns can function as L1 influence markers, the groups are linguistically examined. With that purpose, the personal pronouns found in at least 50% of all OSRAs within each corpus are extracted together with their frequency in the corresponding corpus, as shown below in Table 33.

Rank	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Threshold
1	[se] n=992	[se] n=2584	[we] n=1014	[we] n=1144	[we] n=1229	95-85%
2		[él] n=148	[it] n=523	[it] n=586	[it] n=690	80%
3		[lo] n=102	[they] n=487	[they] n=299	[they] n=412	75-70%
4		[la] n=97				65-60%
5		[nosotros] n=78				50%

Table 33 – Most frequent personal pronouns in the CoRA (present in 50% or higher of the OSRAs) ranked by the number of occurrences and with total frequency in corresponding corpus

By examining the data in Table 33, it is possible to see that all English groups have the same personal pronouns in at least 50% of the OSRAs in each corpus. The ES-EU is the corpus with the largest number of personal pronouns in 50% of the OSRAs in the corresponding corpus, while the PT-EU is the corpus with fewer personal pronouns within the 50% threshold.

Although the general data on personal pronouns do not indicate the existence of potential markers of NLID in this type of POS, the three corpora containing OSRAs written in English are analyzed in terms of distribution, looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. Upon analysis, all personal pronouns are considered to be unlikely to function as NLID markers in OSRAs written in English by the Portuguese/Spanish authors, given that their ranks are the same, despite differences in the percentages within each corpus and most numbers of occurrences being higher in the non-L1 English corpora as shown in Table 34.

N	Personal Pronoun	Corpus	Occurrences in corpus	Rank	% in corpus
1	[we]	EN-GB	1014	1	0,59
		EN-PT^{EU}	1144	1	0,70
		EN-ES^{EU}	1229	1	0,67
2	[it]	EN-GB	523	2	0,31
		EN-PT^{EU}	586	2	0,36
		EN-ES^{EU}	690	2	0,37
3	[they]	EN-GB	487	3	0,28
		EN-PT^{EU}	299	3	0,18
		EN-ES^{EU}	412	3	0,22

Table 34 – Personal pronouns considered to be unlikely to function as NLID markers in OSRAs written in English by the Portuguese/Spanish authors given their same ranks and similar percentage and occurrences in the corresponding corpus

Despite the *a priori* lack of potential to mark L1 influence, it is obvious that Spanish and Portuguese authors use the personal pronouns "we" and "it" more frequently than L1 English authors. As can be seen in Figure 10 below, such high frequencies in English do not reflect the frequencies of the equivalents in the Portuguese and Spanish OSRAs. However, the high frequency of "we" and "it" could result from the lack of an equivalent for the very frequent Portuguese and Spanish pronoun "se" and subsequent adaptation to English.

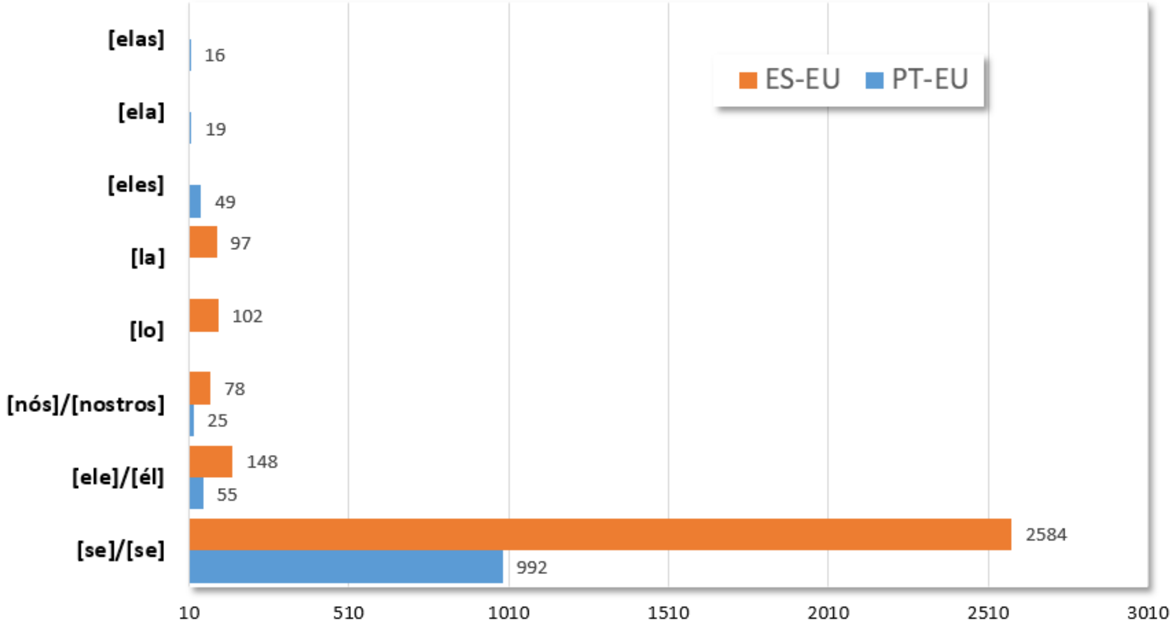


Figure 10 – Frequencies of personal pronouns in the PT-EU and the ES-EU corpora

Conveying the pronoun “se” from Portuguese and Spanish from phrases like (a), (c), (e), and (g) below needs the use of another pronoun in English if a similar syntactic construction is to be kept. Two possible translations of these constructions into English are found in sentences like (b), (d), (f) and (h) extracted from the PT-EU/ES-EU and the EN-PT^{EU}/EN-ES^{EU}.

a. *“**Pode-se concluir** que ainda há um longo caminho a percorrer, mas também se sabe que a idade não é fator intrínseco da fragilidade, embora as doenças crónicas sejam mais comuns em idades avançadas e vaticinadoras de reações adversas como quedas, hospitalização, institucionalização, declínio funcional e morte*[5]*, que a maioria dos idosos não são obrigatoriamente frágeis e que a sua origem não é simplesmente física.” (PT-EU_OSRA_032)*

b. *“Furthermore, because of the high-energy trauma mechanism generally involved, **it can be assumed** that associated injuries occur frequently in children who have sustained fractures of the facial skeleton.” (EN-PTeu_OSRA_044).*

c. *No que diz respeito ao protetor solar usado, cerca de 92% das crianças/adolescentes e 85% dos cuidadores responderam afirmativamente quanto ao uso do mesmo, sendo que **se observaram** diferenças estatisticamente significativas entre as crianças/adolescentes e os cuidadores (p <0,001).(PT-EU_OSRA_044)* *d.* ***We observed*** *that NAMPT is expressed in all tumor types tested, although the leukemia cell lines (NB4, ML2 and HL-60) showed weaker expression (Figure 1B). (EN-PTeu_OSRA_007)*

e. *"Sin embargo,* ***se sabe que*** *en el modelo porcino del IM agudo de isquemia/reperfusión el flujo colateral es muy pequeño o nulo." (ES-EU_OSRA_019)*

f. *"****It is known that*** *adult neurogenesis in different regions decreases exponentially with age." (EN-ESeu_OSRA_025)*

g. ***Se utilizó*** *el programa SPSS versión 13.0.(ES-EU_OSRA_031).*

h. ***We used*** *immunohistochemical staining of α-SMA to evaluate the degree of HSC activation. (EN-ESeu_OSRA_063)*

In the English corpora of the CoRA, the frequency of the structure [it can be + V Past Particle] is not very frequent, i.e., 18 occurrences. However, 89% of all occurrences are found in the non-L1 English corpora, i.e., EN-PT^{EU} and EN-ES^{EU}. The structure [we + V Past Tense] is far more frequent, i.e., 907 occurrences, but again the Portuguese/Spanish authors writing OSRAs in English (EN-PT^{EU} and EN-ES^{EU}) use it more frequently, i.e., 33% and 41%, respectively, than the L1 English authors writing in their L1 (EN-GB), i.e., 26%.

Overall, personal pronouns do not have the potential to function as NLID markers in relation to their mean frequency in the CoRA or distribution of specific pronouns within the corresponding corpus. However, certain lexical combinations with "it" may indicate that the authors are not L1 users of English but of Romance languages like Portuguese and Spanish.

4.3. Variables with only effect of L1 influence – Cross-Language Congruity

In this section, I present the variables for which only one effect of L1 influence is found. The single L1 effect refers to cross-language congruity, i.e., the L1 Portuguese/Spanish authors writing OSRAs in English do not differ significantly from the L1 Portuguese/Spanish authors writing OSRAs in Portuguese/Spanish in relation to these variables.

4.3.1. V3: number of paragraphs

After examining the samples of V3, the following descriptive statistics are obtained:

V3: number of paragraphs							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	28,17	1,793	24,59	31,75	25,00	209,018	14,457
ES-EU	23,78	1,174	21,44	26,13	21,00	89,547	9,463
EN-GB	20,91	0,792	19,33	22,49	19,00	40,741	6,383
EN-PT^{EU}	23,72	1,251	21,22	26,22	23,00	101,703	10,085
EN-ES^{EU}	21,35	0,943	19,47	23,24	20,00	57,763	7,600

The group with the highest median value of V3 is the PT-EU, and the group with the lowest value of V3 is the EN-GB group.

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V3 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test**,** *F* = 2.098, *p* = .150);

II. There are also no statistically significant differences in the number of paragraphs between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = 1.513, *p* = .133.
Both groups have similar means of number of paragraphs (*MD* = 2.37, *SED* = 1.57; 95% CI = 5.468 to -.730 paragraphs);

III. There are statistically significant differences in the number of paragraphs between the EN-PT^{EU} and PT-EU groups, *t*(114.359) = -2.134, *p* = .044.

The PT-EU OSRAs contain significantly more paragraphs than the EN-PT^{EU} OSRAs (*MD* = 4.446; *SED* = 2.186; 95% CI = 8.777 to .115 paragraphs);
there are no statistically significant differences in the number of paragraphs between the EN-ES^{EU} and ES-EU groups, t(128) = 1.615, p = .109.
Both groups have similar means of number of paragraphs (*MD* = 2.43, *SED* = 1.505; 95% CI = 5.410 to -.548 paragraphs);

IV. There are no statistically significant differences in the number of paragraphs between the EN-PT^{EU} and EN-GB groups, *t*(108.185) = 1.909, *p* = .60.
Both groups of OSRAs have fairly similar numbers of paragraphs (*MD* = 2.82; *SED* = 1.480; 95% CI = 5.750 to -.119 paragraphs).

There are also no statistically significant differences in the number of paragraphs between the EN-ES^{EU} and EN-GB groups, *t*(128) = .362, *p* = .718.
The groups have almost the same number of paragraphs (*MD* = .446; *SED* = 1.231; 95% CI = 2.882 to -1.990 paragraphs).

The following table summarizes the effects of L1 influence found in each group of OSRAs:

Effect of L1 Influence	**EN-PT^{EU}**	**EN-ES^{EU}**
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	--	✓
IV. Intralingual contrast	--	--

No effects of L1 influence are found for V3 in OSRAs written in English by the Portuguese authors. Similarly, only one effect of L1 influence is found in relation to V3 in OSRAs written in English by the Spanish authors.

The variances of V3 of the EN-PT^{EU} and the EN-ES^{EU} groups are not significantly different, which means that the distribution of the number of paragraphs of the OSRAs within each group is fairly similar. Also, the overall mean values of this variable are not significantly different between the EN-PT^{EU} and the EN-ES^{EU} groups.

There are no significant differences in the number of paragraphs between the non-L1 (EN-PT^{EU} and EN-ES^{EU}) and the L1 (EN-GB) English groups writing in English since they all have fairly similar means of V3, i.e., 21.99 paragraphs, on average.

Additionally, significant differences are found between the Portuguese authors writing in English and the Portuguese authors writing in their L1 in relation to the number of paragraphs. The L1 PT-EU authors significantly reduce the number of paragraphs they produce when writing OSRAs in English, in relation to number of paragraphs they produce when writing OSRAs in their L1. Similarly, such a reduction in the number of paragraphs contributes to the lack of significant differences between the Portuguese authors writing in English (EN-PT^{EU}) and the English authors writing in their L1 (EN-GB).

The Spanish authors also reduce the number of paragraphs they produce when writing OSRAs in English compared to what they produce in their L1. This downsizing in the number of paragraphs is enough to eliminate significant differences between the Spanish authors writing in English (EN-ES^{EU}) and the L1 English authors writing in their L1 (EN-GB). However, it is not sufficient to show a significant difference between the EN-ES^{EU} and the ES-EU groups, bringing about a cross-language congruity in relation to the number of paragraphs.

Additionally, a comparison of the PT-EU and the ES-EU groups shows that there are also significant differences between these groups in relation to V3 (t(110.333) = 2.046, p = .043) with the PT-EU group producing more paragraphs.

In the CoRA, the number of paragraphs (V3) is fairly uniform in all three groups writing OSRAs in English (i.e., 21.99 paragraphs, on average) and significantly different between all three English groups (EN-GB, EN-PT^{EU,} and EN-ES^{EU}) and the PT-EU group.

Only one effect of L1 influence is observed for only one of the groups, i.e., the lack of difference in the paragraphs division between the Spanish authors writing OSRAs in Spanish and the Spanish authors writing OSRAs in English. This effect could be explained, for example, by the restrictions imposed by the scientific journals where the ES-EU OSRAs are published in relation to the limit in the number of characters in an OSRA.

Therefore, no effects of L1 influence are discussed for the EN-PT^{EU} or the EN-ES^{EU} groups in relation to this variable.

4.3.2. V4: standardized type/token ratio (STTR)

After examining the samples of V3, the following descriptive statistics are obtained:

V4: standardized type/token ratio (per 1000 tokens)							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	40,38	0,42	39,53	41,23	40,50	11,65	3,41
ES-EU	38,00	0,38	37,25	38,75	37,73	9,14	3,02
EN-GB	38,42	0,48	37,46	39,37	38,85	14,90	3,86
EN-PT^{EU}	38,05	0,37	37,32	38,79	38,30	8,80	2,97
EN-ES^{EU}	37,22	0,47	36,28	38,16	37,30	14,35	3,79

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V4 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test**,** *F* = 2.697, *p* = .103);

II. There are no statistically significant differences in the standardized type/token ratio between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = 1.402, *p* = .163.
Both groups have very similar STTR (*MD* = .836, *SED* = .597; 95% CI = 2.02 to -.344);

III. There are statistically significant differences in the standardized type/token ratio between the EN-PT^{EU} and PT-EU groups, *t*(128) = 4.149, *p* = .001.
The PT-EU corpus has a higher type/token ratio than the EN-PT^{EU} corpus (*MD* = 2.33, *SED* = .561, 95% CI = 3.44 to 1.22);

There are no statistically significant differences in the standardized type/token ratio between the EN-ES^{EU} and ES-EU groups, *t*(128) = 1.306, *p* = .194.
Both groups have similar type/token ratios (*MD* = .785, *SED* = .601; 95% CI = 1.975 to -.405);

IV. There are no statistically significant differences in the standardized type/token ratio between the EN-PT^{EU} and EN-GB groups, *t*(128) = .605, *p* = .546.

Both groups have similar type/token ratios (*MD* = .365; *SED* = .604; 95% CI = 1.560 to -.829).

There also are no statistically significant differences in the standardized type/token ratio between the EN-ES^{EU} and EN-GB groups, *t*(128) = 1.791, *p* = .076.
Both groups have similar standardized type/token ratios (*MD* = 1.20; *SED* = .671; 95% CI = 2.53 to -.126).

The following table summarizes the effects of L1 influence found in each group of OSRAs:

Effect of L1 Influence	EN-PT^{EU}	EN-ES^{EU}
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	--	✓
IV. Intralingual contrast	--	--

No effects of L1 influence are found for V4 in OSRAs written in English by the Portuguese authors, and only one possible L1 influence effect is found in relation to this variable in OSRAs written in English by the Spanish authors.

According to Levene's test performed to determine homogeneity of variances in the EN-PT^{EU} and the EN-ES^{EU} groups, OSRAs written in English by PT-EU and ES-EU authors are homogeneous in their V4 internal distribution. These authors behave uniformly within their groups in relation to the standardized type/token ratio. However, the mean values of V4 are not significantly different between these groups, which does not allow one to argue intergroup heterogeneity.

The mean values of V4 are significantly different in OSRAs written in English by the Portuguese authors and in OSRAs written in Portuguese by the L1 authors of that language, with the latter having a higher standardized type/token ratio. However, the mean values of V4 are very similar in OSRAs written in English by the Spanish authors and in OSRAs written in Spanish by the L1 authors of Spanish. The former group has a mean value of standardized type/token ratio of 37.22, and the latter group has a mean value of 38.00, which means that there is congruity between these groups in relation to V4.

Finally, the Portuguese and Spanish authors do not differ significantly from the L1 English authors in relation to the standardized type/token ratio when writing OSRAs in English. Both the Portuguese and the Spanish authors decrease their standardized type/token ratio when they write OSRAs in English in relation to the ratio they present when writing in their respective L1s. However, and as mentioned above, in the case of the Spanish authors, this reduction does not allow for differentiating the EN-ES^{EU} authors from the ES-EU authors.

Overall, only one effect of L1 influence is found, i.e., language congruity between the Spanish authors writing OSRAs in their L1 and the Spanish authors writing OSRAs in English. Therefore, the standardized type/token ratio (STTR) is not a good variable to detect L1 influence in the OSRAs of the CoRA.

4.3.3. V5: number of 1-to-5-letter words

After examining the samples of V5, the following descriptive statistics are obtained:

V5: frequency of 1 to 5-letter words (all values in words/1000 tokens)							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	582,82	3,88	575,07	590,56	581,00	977,19	31,26
ES-EU	607,26	3,45	600,36	614,16	609,00	774,51	27,83
EN-GB	605,98	5,18	595,63	616,34	603,00	1747,24	41,80
EN-PT^{EU}	604,15	3,74	596,67	611,63	599,00	911,44	30,19
EN-ES^{EU}	601,83	3,78	594,29	609,37	605,00	926,59	30,44

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V5 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test, *F* = .021, *p* = .884);

II. There are no statistically significant differences in the frequency of 1-to-5-letter words between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = .437, *p* = .663.
Both groups have similar means of 1-to-5-letter words per thousand words (*MD* = 2.32; *SED* = 5.31; 95% CI = 12.84 to -8.20 1-to-5-letter words per thousand words);

III. There are statistically significant differences in the frequency of 1-to-5-letter words between the EN-PT^{EU} and PT-EU groups, *t*(128) = 3.958, *p* = .001.
EN-PT^{EU} OSRAs contain significantly more 1-to-5-letter words per thousand words than PT-EU OSRAs (*MD* = 21.33; *SED* = 5.39; 95% CI = 32.00 to 10.67 1-to-5-letter words per thousand words);

There are no statistically significant differences in the frequency of 1-to-5-letter words between the EN-ES^{EU} and ES-EU groups, *t*(128) = 1.062, *p* = .290.
Both groups have similar means of 1-to-5-letter words per thousand words (*MD* = 5.43; *SED* = 5.11; 95% CI = 4.69 to -15.55 1-to-5-letter words per thousand words);

IV. There are no statistically significant differences in the frequency of 1-to-5-letter words between the EN-PT^{EU} and EN-GB groups, *t*(128) = .286, *p* = .775.
Both groups have similar means of 1-to-5-letter words per thousand words (*MD* = 1.83; *SED* = 6.39; 95% CI = 14.49 to -10.82 1-to-5-letter words per thousand words).

There are no statistically significant differences in the frequency of 1-to-5-letter words between the EN-ES^{EU} and EN-GB groups, *t*(128) = .648, *p* = .518.
Both groups have similar means of 1-to-5-letter words per thousand words (*MD* = 4.15; *SED* = 6.41; 95% CI = 16.84 to -8.53 1-to-5-letter words per thousand words).

The following table summarizes the effects of L1 influence found in each group of OSRAs:

Effect of L1 Influence	**EN-PT^{EU}**	**EN-ES^{EU}**
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	--	✓
IV. Intralingual contrast	--	--

The frequency of 1-to-5-letter words is not influenced by the authors' L1 in OSRAs written in English by the Portuguese L1 authors since no effects of L1 influence are found. Similarly, in OSRAs written in English by the L1 Spanish authors, the frequency of 1-to-5-letter words is not influenced by the authors' L1 since only one effect of L1 influence is verified.

The Portuguese and Spanish authors writing OSRAs in English do not differ significantly between each other in relation to the variances of the frequency of 1-to-5-letter words. These groups also do not differ in relation to the mean values of the frequency of 1-to-5-letter words per thousand words, which means that no intergroup heterogeneity can be argued.

The Portuguese authors use significantly more 1-to-5-letter words per thousand words when writing OSRAs in English than when they write OSRAs in their L1. Similarly, the Portuguese authors writing OSRAs in English do not differ from the L1 English authors writing in their L1 since both groups have very similar mean values of 1-to-5-letter words per thousand words, i.e., 604.15 and 605.98, respectively.

The Spanish groups behave slightly differently. The Spanish authors writing OSRAs in English do not differ from the Spanish authors writing in their L1 (ES-EU) or from L1 English authors writing in English (EN-GB). The ES-EU group has the highest mean value of 1-to-5-letter words per thousand words (607.26). When the L1 ES-EU authors write OSRAs in English, they decrease the frequency of 1-to-5-letter words per thousand words to 601.83, coming closer to the L1 EN-GB authors writing in English who use 605.98 1-to-5-letter words per thousand words.

Since the results obtained for this variable do not show potential to mark NLID, no further analyses are performed.

4.3.4. V6: number of 6-to-10-letter words

After examining the samples of V6, the following descriptive statistics are obtained:

V6: frequency of 6-to-10-letter words (all values in words/1000 tokens)							
CORPORA	Mean	Std. Error	95% Confidence Interval for Mean		Median	Variance	Std. Deviation
			Lower Bound	Upper Bound			
PT-EU	319,05	3,21	312,64	325,46	318,50	668,74	25,86
ES-EU	304,58	2,81	298,97	310,19	303,20	513,02	22,65
EN-GB	329,94	4,25	321,45	338,43	330,60	1173,75	34,26
EN-PT^{EU}	324,62	3,44	317,76	331,49	329,20	767,84	27,71
EN-ES^{EU}	327,13	3,37	320,40	333,87	327,10	738,21	27,17

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V6 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test, *F* = .035, *p* = .852);

II. There are no statistically significant differences in the frequency of 6-to-11-letter words between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = .521, *p* = .603.
Both groups have similar means of 6-to-10-letter words per thousand words (*MD* = 2.51; *SED* = 4.81; 95% CI = 12.03 to -7.01 6-to-10-letter words per thousand words);

III. There are no statistically significant differences in frequency of 6-to-10-letter words between the EN-PT^{EU} and PT-EU groups, *t*(128) = 1.187, *p* = .238.
Both groups have similar means of 6-to-10-letter words per thousand words (*MD* = 5.58; *SED* = 4.70; 95% CI = 14.88 to -3.72 6-to-10-letter words per thousand words);

There are statistically significant differences in the frequency of 6-to-10-letter words between the EN-ES^{EU} and ES-EU groups, *t*(128) = 5.142, *p* = .001.
The EN-ES^{EU} group has significantly more 6-to-10-letter words per thousand words than the ES-EU group (*MD* = 22.56; *SED* = 4.39, 95% CI = 31.24 to 13.88 6-to-10-letter words per thousand words);

IV. There are no statistically significant differences in the frequency of 6-to-10-letter words between the EN-PT^{EU} and EN-GB groups, *t*(128) = .973, *p* = .333.
Both groups have similar means of 6-to-10-letter words per thousand words (*MD* = 5.32; *SED* = 5.47; 95% CI = 16.13 to -5.50 6-to-10-letter words per thousand words).

There are no statistically significant differences in the frequency of 6-to-10-letter words between the EN-ES^{EU} and EN-GB groups, *t*(128) = .517, *p* = .606.
Both groups have similar means of 6-to-10-letter words per thousand words (*MD* = 2.81; *SED* = 5.42; 95% CI = 13.54 to -7.93 6-to-10-letter words per thousand words).

The following table summarizes the effects of L1 influence found in each group of OSRAs:

Effect of L1 Influence	EN-PT^{EU}	EN-ES^{EU}
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	✓	--
IV. Intralingual contrast	--	--

The frequency of 6-to-11-letter words does not appear to be influenced by the authors' L1 in OSRAs written in English by the Spanish authors since no effects of L1 influence are found. Also, in OSRAs written in English by the Portuguese L1 authors, the frequency of 6-to-11-letter words does not seem to be influenced by the authors' L1 since only one possible effect of L1 influence is found.

The results of Levene's test indicate that the EN-PT^{EU} and the EN-ES^{EU} groups have similar variances of V6. However, these groups are not different in terms of their mean values of 6-to-11-letter words. In fact, none of the three groups writing OSRAs in English (EN-GB, EN-PT^{EU}, EN-ES^{EU}) uses 6-to-11-letter words in a frequency that is significantly different from the other two since their mean values of V6 are very similar, i.e., 329.94, 324.62, 327.13, respectively.

Moreover, the Spanish authors writing OSRAs in English differ significantly from the Spanish authors writing OSRAs in their L1 in relation to the frequency of V6. That is, the Spanish authors increase the frequency of V6 significantly when they write OSRAs in English. The Portuguese authors also increase the frequency of V6 when they write OSRAs in English, but not enough to differ significantly from the Portuguese authors writing in their L1.

No further analyses are carried out with 6-to-10 letter words since no significant results are obtained from the comparisons performed to assess this variable's potential to mark L1 influence in the OSRAs of the CoRA.

4.3.5. V10: number of coordinating conjunctions

After examining the samples of V10, the following descriptive statistics are obtained:

V10: frequency of coordinating conjunctions (all values in words/1000 tokens)							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	39,88	1,05	37,77	41,98	40,64	72,08	8,49
ES-EU	34,39	1,02	32,34	36,43	33,49	68,06	8,25
EN-GB	38,99	1,05	36,89	41,10	38,40	72,25	8,50
EN-PT^{EU}	41,24	0,91	39,41	43,06	39,98	54,17	7,36
EN-ES^{EU}	38,83	1,08	36,66	40,99	38,06	76,39	8,74

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V10 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test**,** *F* = .506, *p* = .478);

II. There are no statistically significant differences in the frequency of coordinating conjunctions between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = 1.699, *p* = .092.
Both groups have similar means of coordinating conjunctions per thousand words (*MD* = 2.41; *SED* = 1.42; 95% CI = 5.21 to -.396 coordinating conjunctions per thousand words);

III. There are no statistically significant differences in frequency of coordinating conjunctions between the EN-PT^{EU} and PT-EU groups, *t*(128) = .975, *p* = .331.
Both groups of OSRAs have approximately the same mean of coordinating conjunctions per thousand words (*MD* = 1.36; *SED* = 1.39, 95% CI = 4.12 to -1.40 coordinating conjunctions per thousand words);

There are statistically significant differences in the frequency of coordinating conjunctions between the EN-ES^{EU} and ES-EU groups, *t*(128) = 2.978, *p* = .003.
The EN-ES^{EU} OSRAs have significantly more coordinating conjunctions per thousand words than the ES-EU OSRAs (*MD* = 4.44; *SED* = 1.49; 95% CI = 7.39 to 1.49 coordinating conjunctions per thousand words);

IV. There are no statistically significant differences in the frequency of coordinating conjunctions between the EN-PT^{EU} and the EN-GB groups, *t*(128) = 1.608, *p* = .110. Both groups have similar means of coordinating conjunctions per thousand words (*MD* = 2.24; *SED* = 1.40; 95% CI = 5.00 to -.517 coordinating conjunctions per thousand words).

There also are no statistically significant differences in the frequency of coordinating conjunctions between the EN-ES^{EU} and the EN-GB groups, *t*(128) = .109, *p* = .913. Both groups have almost identical means of coordinating conjunctions per thousand words (*MD* = .165; *SED* = 1.51; 95% CI = 3.16 to -2.83 coordinating conjunctions per thousand words).

The following table summarizes the effects of L1 influence found in each group of OSRAs:

Effect of L1 Influence	**EN-PT^{EU}**	**EN-ES^{EU}**
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	✓	--
IV. Intralingual contrast	--	--

In OSRAs written in English by the Portuguese or by the Spanish authors, the frequency of coordinating conjunctions does not seem to be influenced by the use these authors make of this variable when writing OSRAs in their L1 since no effects are found for the Spanish group, and only one potential effect is found for the Portuguese group. Moreover, no significant differences are found between the groups writing OSRAs in English, either produced by L1 or non-L1 authors, in relation to the number of coordinating conjunctions per thousand words.

The groups are linguistically examined to verify if such a lack of significant differences reflects the use of specific coordinating conjunctions by the authors. Therefore, the coordinating conjunctions found in at least 50% of all OSRAs within each corpus are extracted together with their frequency in the corresponding corpus, as shown below in Table 35.

Rank	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Threshold
1	[e] n=4008	[y]/[e][16] n=4394	[and] n=4806	[and] n=5160	[and] n=5521	95%
2	[ou] n=621	[o] n=724	[or] n=850	[or] n=617	[or] n=729	85-80%
3	[/] n=398	[además] n=139	[/] n=379	[/] n=330	[/] n=382	75%
4	[no=entanto] n=156	[pero] n=139	[but] n=348	[but] n=253	[but] n=282	70%
5	[mas] n=143			[as=well=as] n=76	[as=well=as] n=79	65%
6	[assim] n=87					60%
7	[contudo] n=87					50%

Table 35 – Most frequent coordinating conjunctions in the CoRA

[16] Orthographic variation of the coordinating conjuction [y]-/ɪ/, which in Spanish changes to [e]-/e/ when the following word begins with the sound /ɪ/. Used to avoid lengthing of the sound /ɪ/ and thus dissonance. E.g., from the CoRA: *necesario e imprescindible*; *soluble e inmovilizada*; *atención e hiperactividad*

As can be verified in Table 35, the number of coordinating conjunctions found in 50% or more of the OSRAs in each corpus is rather small. The Portuguese authors writing OSRAs in their L1 use more coordinating conjunctions more consistently across the OSRAs in terms of frequency. All authors writing in English use the same coordinating conjunctions, except for the non-L1 authors who use "as=well=as" more frequently than L1 authors.

To examine the difference found in the three English corpora in relation to the coordinating conjunctions "as=well=as", the OSRAs are further analyzed looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. After analysis, four coordinating conjunctions are deemed unlikely to function as NLID markers since their ranks are the same and their percentages in the corresponding corpus are very similar, despite differences in the number of occurrences. See Table 36 below.

N	Coordinating Conjunction	Corpus	Occurrences in Corpus	Rank	% in corpus
1	and	EN-GB EN-PT^{EU} EN-ES^{EU}	4806 5160 5521	1 1 1	2,81 3,16 3,00
2	or	EN-GB EN-PT^{EU} EN-ES^{EU}	850 617 729	2 2 2	0,50 0,38 0,40
3	/	EN-GB EN-PT^{EU} EN-ES^{EU}	379 330 382	3 3 3	0,22 0,20 0,21
4	but	EN-GB EN-PT^{EU} EN-ES^{EU}	348 253 282	4 4 4	0,20 0,15 0,15

Table 36 – Coordinating conjunctions considered to be unlikely to function as NLID markers in OSRAs written in English by Portuguese/Spanish authors given their same ranks and a similar percentage in the corresponding corpus

Therefore, only the coordinating conjunction "as=well=as" is further analyzed to verify if it may function as a NLID marker in the OSRAs written in English by the Portuguese/Spanish authors. Because this coordinating conjunction is being assessed as a language transfer marker, its equivalent in Portuguese and Spanish are extracted from the corresponding L1 corpora of the CoRA. In this respect, it is worth mentioning that the coordinating conjunction "as=well=as" usually translates into the phrasal coordinating conjunctions "assim=como" /

"bem=como" / "tal=como" in Portuguese, and "así=como" / "así=como=también" in Spanish. Among the data extracted from the CoRA using the tag <kc>, i.e., coordinating conjunctions, no occurrences are found that correspond to any of these phrasal conjunctions. Since this absence is not common, all occurrences of these phrasal conjunctions are extracted from the parsed PT-EU and ES-EU corpora using their morphological forms. 137 are extracted from the PT-EU corpus and 86 from the ES-EU corpus. A detailed inspection of all the concordances obtained with WordSmith (Scott 2018b) shows that the software VISL tags these morphological forms under the category adverb (ADV), resulting in thirteen different tags for the PT-EU and eleven for the ES-EU groups, shown below in Table 37.

N	PT-EU	ES-EU
1	ADV @ADVL @#AS-<ADVL	ADV @#AS-<ADVL
2	ADV @ADVL @#AS-ADVL	ADV @ADVL @#AS-<ADVL
3	ADV @ADVL @#AS-ADVL>	ADV @ADVL @#AS-A<
4	ADV @ADVL @#AS-N	ADV @ADVL>
5	ADV @ADVL>	ADV @ADVL> @#FS-<ADVL
6	ADV @ADVL> @#FS-	ADV @ADVL> @#FS-ADVL
7	ADV @ADVL> @#FS-<ADVL	ADV @ADVL> @#FS-ADVL>
8	ADV @ADVL> @#FS-ADVL	ADV @COM
9	ADV @ADVL> @#FS-ADVL>	ADV @COM @#AS-<ADVL
10	ADV @UTT @#AS-	ADV @COM @#AS-ADVL>
11	ADV @UTT @#AS-<ADVL	ADV @COM @#AS-AS
12	ADV @UTT @#AS-<ADVL @#AS-NN	
13	ADV @UTT @ADVL> @#FS-<ADVL	

Table 37 – Symbols used by VISL to tag the morphological forms "assim=como" / "bem=como" / "tal=como" in Portuguese, and "así=como" / "así=como=también" in Spanish

This tagging probably responds to a prioritization of the main POS of the phrasal coordinating conjunctions. After close reading of a random sample of 24 of the sentences containing these phrasal conjunctions under the category ADV (n=12 per corpus), only the coordinating function is verified (Matos and Raposo 2013: 1777). Therefore, these are considered for the analysis of "as=well=as" with the KC function. Below are some examples from the CoRA:

a) *La prevalencia de la diabetes mellitus tipo 2 (DM2) tiene tendencia a incrementarse, debido a los cambios alimenticios, **así como** al envejecimiento poblacional, cambios en los criterios diagnósticos y menor mortalidad de los pacientes diabéticos.*

b) *Foram recolhidos dados demográficos, da terapêutica antineoplásica em curso, bem como referentes às variáveis ecocardiográficas.*

Table 38 shows the coordinating conjunction "as=well=as" and its equivalents in Portuguese and Spanish as found in the corresponding L1 corpora under the tag ADV.

N	Coordinating conjunction	Corpus	Occurrences in Corpus	Rank	% in corpus
1	bem=como/assim=como	PT-EU	86	31	0,06
	así como/así como también	ES-EU	137	26	0,08
	as=well=as	EN-GB	32	9	0,02
		EN-PT^{EU}	76	5	0,05
		EN-ES^{EU}	79	5	0,04

Table 38 – The coordinating conjunction "as=well=as" and equivalents in Portuguese/Spanish considered to be likely to function as NLID markers in OSRAs written in English by the Portuguese/Spanish authors

The frequencies of these conjunctions are compared in order to examine if there are significant differences between the groups. The following questions are asked of both the EN-PT^{EU} and the EN-ES^{EU} corpora.

Effect of L1 Influence	L1 influence EN-PT^{EU} questions	L1 influence EN-ES^{EU} questions
I) Intragroup homogeneity	Are the frequencies of the KC in the EN-PT^{EU} / EN-ES^{EU} OSRAs uniformly distributed?	
II) Intergroup heterogeneity	Are the frequencies of the KC in the EN-PT^{EU} and EN-ES^{EU} OSRAs statistically significantly different?	
III) Cross-language congruity	Are the frequencies of the KC in the EN-PT^{EU} and PT-EU OSRAs statistically similar?	Are the frequencies of the KC in the EN-ES^{EU} and ES-EU OSRAs statistically similar?
IV) Intralingual contrast	Are the frequencies of the KC in the EN-PT^{EU} and EN-GB OSRAs statistically significantly different?	Are the frequencies of the KC in the EN-ES^{EU} and EN-GB OSRAs statistically significantly different?

The Mann-Whitney test is used to assess the mean difference between the groups given that the data is not normally distributed and has some outliers. Questions I and II are answered together. The level of significance used is p < .05 for questions II and IV. Since questions I and III look for uniformity and congruity, respectively, a result of p > .05 is associated with a possible effect of L1 influence.

Table 39 below shows the results and mean ranks of the comparisons performed. As can be seen, no significant differences are observed between the English and the Portuguese authors writing OSRAs in English. The Portuguese authors use “bem=como”, “assim=como”, and “tal=como” significantly more frequently in Portuguese than they use “as=well=as” in English. Therefore, no NLID markers can be associated with the frequency of these coordinating conjunctions by the Portuguese authors.

Significant differences indicating the presence of possible NLID markers are observed only for the group of the Spanish authors. These authors use as many “así=como” and “así=como=también” when writing OSRAs in Spanish as they use “as=well=as” when writing OSRAs in English, and they use “as=well=as” significantly more frequently than the L1 English authors writing OSRAs in English. This may indicate that the Spanish authors transfer the use they make of the equivalent of “as=well=as” in Spanish into English.

	L1 Influence Effects								
	I) Intra-L1 homogeneity (Levene's test)	II) Inter-L1 heterogeneity (Mann-Whitney Test)	EN-PT^{EU} and EN-ES^{EU} similar in variance but different in means?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-PT^{EU} group?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-ES^{EU} group?
	Corpora Examined			**Corpora Examined**			**Corpora Examined**		
Coordinating conjunction	EN-PT^{EU} vs. EN-ES^{EU}	EN-PT^{EU} vs. EN-ES^{EU}		EN-PT^{EU} vs. PT-EU	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. ES-EU	EN-ES^{EU} vs. EN-GB	
	Reference *p* values			**Reference *p* values**			**Reference *p* values**		
	p > .05 **AND**	*p* < .05?		*p* > .05?	*p* < .05?		*p* > .05?	*p* < .05?	
bem=como / assim=como así=como / así=como=también as=well=as	*F* = .344 *p* = .560	*Z* = -1.935 *p* = .053 Mean ranks: EN-PT^{EU}= 32.42 EN-ES^{EU}= 41.32	*no*	*Z* = -2.635 *p* = .008 Mean ranks: EN-PT^{EU}= 37.64 PT-EU= 51.51	*Z* = -.754 *p* = .451 Mean ranks: EN-PT^{EU}= 32.64 EN-GB= 29.57	*no*	*Z* = -1.524 *p* = .127 Mean ranks: EN-ES^{EU}= 40.33 ES-EU= 33.26	*Z* = -2.706 *p* = .007 Mean ranks: EN-ES^{EU}= 33.08 EN-GB= 21.93	*yes*

Table 39 – Results of the Mann-Whitney's tests performed to assess mean differences between the groups in relation to the coordinating conjunction [as=well=as] and equivalents in Portuguese and Spanish

4.3.6. V11: number of subordinating conjunctions

After examining the samples of V11, the following descriptive statistics are obtained:

V11: frequency of subordinating conjunctions (all values in words/1000 tokens)							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	10,34	0,54	9,26	11,43	10,27	19,18	4,38
ES-EU	14,10	0,64	12,83	15,37	13,98	26,21	5,12
EN-GB	10,82	0,52	9,79	11,85	10,60	17,39	4,17
EN-PT^{EU}	10,85	0,52	9,81	11,89	10,81	17,56	4,19
EN-ES^{EU}	11,98	0,46	11,06	12,90	11,81	13,76	3,71

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V11 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test**,** *F* = 2.778, *p* = .098);

II. There are no statistically significant differences in the frequency of subordinating conjunctions between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = 1.623, *p* = .107.
Both groups have similar means of subordinating conjunctions per thousand words (*MD* = 1.13; *SED* = .695; 95% CI = 2.50 to -.247 subordinating conjunctions per thousand words);

III. There are no statistically significant differences in frequency of subordinating conjunctions between the EN-PT^{EU} and PT-EU groups, *t*(128) = .671, *p* = .503.
Both groups have fairly similar means of subordinating conjunctions per thousand words (*MD* = .505; *SED* = .753, 95% CI = 1.99 to -.984 subordinating conjunctions per thousand words);

There are statistically significant differences in the frequency of subordinating conjunctions between the EN-ES^{EU} and ES-EU groups, *t*(128) = 2.706, *p* = .008.
The ES-EU OSRAs have significantly more subordinating conjunctions per thousand words than the EN-ES^{EU} OSRAs (*MD* = 2.12; *SED* = .785, 95% CI = 3.68 to .571 subordinating conjunctions per thousand words);

IV. There are no statistically significant differences in the frequency of subordinating conjunctions between the EN-PT^{EU} and EN-GB groups, *t*(128) = .038, *p* = .970.

Both groups have almost identical means of subordinating conjunctions per thousand words (*MD* = .028; *SED* = .733; 95% CI = 1.48 to -1.42 subordinating conjunctions per thousand words).

There are no statistically significant differences in the frequency of subordinating conjunctions between the EN-ES^{EU} and EN-GB groups, *t*(128) = 1.669, *p* = .098. Both groups have very similar means of subordinating conjunctions per thousand words (*MD* = 1.16; *SED* = .692; 95% CI = 2.52 to -.215 subordinating conjunctions per thousand words).

The following table summarizes the effects of L1 influence found:

Effect of L1 Influence	**EN-PT^{EU}**	**EN-ES^{EU}**
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	✓	--
IV. Intralingual contrast	--	--

The Spanish authors use more subordinating conjunctions when they write OSRAs in their L1 than when they write OSRAs in English, and this decrease in relation to English corresponds with the frequency with which the L1 English authors use subordinating conjunctions when writing OSRAs in their L1.

The Portuguese authors, however, use subordinating conjunctions at very similar frequencies when writing OSRAs in their L1, and when writing OSRAs in English. Moreover, the frequency with which these authors use subordinating conjunctions in English is not significantly different from the frequency with which the L1 English authors use subordinating conjunctions when writing OSRAs in English.

Only one effect of L1 influence is found in relation to the number of subordinating conjunctions used by the Portuguese authors when writing OSRAs in English, i.e., cross-language congruity, and no effects are found in OSRAs written by the Spanish authors writing in English. Therefore, no L1 influence can be argued in relation to the frequency of subordinating conjunctions in either group.

The groups are linguistically examined to understand if this lack of L1 influence reflects the use of specific subordinating conjunctions. The subordinating conjunctions found in 50% or more of all OSRAs within each corpus are extracted together with the total frequency of each subordinating conjunction, as shown below in Table 40.

Rank	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Threshold
1	[que] n=732	[que] n=1120	[that] n=913	[that] n=974	[that] n=1252	95%
2	[se] n=190	[según] n=178	[although] n=154	[although] n=178	[although] n=208	85-80%
3	[embora] n=82	[si] n=154	[as] n=140	[while] n=115	[because] n=104	75%
4	[como] n=70	[aunque] n=149	[if] n=113	[as] n=85	[as] n=88	70%
5		[ya=que] n=142	[while] n=110	[because] n=80	[whether] n=88	65%
6		[cuando] n=122	[whether] n=101	[if] n=79	[while] n=88	60%
7		[así=como] n=86			[whereas] n=85	
8		[mientras=que] n=86				50%

Table 40 – Most frequent subordinating conjunctions in the CoRA (present in 50% or higher of the OSRAs) ranked by the number of occurrences and with total frequency in corresponding corpus

As can be noticed in Table 40, the number of subordinating conjunctions found in 50% or higher of the OSRAs in each corpus is small. Portuguese authors writing OSRAs in their L1 use less subordinating conjunctions than all the other groups and the ES-EU is the group with the largest number of subordinating conjunctions in the 50% threshold. Finally, all groups writing in English have similar numbers of subordinating conjunctions within that threshold, despite differences in the number of occurrences.

To examine possible NLID markers within subordinating conjunctions, the OSRAs in the three English corpora are further analyzed, looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. Upon analysis, most subordinating conjunctions are deemed unlikely to function as NLID markers given their similar ranks and their percentage in the corresponding corpus, despite differences in the number of occurrences between the groups, as shown below in Table 41.

Nº	Subordinating conjunction	Corpus	Occurrences in Corpus	Rank	% in corpus
1	although	EN-GB	154	2	0,09
		EN-PT^{EU}	178	2	0,11
		EN-ES^{EU}	208	2	0,11
2	as	EN-GB	140	3	0,08
		EN-PT^{EU}	85	4	0,05
		EN-ES^{EU}	88	4	0,05
3	while	EN-GB	110	5	0,06
		EN-PT^{EU}	115	3	0,07
		EN-ES^{EU}	88	6	0,05
4	because	EN-GB	49	8	0,03
		EN-PT^{EU}	80	5	0,05
		EN-ES^{EU}	104	3	0,06
5	if	EN-GB	113	4	0,07
		EN-PT^{EU}	79	6	0,05
		EN-ES^{EU}	59	9	0,04
6	whether	EN-GB	101	6	0,06
		EN-PT^{EU}	41	8	0,03
		EN-ES^{EU}	88	5	0,05
7	whereas	EN-GB	53	7	0,03
		EN-PT^{EU}	44	9	0,03
		EN-ES^{EU}	85	7	0,05

Table 41 – Subordinating conjunctions considered to be unlikely to function as NLID markers in OSRAs written in English by Portuguese/Spanish authors given their similar ranks and percentage in the corresponding corpus

Therefore, only one subordinating conjunction is further analyzed to verify if it may function as a NLID marker in OSRAs written in English by Portuguese/Spanish authors. Because this subordinating conjunction is being assessed as a language transfer marker, its equivalents in Portuguese and Spanish are extracted from the PT-EU and the ES-EU corpora.

Nº	Subordinating conjunction	Corpus	Occurrences in Corpus	Rank	% in corpus
1	[que]	PT-EU	732	1	0,51
	[que]	ES-EU	1120	1	0,69
	[that]	EN-GB	913	1	0,53
		EN-PT^{EU}	974	1	0,60
		EN-ES^{EU}	1252	1	0,68

Table 42 – Subordinating conjunctions that may function as NLID markers in OSRAs written in English by Portuguese/Spanish authors

The frequencies of "that" and equivalents "que", and "que", are compared to examine significant differences between the groups. The following questions are asked for both the EN-PT^{EU} and the EN-ES^{EU} corpora.

Effect of L1 Influence	L1 influence EN-PT^{EU} questions	L1 influence EN-ES^{EU} questions
I) Intragroup homogeneity	Are the frequencies of [that] in the EN-PT^{EU} / EN-ES^{EU} OSRAs uniformly distributed?	
II) Intergroup heterogeneity	Are the frequencies of the [that] in the EN-PT^{EU} and EN-ES^{EU} OSRAs statistically significantly different?	
III) Cross-language congruity	Are the frequencies of [that] and [que] in the EN-PT^{EU} and PT-EU OSRAs statistically similar?	Are the frequencies of [that] and [que] in the EN-ES^{EU} and ES-EU OSRAs statistically similar?
IV) Intralingual contrast	Are the frequencies of [that] in the EN-PT^{EU} and EN-GB OSRAs statistically significantly different?	Are the frequencies of [that] in the EN-ES^{EU} and EN-GB OSRAs statistically significantly different?

Since the data is not normally distributed and has outliers, the Mann-Whitney test is used to assess for a mean difference between the groups. Questions I and II are answered together. The level of significance used is $p < .05$ for questions II and IV. Because questions I and III look for uniformity and congruity, respectively, a result of *p* > .05 is associated with an L1 effect. Table 43 below shows the results and mean ranks of all comparisons.

	L1 Influence Effects								
	I) Intra-L1 homogeneity (Levene's test)	II) Inter-L1 heterogeneity (Mann-Whitney Test)	EN-PT^{EU} and EN-ES^{EU} similar in variance but different in means?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two L1 influence effects found for EN-PT^{EU}?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two L1 influence effect found for EN-ES^{EU}?
	Corpora Examined			**Corpora Examined**			**Corpora Examined**		
Subordinating conjunction	EN-PT^{EU} vs. EN-ES^{EU}	EN-PT^{EU} vs. EN-ES^{EU}		EN-PT^{EU} vs. PT-EU	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. ES-EU	EN-ES^{EU} vs. EN-GB	
	Reference *p* values			**Reference *p* values**			**Reference *p* values**		
	p > .05 **AND**	*p* < .05?		*p* > .05?	*p* < .05?		*p* > .05?	*p* < .05?	
that/que/que	*F* = .093 *p* = .761	*Z* = -2.502 *p* = .012 Mean ranks: EN-PT^{EU}= 56.71 EN-ES^{EU}= 73.16	*yes*	*Z* = -1.915 *p* = .055 Mean ranks: EN-PT^{EU}= 70.20 PT-EU= 57.71	*Z* = -.085 *p* = .932 Mean ranks: EN-PT^{EU}= 64.72 EN-GB= 65.28	*yes*	*Z* = -.974 *p* = .330 Mean ranks: EN-ES^{EU}= 68.72 ES-EU= 62.28	*Z* = -2.778 *p* = .005 Mean ranks: EN-ES^{EU}= 74.67 EN-GB= 56.33	*yes*

Table 43 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups in relation to the subordinating conjunction "that" and its equivalents "que" and "que" in Portuguese and Spanish

As can be seen in Table 43, for both the EN-PT^{EU} and the EN-ES^{EU} groups, at least two effects of L1 influence are found. The groups have similar within-group variances but are significantly different in relation to the mean values of the number of occurrences of the subordinating conjunctions “that”. Moreover, for both groups, the cross-language congruity effect is found, but only one, i.e., the EN-ES^{EU}, has intralingual contrast with the L1 English authors. In other words, the Portuguese authors writing OSRAs in English use the subordinating conjunction “that” as frequently as the L1 English authors use that conjunction when writing in their L1. Moreover, even though the Portuguese authors writing OSRAs in English use the subordinating conjunction “that” more frequently than they use the subordinating conjunction “que” when writing OSRAs in Portuguese, this difference is not significant. On the other hand, the Spanish authors writing OSRAs in English use the subordinating conjunction “that” as frequently as they use the subordinating conjunction “que” when writing OSRAs in Spanish, and significantly more frequently than the L1 English authors writing OSRAs in their L1.

From the analysis of the parsed corpora files, no other differences can be assessed between the groups in relation to the subordinating conjunction [that] besides the fact that the tagging in Portuguese and Spanish is far more specific than that obtained from the texts written in English, as shown in table Table 44.

Nº	Tags	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	KS @SUB #			911	970	1249
2	KS @SUB @#FS-<ACC	419	653			
3	KS @SUB @#FS-	71	75			
4	KS @SUB @#FS-P<	40	78			
5	KS @SUB @#FS-<SUBJ	75	31			
6	KS @SUB	53	48			
7	KS @COM @#AS-KOMP<		86			
8	KS @PRT-AUX<	6	33			
9	KS @COM @#FS-KOMP<		37			
10	KS @SUB @#FS-KOMP<	6	26			
11	KS @SUB @#FS-<ADVL	13	17			
12	KS @UTT @#AS-KOMP<	23				
13	KS @SUB @#FS-P	3	8			
14	KS @SUB @#FS-<SC	7	3			
15	KS @SUB @#FS-A	6	2			
16	KS @COM @#FS-KOMP		8			

17	KS @SUB @#FS-N	2	4			
18	KS @AS-				4	2
19	KS @SUB @#FS-SUBJ>	2	3			
20	KS @SUB @#FS-KOMP	2	2			
21	KS @COM @#AS-KOMP		2			
22	KS @COM @#AS-KOMPAS<		2			
23	KS @>A	1				
24	KS @SUB @#FS-PA	1				
25	KS @SUB @#FS--PASS	1				
26	KS @UTT @#AS-KOMPN	1				
27	KS &afterpar @SUB			1		
28	KS @AS-ADVL>					1
29	KS @COM @#AS-KOMP @AS<		1			
30	KS @COM @#AS-KOMPAS< @AS<		1			
31	KS @SUB DET S @>N			1		

Table 44 – Tags of the subordinating conjunctions in the five corpora

The tagging @SUB is the most frequent syntactic function described in both the PT-EU and the ES-EU corpora, with 699 and 950 occurrences, corresponding to 96% and 85%, respectively. Therefore, the subordinating function of the conjunction “que” is the most frequent in both corpora, just as it is in both the non-L1 English corpora.

The subordinator “that” in the English corpora does not function as part of the subordinate clause and it correspond mostly to so-called comment clauses that are preceded by transitive verbs (Quirk et al. 1985: 510; 1006). Below, Table 45 shows a list of the most frequent transitive verbs containing the subordinating conjunction “that” with higher numbers of occurrences in the EN-PT^{EU}/EN-ES^{EU} corpora in bold.

Nº	Verb with [that]	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	suggest that …	158	131	**177**
2	show that …	110	**168**	**177**
3	demonstrate that …	43	**51**	**74**
4	find that …	30	**62**	**56**
5	report that …	33	**51**	**37**
6	observe that …	15	**35**	**55**
7	note that …	25	11	24
8	reveal] that	8	**27**	**16**
9	hypothesi(s/z)e[17] that …	12	**16**	**22**
10	know that …	4	**12**	**34**

[17] To indicate that the numbers reflect both orthographic variances, i.e., the American and British.

11	confirm that …	12	**14**	**22**
12	indicate that …	20	3	17
13	conclude that …	6	6	**7**
14	acknowledge that …	10	4	1

Table 45 – Most frequent transitive verbs followed by the subordinating conjunction [that] in the English corpora

Overall, OSRAs written in English by the L1 authors have 47 different transitive verbs followed by the subordinating conjunction "that", while OSRAs written in English by the Portuguese/Spanish authors contain 41 and 42 different verbs followed by the subordinating conjunction "that", respectively. However, the total number of occurrences of these verbs in OSRAs written in English by the Portuguese/Spanish authors is 658 and 776, respectively, while in the EN-GB corpus, the total number of occurrences of transitive verbs followed by the subordinator "that" is 548.

These syntactic constructions in English have their equivalents in Portuguese and Spanish. However, the frequencies found in the PT-EU/ES-EU corpora are not as high as those found in English, as shown in Table 46 below.

N	Verb with [that]	PT-EU	ES-EU
1	indicar que	20	34
2	observar que	9	40
3	considerar que	23	17
4	sugerir que	27	10
5	demostrar que	3	27
6	destacar que	4	23
7	afirmar que	11	5
8	creer que	1	14
9	evidenciar que	5	6
10	esperar que	6	3
11	comprobar que	-	8
12	asumir que	-	5
13	recomendar que	3	2
14	supor/suponer que	2	1

Table 46 – Most frequent transitive verbs followed by the subordinating conjunction [que] in the Portuguese and Spanish corpora

Although, quantitatively, there is evidence of an L1 effect in OSRAs written in English by the Portuguese and especially by the Spanish authors associated with the subordinating conjunction "that", such effect is not maintained in specific syntactic constructions such as the one examined above with transitive verbs. The frequency of the subordinating conjunction "that" after transitive verbs that convey comments may not mark L1 influence, but it may function as a marker of hedging mechanisms (Quirk et al. 1985: 1113) of the non-L1 English authors in the CoRA.

4.4. Variables with only one effect of L1 influence– Intralingual Contrast

This section describes the variables whose results indicate intralingual contrast, i.e., the L1 Portuguese/Spanish authors writing OSRAs in English differ significantly from the L1 English authors writing in their L1 in relation to these variables.

4.4.1. V1: number of commas

After examining the samples of V1, the following descriptive statistics are obtained:

V1: frequency of commas (all values in commas/1000 tokens)							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	58,48	1,31	55,86	61,09	58,30	111,51	10,56
ES-EU	47,89	1,27	45,35	50,43	47,50	103,84	10,19
EN-GB	47,94	1,71	44,53	51,35	45,00	189,34	13,76
EN-PT^{EU}	50,62	1,57	47,49	53,74	50,00	159,52	12,63
EN-ES^{EU}	52,29	1,25	49,79	54,80	53,00	102,21	10,11

The lowest median standardized value of commas is in the EN-GB corpus, and the highest is in the PT-EU corpus.

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V1 of the EN-PT^{EU} and EN-ES^{EU} OSRAs are no significantly different as indicated by Levene's test**,** *F* = 3.587, *p* = 0.60**;**
II. There are no statistically significant differences in the frequency of commas between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = .836, *p* = .405.

Both groups have very similar means of commas per thousand words (*MD* = 1.68; *SED* = 2.01; 95% CI = 5.65 to -2.29 commas per thousand words);

III. There are statistically significant differences in the frequency of commas between the EN-PT^{EU} and PT-EU groups, *t*(128) = 3.852, *p* = .001.
The PT-EU OSRAs contain significantly more commas per thousand words than the EN-PT^{EU} OSRAs (*MD* = 7.86; *SED* = 2.04; 95% CI = 11.90 to 3.82 commas per thousand words);

There also are statistically significant differences in the frequency of commas between the EN-ES^{EU} and ES-EU groups, *t*(127) = 2.463, *p* = .015.
EN-ES^{EU} OSRAs have significantly more commas per thousand words than the ES-EU OSRAs (*MD* = 4.40; *SED* = 1.79; 95% CI = 7.94 to .866 commas per thousand words);

IV. There are no statistically significant differences in the frequency of commas between the EN-PT^{EU} and EN-GB groups, *t*(128) = 1.155, *p* = .250.
Both groups have very similar means of commas per thousand words (*MD* = 2.68; *SED* = 2.32; 95% CI = 7.26 to -1.91 commas per thousand words).

There are statistically significant differences in the frequency of commas between the EN-ES^{EU} and EN-GB groups, *t*(117.477) = 2.055, *p* = .042.
The EN-ES^{EU} group uses significantly more commas per thousand words than the EN-GB group (*MD* = 4.35; *SED* = 2.12; 95% CI = 8.55 to .159 commas per thousand words).

The following table summarizes the effects of L1 influence identified:

Effect of L1 Influence	**EN-PT^{EU}**	**EN-ES^{EU}**
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	--	--
IV. Intralingual contrast	--	✓

From the quantitative perspective, no effects of L1 influence can be argued for the PT-EU or the ES-EU authors in relation to the number of commas they use when writing in

English since no effects are found for the first group of authors (EN-PT^{EU}), and only one is found for the second group of authors (EN-ES^{EU}).

The results of Levene's test performed to determine homogeneity of variances in the EN-PT^{EU} and the EN-ES^{EU} groups indicate that OSRAs written in English by the PT-EU and the ES-EU authors are homogeneous in their V1 internal distribution. In other words, these authors behave uniformly within their groups in relation to the number of commas they use when writing in English. However, these groups do not differ significantly in relation to the frequency of use of commas. That is, no intergroup heterogeneity is found between the EN-PT^{EU} and the EN-ES^{EU} OSRAs. The mean values of commas per thousand words are not significantly different between the two groups (p > .05). The Portuguese authors use 50.62 (*SD* = 12.63; *SE* = 1.57) and the Spanish authors use 52.29 (*SD* = 10.11; *SE* = 1.25) commas per thousand words, on average.

Additionally, the comparison between the EN-PT^{EU} authors and the PT-EU authors shows significant differences between those groups in relation to the frequency of commas. The Portuguese authors use significantly more commas when writing in Portuguese (mean value = 58.48 commas per thousand words) than when writing in English (mean value = 50.62 per thousand words). Similarly, the comparison of the EN-ES^{EU} and the ES-EU authors shows significant differences in the frequency of use of commas between the Spanish authors writing in their L1 and the Spanish authors writing in English. However, contrary to what is observed in the Portuguese group, it is the group of the Spanish authors writing in English that uses more commas per thousand words.

Finally, the comparison between the EN-PT^{EU} and the EN-GB groups shows no intralingual contrast between those groups of authors in relation to the frequency of use of commas since there are no significant differences between their mean values of commas per thousand words. However, the comparison of the EN-ES^{EU} and the EN-GB groups shows a significant difference between the Spanish authors writing in English and the L1 English authors writing in their language in relation to the frequency of commas, with the non-L1 group (EN-ES^{EU}) using significantly more commas per thousand words.

In short, the Portuguese group significantly reduces the number of commas they use when writing in English in relation to their L1 counterpart writing in Portuguese and comes closer enough to the L1 English authors writing in their L1 to not differ significantly from them in relation to the frequency of use of commas. However, the Spanish group significantly increases the number of commas they use when writing OSRAs in English in relation to their L1 counterpart writing in Spanish, but then also with in relation to the L1 English authors writing in their L1. Although both the Portuguese and Spanish authors use more commas than the L1 English authors when writing OSRAs in English, this difference is not significant in the Portuguese group and, despite being significant in the Spanish group, has a small effect size (Cohen's *d* = 0.379). In any event, the difference observed between the L1 English authors and the Spanish authors writing in English in relation to the frequency of use of commas cannot be said to be influenced by the use the Spanish authors make of commas in their L1 because if that were the case, the Spanish authors writing in English would have had a mean value of comma frequency similar to the L1 English authors.

4.4.2. V7: number of 11-to-15-letter words

After examining the samples of V7, the following descriptive statistics are obtained:

V7: frequency of 11 to 15-letter words (all values in words/1000 tokens)							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	92,48	2,03	88,42	96,54	90,30	268,30	16,38
ES-EU	84,36	1,89	80,59	88,12	85,30	231,04	15,20
EN-GB	62,21	2,01	58,20	66,22	61,40	262,12	16,19
EN-PT^{EU}	68,50	1,99	64,53	72,47	66,70	256,64	16,02
EN-ES^{EU}	68,20	2,02	64,16	72,23	65,50	265,04	16,28

The results of the independent sample *t*-test indicate that:

I. The variances of the means of V7 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test**,** *F* = .045, *p* = .833);

II. There are no statistically significant differences in the frequency of 11-to-15-letter words between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = .106, *p* = .916.

Both groups have almost identical means of 11-to-15-letter words (*MD* = .300; *SED* = 2.83; 95% CI = 5.91 to -5.31 11-to-15-letter words per thousand words);

III. There are statistically significant differences in the frequency of 11-to-15-letter words between the EN-PT^{EU} and PT-EU groups, *t*(128) = 8.437, *p* = .001.
The PT-EU OSRAs contain significantly more 11-to-15-letter words than the EN-PT^{EU} OSRAs (*MD* = 23.98; *SED* = 2.84; 95% CI = 29.60 to 18.36 11-to-15-letter words per thousand words);

There are statistically significant differences in the frequency of 11-to-15-letter words between the EN-ES^{EU} and ES-EU groups, *t*(128) = 5.849, *p* = .001.
The ES-EU OSRAs have significantly more 11-to-15-letter words than the EN-ES^{EU} OSRAs (*MD* = 16.16; *SED* = 2.76; 95% CI = 21.63 to 10.69 11-to-15-letter words per thousand words);

IV. There are statistically significant differences in the frequency of 11-to-15-letter words between the EN-PT^{EU} and EN-GB groups, *t*(128) = 2.227, *p* = .028.
The EN-PT^{EU} sample has significantly more 11-to-15-letter words per thousand words than the EN-GB sample (*MD* = 6.29; *SED* = 2.82; 95% CI = 11.88 to .700 11-to-15-letter words per thousand words).

There are statistically significant differences in the frequency of 11-to-15-letter words between the EN-ES^{EU} and EN-GB groups, *t*(128) = 2.103, *p* = .037.
The EN-ES^{EU} sample has significantly more 11-to-15-letter words per thousand words than the EN-GB sample (*MD* = 5.99; *SED* = 2.85; 95% CI = 11.63 to .355 11-to-15-letter words per thousand words).

The following table summarizes the effects of L1 influence found in each group of OSRAs:

Effect of L1 Influence	**EN-PT^{EU}**	**EN-ES^{EU}**
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	--	--
IV. Intralingual contrast	✓	✓

In OSRAs written in English by the Portuguese or the Spanish L1 authors, the frequency of 11-to-15-letter words is not influenced by the frequency these authors make of this variable in their respective L1s. Both the Portuguese and Spanish authors use significantly fewer 11-to-15-letter words when writing OSRAs in their L1s than when they write OSRAs in English. Both groups of authors increase the frequency of V7 when writing in English, and this increment sets them apart also from the L1 English authors writing OSRAs in their L1, i.e., in English.

4.4.3. V8: number of definite articles

After examining the samples of V8, the following descriptive statistics are obtained:

V8: frequency of definite articles (all values in words/1000 tokens)							
CORPORA	**Mean**	**Std. Error**	**95% Confidence Interval for Mean**		**Median**	**Variance**	**Std. Deviation**
			Lower Bound	**Upper Bound**			
PT-EU	129,94	2,00	125,94	133,94	133,88	260,82	16,15
ES-EU	121,30	1,97	117,37	125,23	120,62	251,54	15,86
EN-GB	45,82	1,42	42,98	48,67	44,42	132,02	11,49
EN-PT^{EU}	49,49	1,69	46,11	52,88	47,70	186,60	13,66
EN-ES^{EU}	53,62	1,77	50,08	57,16	55,71	203,63	14,27

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V8 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test, *F* = .657, *p* = .419);

II. There are no statistically significant differences in the frequency of definite articles between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = 1.685, *p* = 0.94.

Both groups have similar means of definite articles per thousand words (*MD* = 4.13; *SED* = 2.45; 95% CI = 8.98 to -.720 definite articles per thousand words);

III. There are statistically significant differences in frequency of definite articles between the EN-PT^{EU} and PT-EU groups, *t*(128) = 30.668, *p* = .001.

The PT-EU OSRAs contain significantly more definite articles per thousand words than the EN-PT^{EU} OSRAs (*MD* = 80.44; *SED* = 2.62; 95% CI = 85.63 to 75.25 definite articles per thousand words);

There are statistically significant differences in the frequency of definite articles between the EN-ES^{EU} and ES-EU groups, *t*(128) = 25.57, *p* = .001.

The ES-EU OSRAs have significantly more definite articles per thousand words than the EN-ES^{EU} OSRAs (*MD* = 67.68; *SED* = 2.65; 95% CI = 72.92 to 62.44 definite articles per thousand words);

IV. There are no statistically significant differences in the frequency of definite articles between the EN-PT^{EU} and EN-GB groups, *t*(128) = 1.657, *p* = .100.
Both groups have similar means of definite articles per thousand words (*MD* = 3.67; *SED* = 2.21; 95% CI = 8.05 to -.711 definite articles per thousand words).

There are statistically significant differences in the frequency of definite articles between the EN-ES^{EU} and EN-GB groups, *t*(122.396) = 3.430, *p* = .001.
The EN-ES^{EU} sample has significantly more definite articles per thousand words than the EN-GB sample (*MD* = 7.96; *SED* = 2.27; 95% CI = 12.29 to 3.30 definite articles per thousand words).

The following table summarizes the effects of L1 influence found in the groups:

<table>
<tr><th>Effect of L1 Influence</th><th>EN-PT^{EU}</th><th>EN-ES^{EU}</th></tr>
<tr><td>I. Intra-L1 homogeneity</td><td colspan="2">--</td></tr>
<tr><td>II. Inter-L1 heterogeneity</td><td colspan="2">--</td></tr>
<tr><td>III. Cross-language congruity</td><td>--</td><td>--</td></tr>
<tr><td>IV. Intralingual contrast</td><td>--</td><td>✓</td></tr>
</table>

The frequency of definite articles in OSRAs written in English by the Portuguese or by the Spanish authors does not seem to be influenced by the frequency with which these authors use definite articles when writing OSRAs in their respective L1s. Both the Portuguese and the Spanish authors use significantly more definite articles when writing OSRAs in their L1 than when they write OSRAs in English.

In the Portuguese group, the decrease in the frequency of definite articles when writing in English gets them closer to the use the L1 English authors make when writing in their L1, and no significant differences are observed between these groups.

Spanish authors, however, reduce the frequency of use of definite articles when writing OSRAs in English to a mean value that sets a significant difference in relation to the

Spanish authors writing OSRAs in Spanish, but not in relation to the L1 English authors writing in English and therefore, the mean values of definite articles per thousand words are significantly different between the EN-ES^{EU} and the EN-GB groups.

4.4.4. V17: number of adverbs

After examining the samples of V17, the following descriptive statistics are obtained:

V17: frequency of adverbs (all values in words/1000 tokens)							
CORPORA	Mean	Std. Error	95% Confidence Interval for Mean		Median	Variance	Std. Deviation
			Lower Bound	Upper Bound			
PT-EU	47,05	1,10	44,85	49,25	45,50	79,03	8,89
ES-EU	33,17	0,96	31,25	35,10	31,70	60,37	7,77
EN-GB	46,70	1,38	43,95	49,45	46,40	122,99	11,09
EN-PT^{EU}	44,47	1,14	42,18	46,76	43,50	85,19	9,23
EN-ES^{EU}	41,43	0,86	39,71	43,15	41,70	48,16	6,94

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V17 of the EN-PT^{EU} and EN-ES^{EU} groups are significantly different (Levene's test, *F* = 4.755, *p* = .031);

II. There are statistically significant differences in the frequency of adverbs between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(118.822) = 2.124, *p* = .036.
The EN-PT^{EU} group uses significantly more adverbs per thousand words than the EN-ES^{EU} group (*MD* = 3.04; *SED* = 1.43; 95% CI = 5.88 to 2.06 adverbs per thousand words);

III. There are no statistically significant differences in the frequency of adverbs between the EN-PT^{EU} and PT-EU groups, *t*(128) = 1.624, *p* = .107.
Both groups use adverbs at similar frequencies (*MD* = 2.58; *SED* = 1.59; 95% CI = 5.73 to -.564 adverbs per thousand words);

There are statistically significant differences in the frequency of adverbs between the EN-ES^{EU} and ES-EU groups, *t*(128) = 6.391, *p* = .001.

The EN-ES^{EU} group uses significantly more adverbs per thousand words than the ES-EU group (*MD* = 8.26; *SED* = 1.29; 95% CI = 10.81 to 5.70 adverbs per thousand words);

IV. There are no statistically significant differences in the frequency of adverbs between the EN-PT^{EU} and EN-GB groups, *t*(128) = 1.248, *p* = .214.
Both groups have similar means of adverbs per thousand words (*MD* = 2.23; *SED* = 1.79; 95% CI = 5.77 to -1.31 adverbs per thousand words).

There are statistically significant differences in the frequency of adverbs between the EN-ES^{EU} and EN-GB groups, *t*(107.420) = 3.251, *p* = .002.
The EN-GB group uses significantly more adverbs per thousand words than the EN-ES^{EU} group (*MD* = 5.28; *SED* = 1.62; 95% CI = 8.49 to 2.06 adverbs per thousand words).

The following table summarizes the effects of L1 influence found in the groups:

Effect of L1 Influence	**EN-PT^{EU}**	**EN-ES^{EU}**
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	✓	--
IV. Intralingual contrast	--	✓

The effects of L1 influence in relation to the frequency of adverbs are found for both the EN-PT^{EU} and the EN-ES^{EU} groups. For the EN-PT^{EU} group, cross-language congruity is found, and for the EN-ES^{EU}, intralingual contrast is found. In other words, the Portuguese authors writing OSRAs in English maintain the frequency of use of adverbs they show when writing OSRAs in their L1. Similarly, the Portuguese authors writing OSRAs in English use adverbs as frequently as the L1 English authors. On the other hand, the Spanish authors writing OSRAs in English use adverbs significantly more frequently than the Spanish authors writing OSRAs in their L1, but not as frequently as the English authors writing in their L1.

Both groups (EN-PT^{EU} and EN-ES^{EU}) are analyzed linguistically to understand if these results reflect the use of specific adverbs and, therefore, may function as L1 influence markers. For that, all adverbs found in at least 50% of all OSRAs within each corpus are extracted into Table 47 together with the total corresponding frequency of each adverb, as shown below.

Rank	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Threshold
1	[não] n=943	[no] n=1002	[not] n=765	[not] n=625	[not] n=831	**95%**
2	[muito] n= 778	[más] n=481	[much] n=468	[also] n=467	[also] n=519	
3	[como] n=560	[como] n=258	[there] n=441	[much] n=376	[however] n=330	
4	[também] n=276	[también] n=195	[also] n=348	[significantly] n=257	[much] n=296	**90%**
5	[apenas] n=217	[sin=embargo] n=157	[however] n=272	[only] n=250	[only] n=217	
6	[ainda] n=160	[además] n=139	[only] n=231	[however] n=238	[significantly] n=203	
7	[no=entanto] n=156	[cuando] n=122	[both] n=170	[there] n=229	[when] n=203	**85%**
8	[estatístico] n= 129	[tanto] n=105	[when] n=159	[when] n=228	[as] n=198	**80%**
9	[cerca=de] n=108	[muy] n=101	[as] n=145	[both] n=156	[there] n=192	**75-70%**
10	[quando] n=108	[así=como] n=86	[where] n=128	[as] n=152	[thus] n=184	**65%**
11	[assim] n=87	[respectivamente] n=59	[significantly] n=126	[respectively] n=126	[therefore] n=167	
12	[nomeadamente] n=87		[therefore] n=122	[thus] n=119	[both] n=166	
13	[contudo] n=87		[previously] n=111	[previously] n=89	[in=addition] n=139	**60%**
14	[já] n=80		[respectively] n=99	[moreover] n=86	[moreover] n=112	
15	[através] n=78		[little] n=88	[therefore] n=77	[previously] n=112	**55%**
16	[pouco] n=70		[very] n=84	[still] n=75	[respectively] n=109	
17	[bem=como] n=61		[for=example] n=78	[furthermore] n=72	[very] n=82	**50%**
18	[só] n=55		[particularly] n=73	[namely] n=71	[furthermore] n=73	
19			[far] n=72	[very] n=65	[where] n=71	
20			[either] n=71	[either] n=56	[out] n=57	
21			[here] n=64	[nevertheless] n=54		
22				[mainly] n=53		
23				[recently] n=52		

Table 47 – Most frequent adverbs in the CoRA (present in 50% of the OSRAs) ranked by the number of occurrences and with corresponding total frequency in the corresponding corpus.

As can be seen, the L1 Portuguese authors writing in English (EN-PT^{EU}) are those with the highest number of adverbs that are present in 50% of the corresponding corpus; whereas the L1 Spanish authors writing in English (EN-ES^{EU}) are those with the smallest number of adverbs present in 50% of the corresponding corpus, despite the difference with the EN-GB group being rather small. Similarly, the L1 Portuguese authors writing OSRAs in their L1 are those with the largest number of adverbs present in 50% of the corresponding corpus, while the L1 Spanish authors writing OSRAs in their L1 have a smaller number of adverbs distributed in 50% of the corresponding corpus. Finally, the three English corpora contain more adverbs found in 50% of the OSRAs of each corpus than the L1 Portuguese and Spanish corpora.

After these first overall observations are obtained, the three corpora of authors writing OSRAs in English are analyzed, looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. Upon analysis, thirteen adverbs (see Table 48) are deemed unlikely to function as NLID markers given their similar ranks and percentage in the corresponding corpus, despite differences in the number of occurrences between the groups.

N	Adverb	Corpus	Occurrences in corpus	Rank	% in corpus
1	very	EN-GB	84	16	0,05
		EN-PT^{EU}	65	19	0,04
		EN-ES^{EU}	82	17	0,04
2	therefore	EN-GB	122	12	0,07
		EN-PT^{EU}	77	15	0,05
		EN-ES^{EU}	167	11	0,09
3	previously	EN-GB	111	13	0,06
		EN-PT^{EU}	89	13	0,05
		EN-ES^{EU}	112	15	0,06
4	respectively	EN-GB	99	14	0,06
		EN-PT^{EU}	126	11	0,08
		EN-ES^{EU}	109	16	0,06
5	recently	EN-GB	50	33	0,03
		EN-PT^{EU}	52	23	0,03
		EN-ES^{EU}	49	31	0,03
6	only	EN-GB	231	6	0,13
		EN-PT^{EU}	250	5	0,15
		EN-ES^{EU}	217	5	0,12
7	particularly	EN-GB	73	18	0,04
		EN-PT^{EU}	66	34	0,04
		EN-ES^{EU}	28	46	0,02
8	not	EN-GB	765	1	0,45
		EN-PT^{EU}	625	1	0,38
		EN-ES^{EU}	831	1	0,45

9	however	EN-GB	272	5	0,16
		EN-PT^{EU}	238	6	0,15
		EN-ES^{EU}	330	3	0,18
10	furthermore	EN-GB	64	26	0,04
		EN-PT^{EU}	72	17	0,04
		EN-ES^{EU}	73	18	0,04
11	both	EN-GB	170	7	0,10
		EN-PT^{EU}	156	9	0,10
		EN-ES^{EU}	166	12	0,09
12	far	EN-GB	72	19	0,04
		EN-PT^{EU}	51	38	0,03
		EN-ES^{EU}	46	35	0,02
13	as	EN-GB	145	9	0,08
		EN-PT^{EU}	152	10	0,09
		EN-ES^{EU}	198	8	0,11

Table 48 – Adverbs unlikely to function as NLID markers given their similar ranks and percentages in the corresponding corpus

Despite differences in ranks, seven adverbs are found in higher frequencies in the L1 English group than in the non-L1 English groups and therefore, these are analyzed to verify their potential to mark possible strategies of avoidance of use by the non-L1 authors. These adverbs are shown in Table 49.

N	Adverb	Corpus	Occurrences in corpus	Rank	% in corpus
1	[where]	EN-GB	128	10	0,07
		EN-PT^{EU}	71	19	0,04
		EN-ES^{EU}	48	30	0,03
2	[either]	EN-GB	71	20	0,04
		EN-PT^{EU}	56	20	0,03
		EN-ES^{EU}	30	48	0,02
3	[there]	EN-GB	441	3	0,26
		EN-PT^{EU}	229	7	0,14
		EN-ES^{EU}	192	9	0,10
4	[much]	EN-GB	468	2	0,27
		EN-PT^{EU}	376	3	0,23
		EN-ES^{EU}	296	4	0,16
5	[for=example]	EN-GB	78	17	0,05
		EN-PT^{EU}	12	55	0,01
		EN-ES^{EU}	19	50	0,01
6	[little]	EN-GB	88	15	0,05
		EN-PT^{EU}	48	39	0,03
		EN-ES^{EU}	36	46	0,02
7	[here]	EN-GB	64	21	0,04
		EN-PT^{EU}	33	51	0,02
		EN-ES^{EU}	49	39	0,03
8	out	EN-GB	69	23	0,04
		EN-PT^{EU}	23	62	0,01
		EN-ES^{EU}	57	20	0,03

Table 49 – Adverbs that could mark strategies of avoidance

The frequencies of the adverbs in Table 49 are compared to examine if there are significant differences between the groups. For all adverbs, the fourth L1 effect of the unified framework (Jarvis 2010, 2000) is tested for both EN-PT^{EU} and EN-ES^{EU} OSRAs, stated as follows:

Effect of L1 Influence	**L1 influence EN-PT^{EU} question**	**L1 influence EN-ES^{EU} question**
IV) Intralingual contrast	Are the frequencies of the adverbs where/ either/there/much/for=example/little/here in the EN-PT^{EU}/ EN-ES^{EU} and the EN-GB corpora statistically significantly different?	

Since the data is not normally distributed and has some outliers, the Mann-Whitney test is used to assess for a mean difference between the groups. The level of significance used is *p* < .05. Fourteen tests are carried out. The number of occurrences of all adverbs is normalized by 100. Table 50 shows the results obtained.

	IV - Intralingual contrast (Mann-Whitney Test)	L1 influence effect IV found for EN-PT^{EU}?	IV - Intralingual contrast (Mann-Whitney Test)	L1 influence effect IV found for EN-ES^{EU}?
	corpora examined			
Adverb	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. EN-GB	
	***p* reference value < .05**			
where	*Z* =-2.847 *p* = .004 *M* rank EN-GB = 39.50 *M* rank EN-PT^{EU} =26.34	*yes*	*Z* = -2.583 *p* = .010 *M* rank EN-GB= 43.36 *M* rank EN-ES^{EU}= 30.97	*yes*
either	*Z* = -.310 *p* = .756 *M* rank EN-GB = 36.88 *M* rank EN-PT^{EU} = 38.24	*no*	*Z* = -.771 *p* = .441 *M* rank EN-GB= 31.51 *M* rank EN-ES^{EU}= 28.48	*no*
there	*Z* =-4.577 *p* = .001 *M* rank EN-GB = 76.71 *M* rank EN-PT^{EU} = 46.34	*yes*	*Z* = -4.152 *p* = .001 *M* rank EN-GB= 70.12 *M* rank EN-ES^{EU}= 44.20	*yes*
much	*Z* = -1.850 *p* = .064 *M* rank EN-GB = 71.58 *M* rank EN-PT^{EU} = 59.42	*no*	*Z* = -2.973 *p* = .003 *M* rank EN-GB= 68.52 *M* rank EN-ES^{EU}= 49.74	*yes*
for=example	*Z* = -1.945 *p* = .089 *M* rank EN-GB = 26.76 *M* rank EN-PT^{EU} = 18.15	*no*	*Z* = -1.684 *p* = .092 *M* rank EN-GB= 29.47 *M* rank EN-ES^{EU}= 22.37	*no*

little	*Z* = -.756 *p* = .450 *M* rank EN-GB = 33.87 *M* rank EN-PT^{EU} = 30.50	*no*	*Z* = -.259 *p* = .795 *M* rank EN-GB= 29.38 *M* rank EN-ES^{EU}= 28.24	*no*
here	*Z* = -1.213 *p* = .225 *M* rank EN-GB = 22.52 *M* rank EN-PT^{EU} = 27.50	*no*	*Z* = -.149 *p* = .882 *M* rank EN-GB= 30.27 *M* rank EN-ES^{EU}= 29.65	*no*
out	*Z* = -2.339 *p* = .044 *M* rank EN-GB = 24.05 *M* rank EN-PT^{EU} = 16.58	*yes*	*Z* = -.962 *p* = .336 *M* rank EN-GB= 30.43 *M* rank EN-ES^{EU}= 26.50	*no*

Table 50 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups, indicating 4 adverbs with potential to mark strategies of avoidance

As can be seen, two adverbs, i.e., "where" and "there", are used significantly more frequently by the L1 (i.e., EN-GB) than by both non-L1 authors (EN-PT^{EU} and EN-ES^{EU}) writing OSRAs in English. Additionally, the adverb "much" is used by the L1 English authors (i.e., EN-GB) significantly more frequently than by the Spanish authors writing OSRAs in English (EN-ES^{EU}); and the adverb "out" is used by the L1 English authors (i.e., EN-GB) significantly more frequently than by the Portuguese authors writing OSRAs in English (EN-PT^{EU}).

Based on the significance of the results, the adverbs "where" and "there" are further analyzed for both groups (the EN-PT^{EU} and the EN-ES^{EU}), while the adverbs "much" and "out" are analyzed only for the EN-ES^{EU} group and the EN-PT^{EU} group, respectively. The analyses are based on the concordances of the parsed files. The concordances are obtained with WordSmith 7.0 (Scott 2018b), from which the syntactic structures containing the adverbs (ADV) are extracted.

The significant differences in the number of occurrences of the adverb "where" between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups are more evident in the syntactic structures shown in Figure 11, which are more frequent in the EN-GB than in the other two groups.

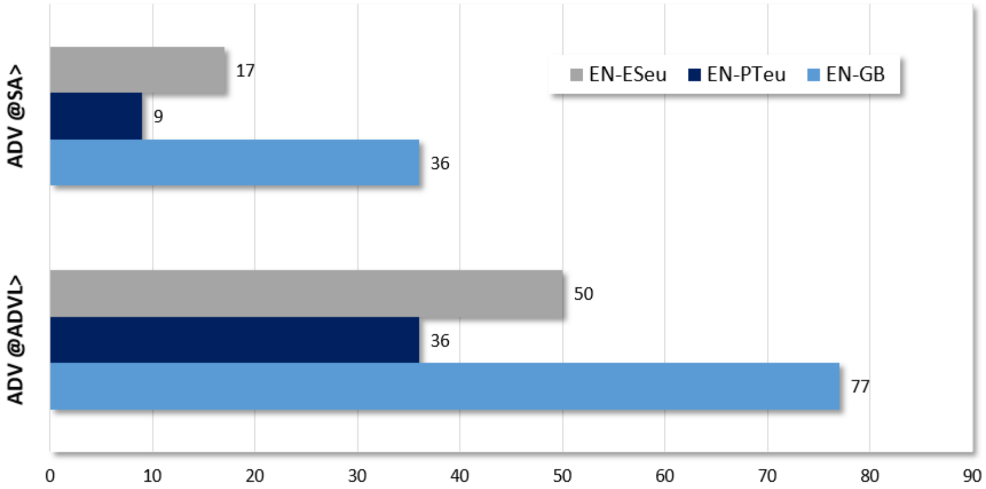


Figure 11 – Syntactic tags of the adverb [where] showing the significant differences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups

The significant differences in the number of occurrences of the adverb "there" between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups are more clearly shown by the syntactic tag [ADV @F-SUBJ>], shown in Table 51, which is more frequent in the EN-GB than in the other two groups.

ADV	Followed by a token with the syntactic function	Example	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	ADV @F-SUBJ>	[there] was/were/is …	433	227	188
3	ADV @ADVL>	[There], 54 bp exons are shown…	1		2
4	ADV @F-	Nor was [there] evidence…	2	1	
5	ADV &afterpar @F-SUBJ>	[there] are still fewer laboratories…	1	1	
6	ADV @N	…the causes of death [there] may differ from…	1		1

Table 51 – Syntactic tags of the adverb "there" showing the significant differences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups

The significant differences between the EN-GB and the EN-ES^{EU} groups in relation to the adverb "much" are more obvious in the use of the comparative structure [ADV COM @<A than], e.g., *much lower than* / *much larger than* / *much higher than*, as shown in Figure 12 below.

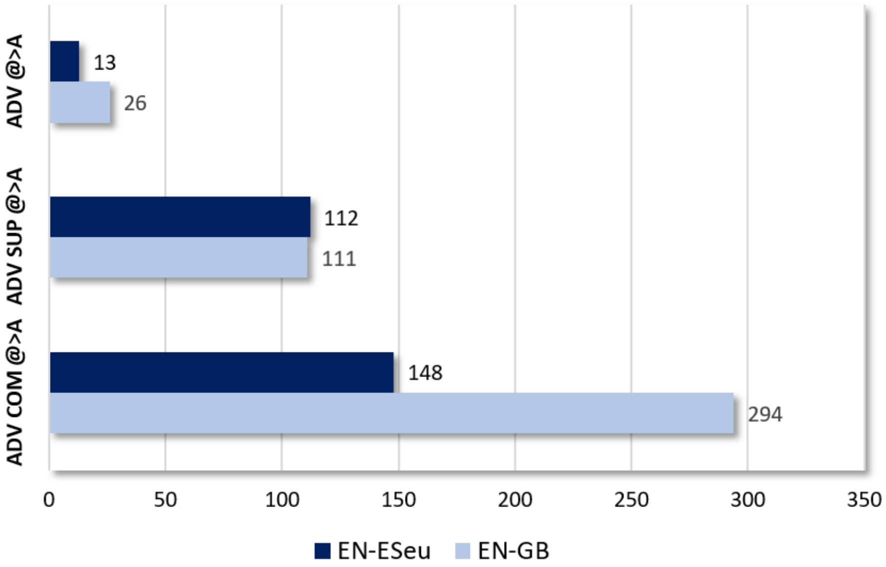


Figure 12 – Syntactic tag of the adverb [much] evincing the significant differences between the EN-GB and the EN-ES^{EU} groups

Finally, the significant differences between the EN-GB and the EN-PT^{EU} groups in relation to the adverb "out" are closely related to the frequency of phrasal verbs with "out" shown in Table 52 below.

N	Phrasal verb	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	rule out	38	4	20
2	carry out	20	11	25
3	point out	2	5	6
4	set out	3	1	1
5	drive out			2
6	find out	1		1
7	spread out	2		
8	roll out	1		
9	seek out	1		
10	stand out			1
11	start out			1
12	wash out		1	

Table 52 – Occurrences of phrasal verbs with "out" in English corpora

The second group of adverbs is examined to determine language transfer. The group comprises nine adverbs distributed in 50% or more of the OSRAs of one or both the non-L1 English corpora and whose number of occurrences and percentage in the corresponding corpus is higher in one or both the non-L1 English corpora than in the L1 English corpus.

Those adverbs are “also”, “in=addition”, “mainly”, “moreover”, “namely”, “nevertheless”, “significantly”, “still”, “thus”, and “when”.

After preliminary analysis, some of these adverbs are analyzed as a group with other synonyms found in the corpora because their translation from English into Portuguese/Spanish can adopt any of the equivalents found in the English corpora and the Portuguese/Spanish corpora. Those two adverb groups are:

- the group consisting of the adverbs “also”, “in=addition”, “moreover”, “additionally”, and “likewise”;
- the group consisting of the adverbs “thus”, “so”, “hence”, and “accordingly”.

After the equivalent(s) of all nine adverbs and the two adverbs groups in the L1 Portuguese and Spanish corpora are analyzed in terms of occurrences, ranks, and percentages in the corresponding corpus, it is found that two adverbs and one adverb group (Table 53) are less frequent in the PT-EU and the ES-EU corpora than in all English corpora, and therefore cannot be used as markers of NLID since the high frequency of their equivalents in OSRAs written in English by the Portuguese/Spanish authors does not mirror their frequency in the L1 Portuguese or the L1 Spanish corpora.

N	Adverb	Corpus	Occurrences in corpus	Rank	% in corpus
1	significativamente	PT-EU	89	12	0,06
	significativamente	ES-EU	87	12	0,05
	significantly	EN-GB	126	11	0,07
		EN-PT^{EU}	257	4	0,16
		EN-ES^{EU}	203	6	0,11
2	quando	PT-EU	108	10	0,08
	cuando	ES-EU	122	7	0,07
	when	EN-GB	159	8	0,09
		EN-PT^{EU}	228	8	0,14
		EN-ES^{EU}	203	7	0,11
3	também / além=disso / igualmente / adicionalmente]/ aliás / inclusive / além	PT-EU	347	4	0,24
	también / además / asímismo / igualmente/ adicionalmente	ES-EU	387	4	0,24
	also / in=addition / moreover / additionally / likewise	EN-GB	427	4	0,25
		EN-PT^{EU}	644	2	0,39
		EN-ES^{EU}	814	2	0,44

Table 53 – Adverbs with a higher number of occurrences in the non-L1 English corpora, but whose equivalents in the PT-EU and the ES-EU are not equally frequent

Therefore, after all the corpora are examined, five adverbs and one group of adverbs are considered to have the potential to function as NLID markers since they occur in higher frequencies in the non-L1 English corpora than in the L1 English corpus. Similarly, their equivalents in the L1 Portuguese or the L1 Spanish corpora, or both, also occur in frequencies higher than those in the EN-GB corpus and similar frequencies to those of the non-L1 English corpora. These adverbs and the group of adverbs are shown in Table 54.

N	Adverb	Corpus	Occurrences in corpus	Rank	% in corpus
1	assim / desta=forma / portanto / consequente / logo / por=conseguinte	PT-EU	139	11	0,10
	por=tanto / así / por=lo=tanto / en=consecuencia / así=pues / por=ende / por=consiguiente / por=esto	ES-EU	220	12	0,14
	thus / so / hence / accordingly	EN-GB	162	33	0,09
		EN-PT^{EU}	180	10	0,11
		EN-ES^{EU}	243	12	0,13
2	principalmente	PT-EU	18	57	0,01
	principalmente	ES-EU	48	22	0,03
	mainly	EN-GB	4	228	0,0023
		EN-PT^{EU}	53	22	0,03
		EN-ES^{EU}	55	26	0,03
3	nomeadamente	PT-EU	87	12	0,061
	concretamente	ES-EU	11	77	0,007
	namely	EN-GB	2	339	0,001
		EN-PT^{EU}	71	18	0,043
		EN-ES^{EU}	5	187	0,003
4	contudo / não=obstante	PT-EU	94	14	0,07
	no=obstante	ES-EU	46	23	0,03
	nevertheless	EN-GB	14	98	0,008
		EN-PT^{EU}	54	21	0,033
		EN-ES^{EU}	40	36	0,022
5	ainda / ainda=assim	PT-EU	172	6	0,12
	todavía / aún	ES-EU	55	53	0,03
	still	EN-GB	27	60	0,02
		EN-PT^{EU}	75	17	0,05
		EN-ES^{EU}	49	32	0,03

Table 54 – Adverbs that may function as NLID markers in OSRAs written in English by L1 Portuguese / L1 Spanish authors

These adverbs and the adverb group are examined to understand if the differences in the groups' frequencies are significant. For all adverbs and the adverb group, the following questions are asked for both EN-PT^{EU} and EN-ES^{EU} OSRAs.

Effect of L1 Influence	L1 influence EN-PT^{EU} questions	L1 influence EN-ES^{EU} questions
I) Intragroup homogeneity II) Intergroup heterogeneity	Are the frequencies of the adverb/adverb group in the EN-PT^{EU} / EN-ES^{EU} OSRAs uniformly distributed? and Are the frequencies of the adverb/adverb group in the EN-PT^{EU} and EN-ES^{EU} OSRAs statistically significantly different?	
III) Cross-language congruity	Are the frequencies of the adverb/adverb group in the EN-PT^{EU} and PT-EU OSRAs statistically similar?	Are the frequencies of the adverb/adverb group in the EN-ES^{EU} and ES-EU OSRAs statistically similar?
IV) Intralingual contrast	Are the frequencies of the adverb/adverb group in the EN-PT^{EU} and EN-GB OSRAs statistically significantly different?	Are the frequencies of the adverb/adverb group in the EN-ES^{EU} and EN-GB OSRAs statistically significantly different?

Questions in relation to effects I and II are answered together. Because the data are not normally distributed and in some cases there are outliers, the non-parametric alternative to the independent samples *t*-test, i.e., the Mann-Whitney test, is used to assess for a mean difference between the groups. The level of significance used is $p < .05$ for questions in relation to effects II and IV. Since questions in relation to the effects I and III look for uniformity and congruity, respectively, a result of $p > .05$ is associated with an existing L1 effect. Table 55 below shows the results and mean ranks of all comparisons.

Adverb/Adverb group	L1 Influence Effects								
	I) Intra-L1 homogeneity (Levene's test)	II) Inter-L1 heterogeneity (Mann-Whitney Test)	EN-PT^{EU} and EN-ES^{EU} similar in variance but different in means?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-PT^{EU}?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-ES^{EU}?
	Corpora Examined			Corpora Examined			Corpora Examined		
	EN-PT^{EU} vs. EN-ES^{EU}	EN-PT^{EU} vs. EN-ES^{EU}		EN-PT^{EU} vs. PT-EU	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. ES-EU	EN-ES^{EU} vs. EN-GB	
	Reference *p* values			Reference *p* values			Reference *p* values		
	p > .05 **AND**	*p* < .05?		*p* > .05?	*p* < .05?		*p* > .05	*p* < .05	
assim / desta=forma / portanto / consequente / logo / por=conseguinte por=tanto / así / por=lo=tanto / en=consecuencia / así=pues / por=ende / por=consiguiente / por=esto [thus / so / hence / accordingly	*F* = 4.912 *p* = .029	*Z* = -1.299 *p* = .194 Mean ranks: EN-PT^{EU}: 52.51 EN-ES^{EU}: 60.35	no	*Z* = -.981 *p* = .327 Mean ranks: EN-PT^{EU}: 56.25 PT-EU: 50.53	*Z* = -.451 *p* = .652 Mean ranks: EN-PT^{EU}: 49.34 EN-GB: 51.92	no	*Z* = -.575 *p* = .565 Mean ranks: EN-ES^{EU}: 52.88 ES-EU: 56.31	*Z* = -.807 *p* = .420 Mean ranks: EN-ES^{EU}: 53.57 EN-GB: 48.88	no
principalmente principalmente mainly	*F* = .560 *p* = .482	*Z* = -1.190 *p* = .234 Mean ranks: EN-PT^{EU}: 30.06 EN-ES^{EU}: 35.10	no	*Z* = -.654 *p* = .513 Mean ranks: EN-PT^{EU}: 24.20 PT-EU: 21.73	*Z* = -.362 *p* = .787 Mean ranks: EN-PT^{EU}: 18.67 EN-GB: 16.67	no	*Z* = -1.030 *p* = .303 Mean ranks: EN-ES^{EU}: 23.05 ES-EU: 27.15	*Z* = -.926 *p* = .414 Mean ranks: EN-ES^{EU}: 17.95 EN-GB: 12.83	no
nomeadamente / concretamente / namely	*F* = 3.508 *p* = .069	*Z* = -2.373 *p* = .024 Mean ranks: EN-PT^{EU}: 21.54 EN-ES^{EU}: 9.50	yes	*Z* = -.023 *p* = .982 Mean ranks: EN-PT^{EU}: 38.06 PT-EU: 37.95	*Z* = -1.547 *p* = .178 Mean ranks: EN-PT^{EU}: 19.12 EN-GB: 8.00	yes	*Z* = -.707 *p* = .768 Mean ranks: EN-ES^{EU}: 7.50 ES-EU: 8.25	*Z* = .000 *p* = 1.00 Mean ranks: EN-ES^{EU}: 4.00 EN-GB: 4.00	yes
contudo / não=obstante / no=obstante / nevertheless	*F* = .114 *p* = .737	*Z* = -.246 *p* = .806 Mean ranks: EN-PT^{EU}: 28.11 EN-ES^{EU}: 29.07	no	*Z* = -2.439 *p* = .015 Mean ranks: EN-PT^{EU}: 28.88 PT-EU: 39.80	*Z* = -.585 *p* = .620 Mean ranks: EN-PT^{EU}: 22.53 EN-GB: 20.25	no	*Z* = -.121 *p* = .904 Mean ranks: EN-ES^{EU}: 24.22 ES-EU: 23.79	*Z* = -.720 *p* = .550 Mean ranks: EN-ES^{EU}: 17.70 EN-GB: 15.40	no
ainda / ainda=assim / todavía / aún / still	*F* = .312 *p* = .578	*Z* = -.705 *p* = .481 Mean ranks: EN-PT^{EU}: 34.79 EN-ES^{EU}: 31.75	no	*Z* = -3.330 *p* = .001 Mean ranks: EN-PT^{EU}: 34.75 PT-EU: 52.64	*Z* = -2.025 *p* = .043 Mean ranks: EN-PT^{EU}: 32.33 EN-GB: 24.13	no	*Z* = -.444 *p* = .657 Mean ranks: EN-ES^{EU}: 29.05 ES-EU: 30.85	*Z* = -1.405 *p* = .160 Mean ranks: EN-ES^{EU}: 26.46 EN-GB: 21.75	no

Table 55 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups in relation to the five adverbs, and the adverbial group found to have the potential to function as NLID marker

Two effects of L1 influence are found only for one adverb, i.e., “namely” and its equivalents “nomeadamente” in Portuguese and “concretamente” in Spanish. Those effects refer to intra-L1 homogeneity, inter-L1 heterogeneity, and cross-language congruity. The fourth effect of L1 influence, i.e., intralingual contrast, is not found at significant levels.

The variance of the frequency of use of the adverb “namely” in OSRAs written in English by the L1 Portuguese authors (EN-PT^{EU}) does not differ from the variance of the frequency of use of the same adverb in OSRAs written in English by the L1 Spanish authors (EN-ES^{EU}), which means both corpora have a homogenous distribution of that adverb. Similarly, the means of the frequencies of “namely” in the EN-PT^{EU} and the EN-ES^{EU} OSRAs are statistically significantly different, meaning that the frequencies belong to different corpora. When compared to the corresponding L1 corpora, i.e., PT-EU and ES-EU, no significant differences are found either, which means that the L1 Portuguese authors and the L1 Spanish authors writing OSRAs in English use the adverb “namely” at frequencies that are similar to those of the equivalent adverbs “nomeadamente” and “concretamente” when writing OSRAs in their L1. Finally, despite the differences in the frequencies of the adverb “namely” in OSRAs written in English by the L1 and the non-L1 authors, not being statistically significant, as shown in Figure 13 below, graphically, it is possible to appreciate the distance between the groups in relation to this adverb.

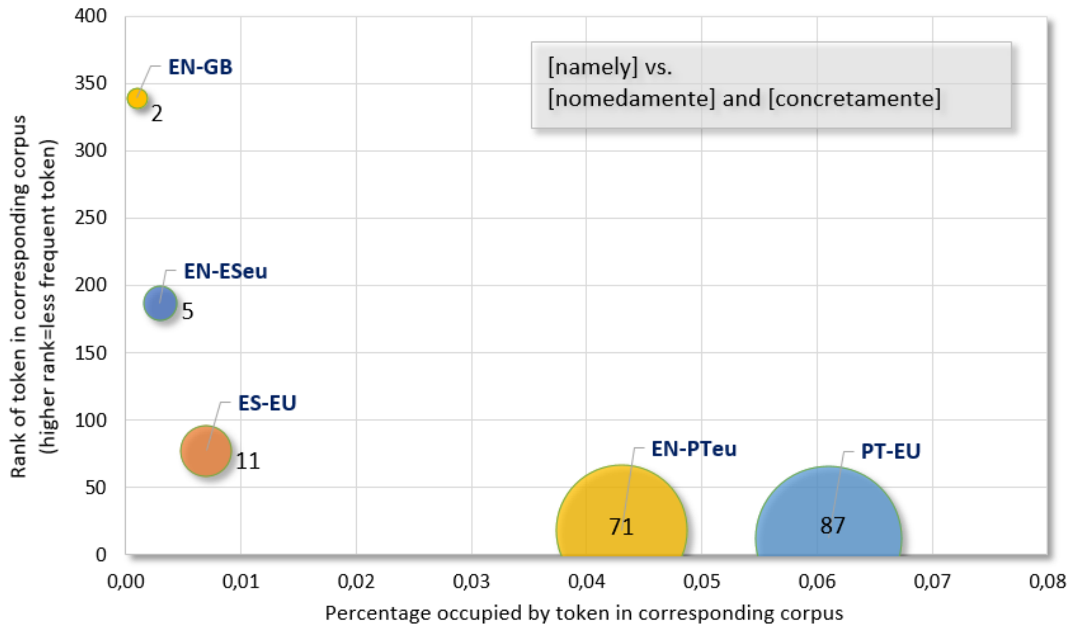


Figure 13 – Distribution of the adverb [namely] vs. [nomeadamente] and [concretamente] in the PT-EU and the ES-EU corpora of the CoRA

Also, a cross-tabulation of the data on the frequency of use of “nomeadamente” in Portuguese, “concretamente” in Spanish, and “namely” in all English corpora using SPSS (Table 56 below) shows that the Portuguese authors’ actual numbers of OSRAs containing a given number of occurrences of the referred adverb are better distributed with regards to the expected[18] numbers of OSRAs containing a given number of occurrences.

Corpus	Number of OSRAs with occurrences of the examined adverb	number of occurrences of “nomeadamente”/ “concretamente”/ “namely”							Total of OSRAs	Total Occurrences
		1.00	2.00	3.00	4.00	6.00	7.00	8.00		
PT-EU	actual	16	15	6	1	2	1		**41**	**87**
	expected	20	13	4	2	1	0,45	0,45		
ES-EU	actual	9	1						**10**	**11**
	expected	5	3	1	0,43	0,22	0,11	0,11		
EN-GB	actual	2							**2**	**2**
	expected	1	1	0,217	0,09	0,04	0,02	0,02		
EN-PT^{EU}	actual	13	13	4	3			1	**34**	**71**
	expected	17	11	4	1	1	0,37	0,37		
EN-ES^{EU}	actual	5							**5**	**5**
	expected	2	2	1	0,22	0,11	0,05	0,05		
Total		**45**	**29**	**10**	**4**	**2**	**1**	**1**	**92**	**176**

Table 56 – Distribution of the adverb "namely" in the English corpora, “nomeadamente" in the L1 Portuguese corpus, and "concretamente" in the L1 Spanish corpus according to the actual and the expected number of OSRAs per number of occurrences, calculated with SPSS

The analyses show that in relation to the category adverbs, there are effects of L1 influence associated with the overall frequency of adverbs, and also with the frequency of specific adverbs, which can be separated into two groups: adverbs that indicate significant differences between the groups in relation to possible strategies of avoidance (i.e., where, there, much, out) and adverbs that can be associated with language transfer (namely vs. nomeadamente).

[18] The expected numbers of OSRAS within each occurrence subgroup (i.e., 1.00, 2.00 etc.) is the number we would expect given the overall distribution of the data in the CoRA, calculated automatically by SPSS.

4.4.5. V18: number of nouns

After examining the samples of V18, the following descriptive statistics are obtained:

V18: frequency of nouns (all values in words/1000 tokens)							
CORPORA	Mean	Std. Error	95% Confidence Interval for Mean		Median	Variance	Std. Deviation
			Lower Bound	Upper Bound			
PT-EU	271,11	2,54	266,04	276,18	271,20	419,43	20,48
ES-EU	173,48	2,20	169,07	177,88	173,20	315,77	17,77
EN-GB	329,63	3,30	323,04	336,22	329,30	707,56	26,60
EN-PT^{EU}	344,51	3,38	337,77	351,26	345,20	741,47	27,23
EN-ES^{EU}	345,74	3,23	339,29	352,18	346,30	676,52	26,01

The results of the independent sample *t*-test indicate that:

I. The variances of the means of V18 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test, *F* = .929, *p* = .337);

II. There are also no statistically significant differences in the frequency of nouns between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = .262, *p* = .794.
Both groups have almost identical mean frequencies of nouns per thousand words (*MD* = 1.22; *SED* = 4.67; 95% CI = 10.46 to -8.02 nouns per thousand words);

III. There are statistically significant differences in the frequency of nouns between the EN-PT^{EU} and PT-EU groups, *t*(118.843) = 17.371, *p* = .001.
The EN-PT^{EU} group uses significantly more nouns per thousand words than the PT-EU group (*MD* = 73.40; *SED* = 4.23; 95% CI = 81.77 to 65.04 nouns per thousand words);

There also are statistically significant differences in the frequency of nouns between the EN-ES^{EU} and ES-EU groups, *t*(113.066) = 44.089, *p* = .001.
The EN-ES^{EU} OSRAs have significantly more nouns per thousand words than the ES-EU OSRAs (*MD* = 172.26; *SED* = 3.91; 95% CI = 180.00 to 164.52 nouns per thousand words);

IV. There are statistically significant differences in the frequency of nouns between the EN-PT^{EU} and EN-GB groups, *t*(128) = 3.153, *p* = .002.

The EN-PT^{EU} sample has significantly more nouns per thousand words than the EN-GB sample (*MD* = 14.89; *SED* = 4.72; 95% CI = 24.23 to 5.54 nouns per thousand words).

There also are statistically significant differences in the frequency of nouns between the EN-ES^{EU} and EN-GB groups, *t*(128) = 3.491, *p* = .001.
The EN-ES^{EU} sample has significantly more nouns per thousand words than the EN-GB sample (*MD* = 16.10; *SED* = 4.61; 95% CI = 25.24 to 6.98 nouns per thousand words).

The following table summarizes the effects of L1 influence found:

Effect of L1 Influence	EN-PT^{EU}	EN-ES^{EU}
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	--	--
IV. Intralingual contrast	✓	✓

For V18, only one effect of L1 influence is found for both the EN-PT^{EU} and the EN-ES^{EU} groups. The L1 Portuguese and the Spanish authors writing OSRAs in English (EN-PT^{EU} and EN-ES^{EU}) differ significantly from the L1 English authors writing OSRAs in their native language in relation to the frequency of use of nouns.

Both the EN-PT^{EU} and the EN-ES^{EU} groups use significantly more nouns per thousand words than the L1 English authors. Similarly, these two groups of non-L1 English authors also differ significantly from their respective L1 counterparts (PT-EU and ES-EU). The EN-PT^{EU} and the EN-ES^{EU} groups also use significantly more nouns per thousand words than the PT-EU and ES-EU authors writing OSRAs in their respective L1s.

Overall, the L1 Portuguese and the Spanish authors can be said to overdo the use of nouns in English. They increase the frequency of nouns when writing OSRAs in English to a point where they differ significantly from the L1 Portuguese and Spanish authors writing in their respective L1s and from the L1 English authors writing in their L1.

Table 57 below shows the nouns in the CoRA that are present in at least 50% of the OSRAs, ranked by the number of occurrences and with the corresponding total frequency in the corresponding corpus.

Rank	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Threshold
1	[%] n=2013	[%] n=1529	[%] n=1493	[%] n=1458	[patient] n=1311	95%
2	[estudo] n=931	[estudio] n=873	[patient] n=998	[patient] n=1416	[cell] n=1012	
3	[ano] n=498	[año] n=406	[study] n=728	[study] n=820	[study] n=951	
4	[caso] n=492	[caso] n=363	[year] n=510	[cell] n=653	[%] n=936	90%
5	[p] n=444	[grupo] n=346	[datum] n=465	[level] n=566	[level] n=812	
6	[risco] n=413	[edad] n=296	[group] n=460	[group] n=466	[P] n=554	
7	[grupo] n=407	[diferencia] n=252	[case] n=444	[result] n=399	[result] n=552	85%
8	[resultado] n=399	[casos] n=248	[figure] n=350	[disease] n=372	[effect] n=548	80%
9	[valor] n=347	[dato] n=241	[disease] n=347	[effect] n=310	[expression] n=521	75-70%
10	[população] n=288	[factor] n=222	[analysis] n=346	[age] n=298	[group] n=432	65%
11	[idade] n=285	[nivel] n=219	[effect] n=325	[table] n=289	[gene] n=364	
12	[fator] n=254	[p] n=218	[age] n=313	[risk] n=281	[treatment] n=330	
13	[prevalência] n=243	[muestra] n=188	[difference] n=313	[P] n=278	[response] n=310	60%
14	[doença] n=231	[análisis] n=184	[risk] n=310	[analysis] n=267	[activity] n=303	
15	[nível] n=228	[enfermedad] n=181	[rate] n=302	[case] n=266	[difference] n=293	55%
16	[saúde] n=212	[forma] n=151	[result] n=279	[control] n=260	[protein] n=289	
17	[análise] n=209	[efecto] n=149	[population] n=270	[year] n=260	[increase] n=262	50%
18	[diagnóstico] n=206	[centro] n=139	[table] n=254	[factor] n=229	[case] n=244	
19	[tratamento] n=194	[día] n=129	[treatment] n=246	[treatment] n=215	[role] n=236	
20	[tabela] n=183	[número] n=127	[number] n=245	[difference] n=208	[disease] n=223	
21	[número] n=180	[objetivo] n=119	[time] n=229	[datum] n=201	[tissue] n=222	
22	[diferença] n=179	[característica] n=116	[change] n=220	[population] n=195	[datum] n=220	
23	[tempo] n=173	[criterio] n=111	[increase] n=205	[rate] n=185	[factor] n=220	
24	[amostra] n=172	[diagnóstico] n=111	[level] n=203	[sample] n=173	[control] n=218	
25	[avaliação] n=169	[figura] n=94	[finding] n=184	[time] n=172	[table] n=211	
26	[associação] n=164	[limitación] n=89	[evidence] n=163	[increase] n=171	[number] n=210	
27	[dado] n=148	[comparación] n=86	[type] n=163	[association] n=170	[change] n=208	
28	[aumento] n=147	[método] n=83	[UK] n=157	[role] n=167	[model] n=191	
29	[maioria] n=145	[información] n=69	[factor] n=156	[number] n=162	[time] n=180	
30	[tipo] n=140	[mayoría] n=69	[health] n=154	[activity] n=150	[risk] n=176	
31	[forma] n=136	[cuenta] n=66	[outcome] n=141	[tissue] n=133	[analysis] n=174	

32	[alteração] n=119	[momento] n=60	[sample] n=140	[finding] n=132	[sample] n=173	**50%**
33	[utilização] n=113		[use] n=122	[function] n=130	[age] n=171	
34	[sexo] n=112		[test] n=118	[type] n=130	[function] n=171	
35	[limitação] n=110		[proportion] n=114	[model] n=127	[population] n=169	
36	[período] n=107		[range] n=91	[development] n=125	[year] n=168	
37	[trabalho] n=104		[approach] n=77	[use] n=123	[mechanism] n=162	
38	[média] n=100		[comparison] n=76	[condition] n=118	[system] n=160	
39	[informação] n=98		[research] n=73	[change] n=111	[development] n=149	
40	[literatura] n=97		[limitation] n=68	[presence] n=110	[type] n=139	
41	[característica] n=95		[part] n=52	[characteristic] n=71	[condition] n=133	
42	[facto] n=95			[limitation] n=68	[value] n=131	
43	[objetivo] n=80				[finding] n=123	
44	[relação] n=79				[process] n=112	
45	[autor] n=72				[presence] n=104	
46					[reduction] n=94	
47					[contrast] n=92	
48					[evidence] n=83	
49					[size] n=82	
50					[characteristic] n=80	
51					[use] n=79	
52					[fact] n=59	
53						

Table 57 – Most frequent nouns in the CoRA (present in at least 50% of the OSRAs) ranked by the number of occurrences and with the corresponding total frequency in corpus

As can be seen in Table 57, the Spanish authors writing OSRAs in English have the most extensive list of nouns distributed in 50% or more of the corresponding corpus. However, the Spanish authors writing OSRAs in their L1 have the shortest list of nous distributed within the 50% threshold. The other three groups have similar numbers of nouns distributed within the 50% threshold.

After the initial observations are made, the three corpora of authors writing OSRAs in English are analyzed, looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. After analysis, the nouns presented in Table 58 below are deemed unlikely to function as NLID markers given their similar ranks and percentage in the corresponding corpus, despite differences in the number of occurrences between the groups.

N	Noun	Corpus	Occurrences in corpus	Rank	% in corpus
1	approach	EN-GB	77	38	0,04
		EN-PT^{EU}	77	109	0,05
		EN-ES^{EU}	52	223	0,03
2	association	EN-GB	142	52	0,08
		EN-PT^{EU}	170	27	0,10
		EN-ES^{EU}	138	65	0,07
3	change	EN-GB	220	23	0,13
		EN-PT^{EU}	111	39	0,07
		EN-ES^{EU}	208	28	0,11
4	characteristic	EN-GB	61	170	0,04
		EN-PT^{EU}	62	43	0,04
		EN-ES^{EU}	80	51	0,04
5	comparison	EN-GB	76	39	0,04
		EN-PT^{EU}	39	286	0,02
		EN-ES^{EU}	45	275	0,03
6	conclusion	EN-GB	44	45	0,03
		EN-PT^{EU}	68	42	0,04
		EN-ES^{EU}	60	54	0,03
7	condition	EN-GB	62	164	0,10
		EN-PT^{EU}	118	38	0,07
		EN-ES^{EU}	133	42	0,07
8	contrast	EN-GB	54	198	0,03
		EN-PT^{EU}	38	292	0,02
		EN-ES^{EU}	92	48	0,05
9	control	EN-GB	214	30	0,13
		EN-PT^{EU}	260	16	0,16
		EN-ES^{EU}	218	25	0,12
10	development	EN-GB	84	96	0,05
		EN-PT^{EU}	125	36	0,08
		EN-ES^{EU}	149	40	0,08

11	difference	EN-GB	313	13	0,18
		EN-PT^{EU}	208	20	0,13
		EN-ES^{EU}	293	16	0,16
12	discussion	EN-GB	69	41	0,04
		EN-PT^{EU}	71	41	0,04
		EN-ES^{EU}	67	53	0,04
13	finding	EN-GB	184	26	0,11
		EN-PT^{EU}	132	32	0,08
		EN-ES^{EU}	123	44	0,07
14	function	EN-GB	119	66	0,07
		EN-PT^{EU}	130	33	0,08
		EN-ES^{EU}	171	35	0,09
15	group	EN-GB	460	6	0,27
		EN-PT^{EU}	466	6	0,29
		EN-ES^{EU}	432	11	0,23
16	increase	EN-GB	205	24	0,12
		EN-PT^{EU}	171	26	0,10
		EN-ES^{EU}	262	18	0,14
17	limitation	EN-GB	68	43	0,04
		EN-PT^{EU}	54	45	0,03
		EN-ES^{EU}	55	214	0,03
18	number	EN-GB	245	21	0,14
		EN-PT^{EU}	162	29	0,10
		EN-ES^{EU}	210	27	0,11
19	part	EN-GB	52	44	0,03
		EN-PT^{EU}	28	439	0,02
		EN-ES^{EU}	46	273	0,02
20	process	EN-GB	46	253	0,03
		EN-PT^{EU}	51	210	0,03
		EN-ES^{EU}	112	45	0,06
21	range	EN-GB	91	37	0,05
		EN-PT^{EU}	62	163	0,04
		EN-ES^{EU}	42	314	0,02
22	reduction	EN-GB	76	120	0,04
		EN-PT^{EU}	78	108	0,05
		EN-ES^{EU}	94	47	0,05
23	research	EN-GB	73	40	0,04
		EN-PT^{EU}	30	407	0,02
		EN-ES^{EU}	44	291	0,02
24	risk	EN-GB	310	14	0,18
		EN-PT^{EU}	281	12	0,17
		EN-ES^{EU}	176	31	0,10
25	sample	EN-GB	140	33	0,08
		EN-PT^{EU}	173	24	0,11
		EN-ES^{EU}	173	33	0,09
26	size	EN-GB	44	266	0,03
		EN-PT^{EU}	46	241	0,03
		EN-ES^{EU}	82	50	0,04

27	table	EN-GB EN-PT^{EU} EN-ES^{EU}	254 289 211	18 11 26	0,15 0,18 0,11
28	time	EN-GB EN-PT^{EU} EN-ES^{EU}	229 172 180	22 25 30	0,13 0,11 0,10
29	treatment	EN-GB EN-PT^{EU} EN-ES^{EU}	246 215 330	20 19 13	0,14 0,13 0,18
30	type	EN-GB EN-PT^{EU} EN-ES^{EU}	163 130 139	28 34 41	0,10 0,08 0,08
31	use	EN-GB EN-PT^{EU} EN-ES^{EU}	122 123 79	34 37 52	0,07 0,08 0,04

Table 58 – Nouns found to be unlikely to function as NLID markers given their similar ranks and their percentage in the corresponding corpus

Another group of nouns is also ruled out after examining the OSRAs where they are used and concluding that the differences in frequency, ranks, and/or percentages in corresponding corpora are context-related. That is, these nouns are more frequently used in one corpus or another because of the topic being discussed in the OSRAs, or because they are terms related to the matter in the analysis. These nouns are presented below in Table 59.

N	Noun	Corpus	Occurrences in corpus	Rank	% in corpus
1	%	EN-GB EN-PT^{EU} EN-ES^{EU}	1493 1458 936	1 1 4	0,87 0,89 0,51
2	age	EN-GB EN-PT^{EU} EN-ES^{EU}	313 298 171	12 10 34	0,18 0,18 0,09
3	case	EN-GB EN-PT^{EU} EN-ES^{EU}	444 266 244	7 15 19	0,26 0,16 0,13
4	cell	EN-GB EN-PT^{EU} EN-ES^{EU}	271 653 1012	19 4 2	0,16 0,40 0,55
5	datum	EN-GB EN-PT^{EU} EN-ES^{EU}	465 201 220	5 21 23	0,27 0,12 0,12
6	disease	EN-GB EN-PT^{EU} EN-ES^{EU}	347 372 223	9 8 21	0,20 0,23 0,12

7	effect	EN-GB	325	11	0,19
		EN-PT^{EU}	310	9	0,19
		EN-ES^{EU}	548	8	0,30
8	expression	EN-GB	137	59	0,08
		EN-PT^{EU}	240	20	0,14
		EN-ES^{EU}	521	10	0,28
9	factor	EN-GB	156	30	0,09
		EN-PT^{EU}	229	18	0,14
		EN-ES^{EU}	220	24	0,12
10	figure	EN-GB	350	8	0,20
		EN-PT^{EU}	187	28	0,11
		EN-ES^{EU}	198	34	0,11
11	gene	EN-GB	379	8	0,22
		EN-PT^{EU}	248	19	0,15
		EN-ES^{EU}	364	12	0,20
12	health	EN-GB	154	31	0,09
		EN-PT^{EU}	41	271	0,03
		EN-ES^{EU}	89	103	0,05
13	mechanism	EN-GB	80	103	0,05
		EN-PT^{EU}	99	74	0,06
		EN-ES^{EU}	162	38	0,09
14	model	EN-GB	194	35	0,11
		EN-PT^{EU}	127	35	0,08
		EN-ES^{EU}	191	29	0,10
15	outcome	EN-GB	141	32	0,08
		EN-PT^{EU}	83	101	0,05
		EN-ES^{EU}	70	150	0,04
16	P	EN-GB	175	39	0,10
		EN-PT^{EU}	278	13	0,17
		EN-ES^{EU}	554	6	0,30
17	patient	EN-GB	998	2	0,58
		EN-PT^{EU}	1311	1	0,71
		EN-ES^{EU}	1416	2	0,87
18	population	EN-GB	270	17	0,16
		EN-PT^{EU}	195	22	0,12
		EN-ES^{EU}	169	36	0,09
19	proportion	EN-GB	114	36	0,07
		EN-PT^{EU}	44	253	0,03
		EN-ES^{EU}	38	355	0,02
20	protein	EN-GB	167	43	0,10
		EN-PT^{EU}	241	21	0,15
		EN-ES^{EU}	289	17	0,16
21	rate	EN-GB	302	15	0,18
		EN-PT^{EU}	185	23	0,11
		EN-ES^{EU}	148	56	0,08
22	system	EN-GB	55	196	0,03
		EN-PT^{EU}	60	172	0,04
		EN-ES^{EU}	160	39	0,09

23	test	EN-GB	118	35	0,07
		EN-PT^{EU}	19	647	0,01
		EN-ES^{EU}	142	60	0,08
24	tissue	EN-GB	51	220	0,03
		EN-PT^{EU}	133	31	0,08
		EN-ES^{EU}	222	22	0,12
25	UK	EN-GB	157	29	0,09
		EN-PT^{EU}	4	2027	0,00
		EN-ES^{EU}	2	3382	0,00
26	year	EN-GB	510	4	0,30
		EN-PT^{EU}	260	17	0,16
		EN-ES^{EU}	168	37	0,09

Table 59 – Nouns whose differences in frequency, ranks and or percentages in corpora are context-related and therefore, are not considered to be likely to function as NLID markers in OSRAs written in English by the L1 Portuguese/Spanish authors

After the preliminary analysis is completed, two nouns are found in the EN-GB corpus more frequently than in the EN-PT^{EU} and EN-ES^{EU} corpora and therefore are analyzed to verify their potential to mark possible strategies of avoidance of use by the non-L1 authors. These nouns are shown in Table 60.

N	Noun	Corpus	Occurrences in corpus	Rank	% in corpus
1	[analysis]	EN-GB	346	10	0,20
		EN-PT^{EU}	267	14	0,16
		EN-ES^{EU}	174	32	0,09
2	[evidence]	EN-GB	163	27	0,10
		EN-PT^{EU}	60	169	0,04
		EN-ES^{EU}	83	49	0,05

Table 60 – Nouns that could mark strategies of avoidance in non-L1 English authors

The frequencies of the nouns "analysis" and "evidence" are compared to examine significant differences between the groups. The fourth L1 effect of the unified framework (Jarvis 2010, 2000) is tested for both the EN-PT^{EU} and the EN-ES^{EU} OSRAs, stated as follows:

Effect of L1 Influence	L1 influence EN-PT^{EU} question	L1 influence EN-ES^{EU} question
IV) Intralingual contrast	Are the frequencies of the noun "analysis/evidence" in the EN-PT^{EU}/ EN-ES^{EU} and the EN-GB corpora statistically significantly different?	

Since the data is not normally distributed and has some outliers, the Mann-Whitney test is used to assess for a mean difference between the groups. The level of significance used is *p* < .05. Two tests are performed. The numbers of occurrences are normalized by 100. Table 61 shows the results obtained.

	IV - Intralingual contrast (Mann-Whitney Test)	Effect of L1 influence IV found for EN-PT^{EU}?	IV - Intralingual contrast (Mann-Whitney Test)	Effect of L1 influence IV found for EN-ES^{EU}?
	corpora examined			
Noun	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. EN-GB	
	***p* reference value < .05**			
[analysis]	*Z* =-2.598 *p* = .009 *M* rank EN-GB = 63.95 *M* rank EN-PT^{EU} = 48.19	*yes*	*Z* = -3.319 *p* = .001 *M* rank EN-GB= 61.03 *M* rank EN-ES^{EU}= 41.66	*yes*
[evidence]	*Z* =-1.755 *p* = .079 *M* rank EN-GB = 43.83 *M* rank EN-PT^{EU} = 34.64	*no*	*Z* = -1.936 *p* = .053 *M* rank EN-GB= 50.01 *M* rank EN-ES^{EU}= 39.60	*no*

Table 61 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups concerning the noun [analysis] deemed as likely to mark strategies of avoidance

As can be seen in Table 61, the Mann Whitney tests indicate that there are no statistically significant differences in the ranked frequencies of the noun [evidence] and its equivalents [evidência] in Portuguese and "evidencia" in Spanish between the L1 and the non-L1 English groups writing in English. On the other hand, the noun [analysis] is significantly more frequently used by the L1 (i.e., EN-GB) than by the non-L1 authors (EN-PT^{EU} and EN-ES^{EU}) writing OSRAs in English. Based on the significance of the results, the noun [analysis] is further examined for both groups (EN-PT^{EU} and EN-ES^{EU}). The examinations are based on the concordances of the parsed files. The concordance are obtained with WordSmith 7.0 (Scott 2018b), from which the syntactic tags containing the nouns (N) are extracted. As can be seen below in Table 62, most of the syntactic structures associated with the noun "analysis" in the EN-GB corpus contribute to the significant differences in the number of occurrences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups.

N	Syntactic structure of [analysis]	Function	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	N S NOM @P	*...from analysis of the …*	119	101	60

3	N S NOM @SUBJ>	*…panel analysis will evolve…*	97	105	67
4	N S NOM @	*…we ran a conservative sensitivity analysis considering that…*	32	33	13
5	N P NOM @P	*…were identified in the […] population from statistical analyses of data*	34	5	15
6	N P NOM @SUBJ>	*The above analyses showed that…*	26	7	8
7	N S NOM @N	*This analysis uses prescribed medications,...*	11	5	3
8	N S NOM @>N	*by applying statistical analysis techniques*	11	4	3
9	N P NOM @	*We carried out two separate analyses:*	5	2	2
10	N P NOM @NPHR	*Supplementary analyses: awareness*	3	1	
11	N S NOM @NPHR	*…so we limited subsequent analysis to 25 patients…*	2	1	1
12	N P NOM @N	*…(analyses performed within each cohort and study-specific estimates pooled in a meta-analysis)…*	1	1	
13	N S NOM @>A	*Our analysis restricted to current smokers did not detect*	1	1	
14	N S NOM @PRED>	*Analysis of the Diagnostic Validity of the Point-of-Care Test*		1	1
15	N P NOM @>N	*… each corresponding to ≥30 partnerships upon which analyses hereon are based.*	1		
16	N P NOM @P< @	*Tables 4 and 5 exhibit the results of bivariate and multivariable linear regression analyses of the predictors*			1
17	N S NOM @ACC>	*Analysis of this data showed that mRNA abundance is dependent on*	1		
18	N S NOM @APP	*…DNA and exon sequence analysis of collagen type Ia1 genes did not yield any clues…*	1		

Table 62 – Syntactic tags of the noun [analysis] evincing the significant differences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups

Since the noun "analysis" has a general meaning but may also have field-related meanings or be part of terms, an examination of the collocations is carried out. After examining the most frequent collocates to the left and to the right of the noun "analysis", the conclusion is reached that the significant differences between the groups respond to collocational or terminological uses. As shown in Table 63 below, the terminological uses of the noun "analysis" are more frequently found in words situated to the left of the noun than to the right.

N	Collocations to the left of the word "analysis"	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	meta-analysis	18	10	9
2	this analysis	20	6	3
3	multivariate analysis	4	8	8
4	regression analysis	2	9	8
5	univariate analysis	1	12	4
6	statistical analysis	5	10	1
7	sensitivity analyses	10	2	3
8	blot analysis	3	3	6
9	sequencing analysis		9	1
10	further analysis	6	3	
11	multivariable analysis	7	2	
12	[19][N]-based analysis	2	2	
13	from analysis	2	1	1
14	immunoblot analysis	2		2
15	meta analysis		3	

Table 63 – Collocations to the left of the noun "analysis" by increasing order

Collocational uses appeared more frequently to the right of "analysis", as shown in Table 64.

N	Collocations to the right of the word "analysis"	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	analysis showed	6	11	13
2	analysis revealed	4	12	11
3	analysis using	4	4	5
4	analyses were	8		1
5	analysis that	3	2	3
6	analysis demonstrated	1	6	

[19] E.g. patient-based, panel-pased, distance-based.

7	analysis with	1	1	5
8	analysis from		1	5
9	analyses demonstrated		2	

Table 64 – Collocations to the right of the noun [analysis] by increasing order

Despite the significant differences between the groups in relation to the frequency of the noun “analysis”, it cannot be affirmed that the non-L1 authors avoid using the word, especially because the collocations to the right of “analysis” are more frequent in non-L1 than in L1 English authors. The significant differences are most like to be the consequence of using certain scientific methods that have to be reported using specific terms.

After examining the noun “analysis”, the remaining nouns are selected for examination based on their possible non-terminological meaning and after consideration of their (a) occurrences, (b) ranks, and (c) percentages of occurrences being higher in both or one of the non-L1 groups (EN-PT^{EU} and EN-ES^{EU}) than in the L1 (EN-GB) and therefore, have the potential to function as NLID markers of language transfer if similar occurrences, ranks, and percentages are verified in the L1 PT-EU and ES-EU corpora. Since these nouns are chosen as possible markers of language transfer, their equivalents in the L1 Portuguese/Spanish corpora are extracted, and their frequencies are compared. Table 65 shows the nouns that after comparison are not further analyzed either because the frequencies of their equivalents in the corresponding L1 corpora do not justify the high frequencies found in the non-L1 English corpora (one or both), or the ES-EU corpus does not have occurrences of the given equivalent, and therefore, the L1 effect comparison cannot be performed.

N	Noun	Corpus	Occurrences in corpus	Rank	% in corpus
1	atividade	PT-EU	84	80	0,06
	actividad	ES-EU	137	22	0,08
	activity	EN-GB	167	42	0,10
		EN-PT^{EU}	150	30	0,09
		EN-ES^{EU}	303	15	0,16
3	nível	PT-EU	228	16	0,16
	nivel	ES-EU	219	11	0,13
	level	EN-GB	203	25	0,12
		EN-PT^{EU}	566	5	0,35
		EN-ES^{EU}	812	5	0,44

5	resposta	PT-EU	52	169	0,04
	respuesta	ES-EU	0	NA	NA
	response	EN-GB	197	34	0,12
		EN-PT^{EU}	77	111	0,05
		EN-ES^{EU}	310	14	0,17
6	resultado	PT-EU	399	8	0,27
	resultado	ES-EU	0	NA	NA
	result	EN-GB	279	16	0,16
		EN-PT^{EU}	399	7	0,24
		EN-ES^{EU}	552	7	0,30
7	papel / função	PT-EU	81	172	0,06
	función[20]	ES-EU	44	137	0,03
	role	EN-GB	74	125	0,04
		EN-PT^{EU}	167	28	0,10
		EN-ES^{EU}	236	20	0,13
9	valor	PT-EU	347	9	0,24
	valor	ES-EU	0	NA	NA
	value	EN-GB	56	190	0,03
		EN-PT^{EU}	144	43	0,09
		EN-ES^{EU}	131	43	0,07

Table 65 – Nouns that upon analysis are disregarded as NLID markers of language transfer

Therefore, three nouns are examined in Portuguese/Spanish as found in the corresponding L1 corpora.

N	Noun	Corpus	Occurrences in corpus	Rank	% in corpus
1	facto	PT-EU	95	72	0,07
	hecho	ES-EU	37	172	0,02
	fact	EN-GB	14	760	0,01
		EN-PT^{EU}	41	270	0,03
		EN-ES^{EU}	59	55	0,03
2	presença / existência[21]	PT-EU	133	78	0,09
	existencia[22]	ES-EU	32	192	0,02
	presence	EN-GB	51	219	0,03
		EN-PT^{EU}	110	40	0,07
		EN-ES^{EU}	104	46	0,06
3	estudo	PT-EU	931	2	0,65
	estudio	ES-EU	873	2	0,54
	study	EN-GB	728	3	0,43
		EN-PT^{EU}	820	3	0,50
		EN-ES^{EU}	951	3	0,52

[20] No occurrences of the noun “papel” (an equivalent of “role”) were found in the ES-EU corpus

[21] Both possible translation into Portuguese are considered.

[22] No direct translation, i.e., [presencia] was found, therefore a common synonym is considered.

Table 66 – Nouns that could function as NLID markers of language transfer

The frequencies of these nouns, i.e., “fact”, “presence”, “study”, and equivalents, are compared to examine significant differences between the groups. For all nouns, the following questions are asked for both the EN-PT^{EU} and the EN-ES^{EU} corpora.

Effect of L1 Influence	**L1 influence** **EN-PT^{EU} questions**	**L1 influence** **EN-ES^{EU} questions**
I) Intragroup homogeneity	Are the frequencies of the nouns fact/presence/study, in the EN-PT^{EU} / EN-ES^{EU} OSRAs uniformly distributed?	
II) Intergroup heterogeneity	Are the frequencies of the nouns fact/presence/study, in the EN-PT^{EU} and EN-ES^{EU} OSRAs statistically significantly different?	
III) Cross-language congruity	Are the frequencies of the nouns fact/presence/study in the EN-PT^{EU} and equivalents in the PT-EU OSRAs statistically similar?	Are the frequencies of the nouns fact/presence/study in the EN-ES^{EU} and equivalents in the ES-EU OSRAs statistically similar?
IV) Intralingual contrast	Are the frequencies of the nouns fact/presence/study in the EN-PT^{EU} and EN-GB OSRAs statistically significantly different?	Are the frequencies of the nouns fact/presence/study in the EN-ES^{EU} and EN-GB OSRAs statistically significantly different?

The Mann-Whitney test is used to assess for a mean difference between the groups given that the data is not normally distributed and has some outliers. Questions in relation to the effects I and II are answered together. The level of significance used is $p < .05$ for questions in relation to effects II and IV. Because questions in relation to effects I and III look for uniformity and congruity, respectively, a result of $p > .05$ is associated with an L1 effect. Table 67 below shows the results and mean ranks of all comparisons.

Noun	L1 Influence Effects								
	I) Intra-L1 homogeneity (Levene's test)	II) Inter-L1 heterogeneity (Mann-Whitney Test)	EN-PT^{EU} and EN-ES^{EU} similar in variance but different in means?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-PT^{EU}?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-ES^{EU}?
	Corpora Examined			Corpora Examined			Corpora Examined		
	EN-PT^{EU} vs. EN-ES^{EU}	EN-PT^{EU} vs. EN-ES^{EU}		EN-PT^{EU} vs. PT-EU	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. ES-EU	EN-ES^{EU} vs. EN-GB	
	Reference *p* values			Reference *p* values			Reference *p* values		
	p > .05 **AND**	*p* < .05?		*p* > .05?	*p* < .05?		*p* > .05?	*p* < .05?	
facto/hecho/fact	*F* = 1.092 *p* = .300	*Z* = -.392 *p* = .695 Mean ranks: EN-PT^{EU}= 30.98 EN-ES^{EU}= 29.38	*no*	*Z* =-1.513 *p* = .130 Mean ranks: EN-PT^{EU}= 28.04 PT-EU=35.00	*Z* =-1.250 *p* = .291 Mean ranks: EN-PT^{EU}=18.78 EN-GB=14.82	*no*	*Z* = -1.518 *p* = .129 Mean ranks: EN-ES^{EU}=33.94 ES-EU= 28.12	*Z* =-1.089 *p* = .351 Mean ranks: EN-ES^{EU}=25.04 EN-GB=20.59	*no*
presença/existência existencia presence	*F* = .220 *p* = .641	*Z* = -.735 *p* = .463 Mean ranks: EN-PT^{EU}=35.20 EN-ES^{EU}=31.80	*no*	*Z* = -1.246 *p* = .213 Mean ranks: EN-PT^{EU}=44.27 PT-EU=37.85	*Z* =-2.252 *p* = .024 Mean ranks: EN-PT^{EU}=32.47 EN-GB=22.80	*yes*	*Z* =-1.197 *p* = .231 Mean ranks: EN-ES^{EU}=26.64 ES-EU=21.63	*Z* =-1.557 *p* = .119 Mean ranks: EN-ES^{EU}=31.21 EN-GB=24.61	*no*
estudo/estudio/study	*F* = 1.769 *p* = .186	*Z* = .914 *p* = .361 Mean ranks: EN-PT^{EU}=62.48 EN-ES^{EU}=68.52	*no*	*Z* = -1.181 *p* = .238 Mean ranks: EN-PT^{EU}= 61.15 PT-EU= 68.91	*Z* = -1.250 *p* = .211 Mean ranks: EN-PT^{EU}=69.08 EN-GB=60.86	*no*	*Z* =-.573 *p* = .566 Mean ranks: EN-ES^{EU}=67.39 ES-EU=63.61	*Z* = -2.063 *p* = .039 Mean ranks: EN-ES^{EU}= 71.73 EN-GB= 58.16	*yes*

Table 67 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups in relation to the frequency of the nouns deemed likely to function as NLID markers

As shown in Table 67, the results of the Mann-Whitney tests indicate statistically significant differences in the ranked frequencies of the noun [presence] between the L1 (EN-GB) and the non-L1 (EN-PT^{EU}) English authors writing in English; and in the ranked frequencies of the noun [study] between the L1 (EN-GB) and the non-L1 (EN-ES^{EU}) English authors writing in English. Additionally, no differences are verified between the Portuguese authors writing in their L1 and the Portuguese authors writing in English in relation to the frequencies of the noun [presence] and its equivalents in Portuguese [presença] and [existência]. Similarly, no differences are found between the Spanish authors writing in their L1 and the Spanish authors writing in English in relation to the frequencies of the noun [study] and its equivalent in Spanish [estudio].

Since the effects III and IV, i.e., cross-language congruity and intralingual contrast, are found for both groups, the nouns [presence] and [study] could function as NLID markers for the Portuguese and Spanish authors writing OSRAs in English, respectively.

After analysis, it is verified that all syntactic functions of the noun [presence]/[existence] in English and their equivalents [presença]/[existência] found in the EN-PT^{EU,} and the PT-EU corpora contribute to the significant differences between the Portuguese and British authors in relation to these nouns since their frequencies are higher in the EN-PT^{EU} and the PT-EU corpora than in the EN-GB corpus (Table 68 below).

N	Tags	Corpus	[existence] [existência]	[presence] [presença]
1	N F S @P<	PT-EU	16	35
2	N F S @SUBJ>		8	22
3	N F S @<ACC		14	12
4	N S NOM @P	EN-GB	-	26
5	N S NOM @		-	13
6	N S NOM @SUBJ>		1	12
7	N S NOM @P	EN-PT^{EU}	4	65
8	N S NOM @SUBJ>		1	22
9	N S NOM @		3	18

Table 68 – Syntactic functions of the noun “presence”/“existence” in English and equivalents “presença”/“existencia” in Portuguese

The analysis of the collocations of the nouns “presence” and “existence” using WordSmith (Scott 2018b) shows that the expressions in Table 69 are the most frequently found containing those nouns. Therefore, both “presence” and “existence” may function as markers of L1 transfer in Portuguese authors writing OSRAs in English, especially in combination with the preposition “of”.

Trigram with “presence”/“existence”	PT-EU	EN-GB	EN-PT^EU
the presence of the existence of		36 1	92 8
a presença de a existência de	46 28		

Table 69 – Most frequent word combinations with [presence]/[existence] in the three English corpora

Also, graphically, it is possible to appreciate the distance between the EN-GB and the EN-PT^{EU} OSRAs and the proximity between EN-PT^{EU} and PT-EU OSRAS concerning the noun [presence] and its equivalents in [presença]/[existência] in Portuguese (Figure 14).

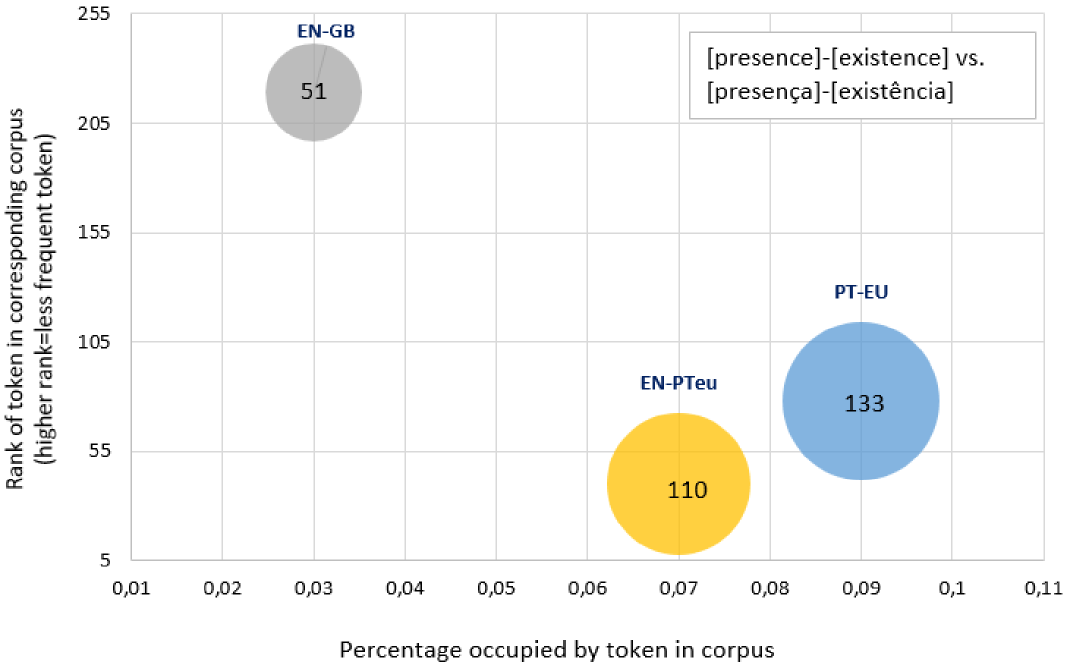


Figure 14 – Distribution of the nouns “presence”/“existence” and “presença”/“existência”

Finally, the analysis of the syntactic structures of the noun "study" in the EN-GB and the EN-ES^{EU} corpora and its equivalent in Spanish "estudio" in the ES-EU corpus shows that the most frequent structures behind the significant differences between the EN-GB and the EN-ES^{EU} corpora concern the subject of a sentence and the argument of prepositions such as "in"/ "of"/ "with", as shown in Table 70 below.

N	Syntactic Structures of [study]	ES-EU	EN-GB	EN-ES^{EU}
1	N M S @P< / N M P @P<	551		
2	N M S @SUBJ> / N M P @SUBJ>	184		
5	N S NOM @P / N P NOM @P		360	513
6	N P NOM @SUBJ> / N S NOM @SUBJ>		248	304

Table 70 – Most frequent syntactic structures of the noun "study" in the EN-GB and the EN-ES^{EU} corpora and its equivalent in Spanish "estudio" in the ES-EU corpus that show the significant differences between the groups

These syntactic functions correspond mostly to the expressions listed below in Table 71, also extracted with WordSmith from the EN-GB, the EN-ES^{EU}, and the ES-EU corpora.

Bigrams with [estudio]/[study]	ES-EU	EN-GB	EN-ES^{EU}
este estudio	88		
nuestro estudio	84		
presente estudio	48		
estudio(s) de	112		
estudio(s) con	9		
this study		117	120
present study		32	84
current study		10	18
recent study		6	15
study/studies of		44	39
study/studies in		22	35
study/studies with		12	26
Total	**341**	**243**	**337**

Table 71 – Most frequent expressions with the noun [study] in the EN-GB and the EN-ES^{EU} corpora and [estudio] in the ES-EU corpus behind the most frequent syntactic structures that show the significant differences between the groups

Despite the statistical data indicating a possible L1 effect in relation to the frequency and use of the noun "study" in the Spanish authors, the word is very frequent in scientific writing. Therefore, its usefulness to mark L1 in OSRAs written in English by these authors would be better served if “study” is combined with a proposition like “of” and verified for the syntactic functions mentioned above.

4.5. Variables with two effects of L1 influence

4.5.1. V9: number of indefinite articles

After examining the samples of V9, the following descriptive statistics are obtained:

V9: frequency of indefinite articles (all values in words/1000 tokens)							
CORPORA	Mean	Std. Error	95% Confidence Interval for Mean		Median	Variance	Std. Deviation
			Lower Bound	Upper Bound			
PT-EU	16,84	0,65	15,55	18,14	16,77	27,35	5,23
ES-EU	19,60	0,71	18,18	21,02	19,13	32,72	5,72
EN-GB	19,88	0,73	18,43	21,34	20,04	34,69	5,89
EN-PT^{EU}	17,86	0,70	16,47	19,25	17,21	31,58	5,62
EN-ES^{EU}	17,55	0,70	16,15	18,95	17,02	32,04	5,66

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V9 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene’s test, *F* = .013, *p* = .908);

II. There are no statistically significant differences in the frequency of indefinite articles in the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = .314, *p* = .754.
Both groups have similar means of indefinite articles per thousand words (*MD* = .310; *SED* = .989; 95% CI = 2.27 to -1.65 indefinite articles per thousand words);

III. There are no statistically significant differences in the frequency of indefinite articles between the EN-PT^{EU} and PT-EU groups, *t*(128) = 1.068, *p* = .288.
Both groups have very similar means of indefinite articles per thousand words (*MD* = 1.02; *SED* = .952; 95% CI = 2.90 to -.867 indefinite articles per thousand words);

There are statistically significant differences in the frequency of indefinite articles between the EN-ES^{EU} and ES-EU groups, *t*(128) = 2.049, *p* = .042.

The ES-EU OSRAs have significantly more indefinite articles per thousand words than the EN-ES^{EU} OSRAs (*MD* = 2.05; *SED* = .998, 95% CI = 4.02 to .070 indefinite articles per thousand words);

IV. There are statistically significant differences in the frequency of indefinite articles between the EN-PT^{EU} and EN-GB groups, *t*(128) = 2.005, *p* = .047.
The EN-GB group uses significantly more indefinite articles per thousand words than the EN-PT^{EU} group (*MD* = 2.02; *SED* = 1.00; 95% CI = 4.02 to .026 indefinite articles per thousand words).

There are statistically significant differences in the frequency of indefinite articles between the EN-ES^{EU} and EN-GB groups, *t*(128) = 2.305, *p* = .023.
The EN-GB group uses significantly more indefinite articles per thousand words than the EN-ES^{EU} group (*MD* = 2.33; *SED* = 1.01; 95% CI = 4.34 to .330 indefinite articles per thousand words).

The following table summarizes the effects of L1 influence found in each group of OSRAs:

Effect of L1 Influence	**EN-PT^{EU}**	**EN-ES^{EU}**
I. Intra-L1 homogeneity	--	
II. Inter-L1 heterogeneity	--	
III. Cross-language congruity	✓	--
IV. Intralingual contrast	✓	✓

The frequency of indefinite articles in OSRAs written in English by the Spanish authors cannot be said to be influenced by the frequency of use these authors make of indefinite articles when writing OSRAs in their L1. The Spanish authors use significantly more indefinite articles per thousand words when they write OSRAs in their L1 than when they write OSRAs in English. That is, the Spanish authors significantly diminish the frequency of indefinite articles when writing OSRAs in English, and this decrease makes them be significantly different from the L1 English authors writing in English.

However, as can be seen in Figure 15 below, in the case of the Portuguese authors writing OSRAs in English, it can be argued that the frequency with which they use indefinite articles may be influenced by the frequency with which they use indefinite articles when writing OSRAs in their L1. The Portuguese authors writing OSRAs in Portuguese use indefinite articles as frequently as the Portuguese authors writing OSRAs in English but significantly less frequently than the L1 English authors writing in English.

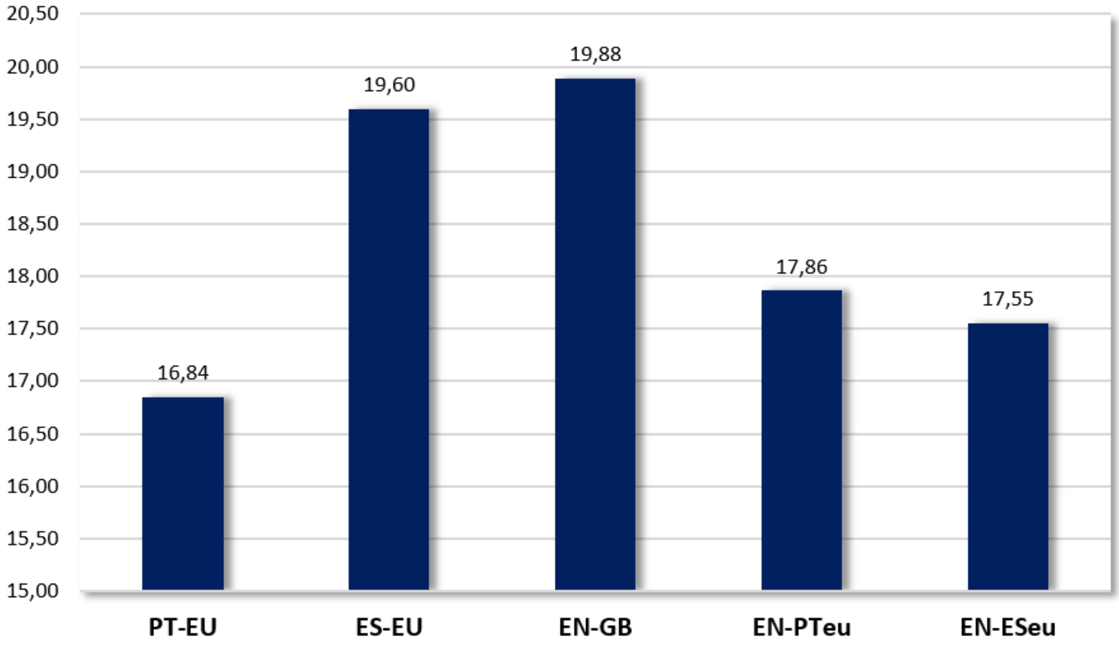


Figure 15 – Mean values of the indefinite articles in the five corpora of the CoRA

Since variation within the grammatical category "indefinite article" is very restricted, i.e., a / an, no further linguistics analysis is performed.

4.5.2. V13: number of demonstrative pronouns

After examining the samples of V13, the following descriptive statistics are obtained:

V13: frequency of demonstrative pronouns (all values in words/1000 tokens)

CORPORA	Mean	Std. Error	95% Confidence Interval for Mean		Median	Variance	Std. Deviation
			Lower Bound	Upper Bound			
PT-EU	12,15	0,56	11,03	13,28	11,60	20,61	4,54
ES-EU	9,08	0,44	8,21	9,96	8,70	12,39	3,52
EN-GB	12,84	0,54	11,76	13,93	11,80	19,10	4,37
EN-PT^{EU}	9,91	0,45	9,01	10,80	9,70	13,10	3,62
EN-ES^{EU}	11,30	0,41	10,49	12,12	11,20	10,76	3,28

The lowest mean standardized value of demonstrative pronouns is in the ES-EU corpus, and the highest is in the EN-GB corpus.

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V13 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test**,** *F* = .766, *p* = .383);

II. Still, there are statistically significant differences in the frequency of demonstrative pronouns between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = 2.303, *p* = .023.
The EN-ES^{EU} group uses significantly more demonstrative pronouns per thousand words than the EN-PT^{EU} groups (*MD* = 1.40, *SED* = .606; 95% CI = 2.59 to .197 demonstrative pronouns per thousand words). However, the effect size of such difference is small (Cohen's *d* = .407)

III. There are statistically significant differences in the frequency of demonstrative pronouns between the EN-PT^{EU} and PT-EU groups, *t*(121.909) = 3.122, *p* = .002.
The PT-EU group uses significantly more demonstrative pronouns per thousand words than the EN-PT^{EU} group (*MD* = 2.25; *SED* = .720; 95% CI = 3.67 to .822 demonstrative pronouns per thousand words). The effect size of such difference is medium (Cohen's *d = .*565)
There are statistically significant differences in the frequency of demonstrative pronouns between the EN-ES^{EU} and ES-EU groups, *t*(128) = 3.716, *p* = .001.
The EN-ES^{EU} group uses demonstrative pronouns more frequently than the ES-EU group (*MD* = 2.22; *SED* = .597, 95% CI = 3.40 to 1.04 demonstrative pronouns per thousand words). Additionally, the effect size is medium (Cohen's *d* = 0.657).

IV. There are statistically significant differences in the frequency of demonstrative pronouns between the EN-PT^{EU} and EN-GB groups, *t*(128) = 4.170, *p* = .001. The EN-GB group uses significantly more demonstrative pronouns per thousand words than the EN-PT^{EU} group (*MD* = 2.94; *SED* = .704; 95% CI = 4.33 to 1.54 demonstrative pronouns per thousand words). The effect size calculated is medium (Cohen's *d* = 0.737)

There also are statistically significant differences in the frequency of demonstrative pronouns between the EN-ES^{EU} and EN-GB groups, *t*(118.740) = 2.270, *p* = .025. The EN-GB group uses significantly more demonstrative pronouns per thousand words than the EN-ES^{EU} group (*MD* = 1.54; *SED* = .678; 95% CI = 2.88 to .197 demonstrative pronouns per thousand words). The effect size calculated is small (Cohen's *d* = 0.417).

The following table summarizes the effects of L1 influence found:

Effect of L1 Influence	**EN-PT^{EU}**	**EN-ES^{EU}**
I. Intra-L1 homogeneity	✓	
II. Inter-L1 heterogeneity	✓	
III. Cross-language congruity	--	--
IV. Intralingual contrast	✓	✓

The results of the statistical examination indicate possible effects of L1 influence on demonstrative pronouns by both the PT-EU and the ES-EU L1 users when writing OSRAs in English. However, the results obtained are not related to higher frequencies of demonstrative pronouns found in OSRAs written in English by the L1 Portuguese/Spanish. Although the EN-PT^{EU} and the EN-ES^{EU} groups differ significantly between each other and from the L1 English authors (EN-GB) in relation to the frequency of demonstrative pronouns – which allows for intergroup heterogeneity and intralingual contrast to be argued– they differ because the L1 English authors (EN-GB) use demonstrative pronouns significantly more frequently than the non-L1 English corpora. Table 72 below shows the demonstrative pronouns within each corpus and their corresponding frequencies. Overall, the EN-PT^{EU} authors show a significant decrease in the number of demonstratives they use compared to the L1 PT-EU writing in their L1, and also compared to the L1 EN-GB authors writing in English. The EN-ES^{EU} authors, however, show a significant increase in the frequency of this variable when compared to the L1 ES-EU authors, coming closer to the L1 EN-GB authors, though not enough as to not have a significant difference between them.

demonstrative pronoun	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}
a / as	79 / 26				
aquela / aquelas	5 / 5				
aquele / aqueles	1 / 35				
aquilo	1				
essa / essas	40 / 11				
esse / esses	46 / 20				
esta / estas	348 / 97				
este / estes	532 / 292				
isso	31				
isto	36				
o / os	75 / 71				
tais	3				
aquel		2			
aquella / aquellas / aquellos		2 / 13 / 44			
el		5			
esa / esas		12 / 8			
ese		25			
eso / esos		7 / 16			
esta / estas		11 / 1142			
este		16			
esto / estos		67 / 13			
la / las		6 / 2			
lo / los		35 / 12			
tal / tales		11 / 1			
that / those			163 / 352	139 / 125	187 / 186
this / these			1065 / 539	813 / 476	1127 / 589
such			89	38	38
Total	**1754**	**1450**	**2208**	**1591**	**2127**

Table 72 – Demonstrative pronouns per corpus in the CoRA

In all the corpora, proximal demonstrative pronouns, i.e., those expressing proximity (E.g., esta, estas, this, these), are more frequent than distal demonstrative pronouns, i.e., those expressing distance (E.g., aquella, essas, that, those) (Stirling and Huddleston 2002: 1504) as shown in Table 73 below.

demonstrative forms	**PT-EU**	**ES-EU**	**EN-GB**	**EN-PT^{EU}**	**EN-ES^{EU}**
distal	195	129	515	264	373
proximal	1559	1321	1693	1327	1754
Total	**1754**	**1450**	**2208**	**1591**	**2127**

Table 73 – Distribution of demonstrative forms in the CoRA

As can be seen in Table 73, the EN-GB authors resort to distal demonstrative pronouns more frequently than all the other authors, while proximal demonstrative are slightly more frequent in the EN-ES^{EU} corpus, followed by the EN-GB corpus.

To understand if NLID markers can be found beyond the total frequencies, the demonstrative pronouns (i.e., <dem>, as tagged by VISL) in the three English corpora are analyzed in terms of distribution, looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. The analysis considers not only proximal and distal demonstrative pronouns but also their singular and plural variations separately. Table 74 below, shows the most frequent demonstrative pronouns in the CoRA.

Rank	**PT-EU**	**ES-EU**	**EN-GB**	**EN-PT^{EU}**	**EN-ES^{EU}**	**Threshold**
1	este n=532	estas n=1142	this n=1065	this n=813	this n=1127	**100%**
2	esta n=348	esto n=67	these n=539	these n=476	these n=589	**70-95%**
3	estes n=292		those n=352	that n=139	that n=187	
4	estas n=97		that n=163	those n=125	those n=186	**60-70%**
5	o n=75		such n=89			**50%**

Table 74 – Most frequent demonstrative pronouns in the CoRA (present in at least 50% of the OSRAs) ranked by the number of occurrences

After the analysis of the most frequent demonstrative pronouns in the English corpora of the CoRA, two pronouns are considered unlikely to function as NLID markers in OSRAs written in English by the Portuguese/Spanish authors since the percentage they represent within each of their corresponding corpus and the rank are similar among the groups, despite the numbers of occurrences being slightly higher in either the EN-GB or the EN-ES^{EU} corpora. Table 75 shows these demonstrative pronouns.

<dem>	Corpus	Occurrences	Rank	%
these	EN-GB	539	2	0,31
	EN-PT^{EU}	476	2	0,29
	EN-ES^{EU}	589	2	0,32
that	EN-GB	160	3	0,09
	EN-PT^{EU}	136	3	0,08
	EN-ES^{EU}	187	3	0,10

Table 75 – Demonstrative pronouns deemed unlikely to function as NLID markers given their similar ranks, occurrences, and percentage in the corresponding corpus

Therefore, only three demonstrative pronouns are further analyzed. However, since their frequency is still higher in the L1 than in the non-L1 English OSRAs, the examination seeks to understand if there are strategies of avoidance of use by the non-L1 English authors who are L1 Portuguese/Spanish users. These demonstrative pronouns are shown below in Table 76.

<dem>	Corpus	Occurrences	Rank	%
this	EN-GB	1065	1	0.62
	EN-PT^{EU}	813	1	0.50
	EN-ES^{EU}	1127	1	0.61
those	EN-GB	355	2	0.21
	EN-PT^{EU}	128	2	0.08
	EN-ES^{EU}	186	2	0.10
such	EN-GB	89	5	0.05
	EN-PT^{EU}	38	33	0.02
	EN-ES^{EU}	38	37	0.02

Table 76 – The demonstrative pronouns “this”, “those”, and “such” assessed for avoidance strategies by EN-PT^{EU}/EN-ES^{EU} authors

The frequencies of the demonstrative pronouns “this”, “those”, “such” are compared to examine if there are significant differences between the groups. The fourth L1 effect of the unified framework (*Jarvis 2010*, *2000*) is tested for both the EN-PT^{EU} and the EN-ES^{EU} OSRAs, stated as follows:

Effect of L1 Influence	**L1 influence EN-PT^{EU} question**	**L1 influence EN-ES^{EU} question**
IV) Intralingual contrast	Are the frequencies of the demonstrative pronoun "this"/"those"/"such" in the EN-PT^{EU}/ EN-ES^{EU} and the EN-GB corpora statistically significantly different?	

Since the data is not normally distributed and has outliers, the Mann-Whitney test is used to assess for a mean difference between the groups. The level of significance used is *p* < .05. Two tests are performed. The numbers of occurrences are normalized by 100. Table 77 shows the results obtained.

	IV - Intralingual contrast (Mann-Whitney Test)	Effect of L1 influence IV found for the EN-PT^{EU}?	IV - Intralingual contrast (Mann-Whitney Test)	Effect of L1 influence IV found for the EN-ES^{EU}?
	corpora examined			
<dem>	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. EN-GB	
	***p* reference value < .05**			
this	*Z* = -2.979 *p* = .003 *M* rank EN-GB = 75.33 *M* rank EN-PT^{EU} = 55.67	*yes*	*Z* = -.198 *p* = .843 *M* rank EN-GB= 64.85 *M* rank EN-ES^{EU}= 66.11	*no*
those	*Z* = -4.422 *p* = .001 *M* rank EN-GB = 64.89 *M* rank EN-PT^{EU} = 38.89	*yes*	*Z* = -3.042 *p* = .002 *M* rank EN-GB= 66.81 *M* rank EN-ES^{EU}= 48.19	*yes*
such	*Z* = -1.152 *p* = .249 *M* rank EN-GB = 33.61 *M* rank EN-PT^{EU} = 28.26	*no*	*Z* = -1.009 *p* = .313 *M* rank EN-GB= 35.61 *M* rank EN-ES^{EU}= 30.91	*no*

Table 77 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups concerning the demonstrative pronoun "those" deemed as likely to mark strategies of avoidance

As can be seen in Table 77, there are no significant differences between the EN-GB authors and the EN-PT^{EU} or the EN-ES^{EU} authors in relation to the frequency of "such", despite native authors using that pronoun more frequently. There are also no significant differences between the EN-GB authors and the Spanish authors writing OSRAs in English in relation to the frequency of the demonstrative pronoun "this". However, significant differences are found between the EN-GB authors and the Portuguese authors writing

OSRAs in English, with the former using the demonstrative pronoun “this” significantly more frequently. Also, the difference in the frequency of the demonstrative pronoun “those” between the L1 English authors writing OSRAs in their L1 and the Portuguese and the Spanish authors writing OSRAs in English is significant. The L1 English authors use the demonstrative pronoun “those” significantly more frequently than the Portuguese and the Spanish authors.

To understand how all English groups use the demonstrative pronouns “this” and “those”, the parsed files of the OSRAs are examined using WordSmith (Scott 2018b).

Table 78 below shows the syntactic structures of the demonstrative pronoun “this” and its distribution across all English corpora. As can be seen, except for the use of “this” as a determiner followed by a noun (N), the L1 English authors use the most frequent syntactic structures with “this” (i.e., 2-4) more frequently than the Portuguese/Spanish authors writing OSRAs in English.

N	Tags with “this”	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	<dem> DET S @>N	643	621	**899**
2	<dem> INDP S @SUBJ>	**333**	166	189
3	<dem> INDP S @P	**51**	17	21
4	<dem> INDP S @	**34**	3	11
5	<dem> DET S @P	1	2	
6	<dem> DET S @SUBJ>	2	1	
7	<dem> INDP S @NPHR		1	2
8	<dem> DET S @SUBJ> @P<			2
9	<dem> DET S &afterpar @>N			1
10	<dem> DET S @<ACC			1
11	<dem> INDP S &afterpar @SUBJ>			1
12	<dem> INDP S &headstop @NPHR		1	
13	<dem> INDP S @ACC>		1	
14	<dem> INDP S @N	1		

Table 78 – Syntactic structures of the demonstrative pronoun “this” in the three English corpora

After examination of the data in Table 78, two analyses are carried out. Given the higher frequency of the demonstrative pronoun “this” when used as a determiner and followed by a noun by the Spanish authors writing in English compared to the L1 English authors, one analysis seeks to understand if such syntactic function of the demonstrative pronoun “this” can mark L1 influence in non-L1 English authors who are L1 Spanish.

Therefore, the frequencies of the demonstrative pronoun “this” in its [<dem> DET S @>N] function and its equivalents in Spanish are compared to examine significant differences between the groups. Based on the unified framework (Jarvis 2010, 2000), the following questions are asked for the EN-PT^{EU} and the EN-ES^{EU} corpora.

Effect of L1 Influence	**L1 influence** **EN-PT^{EU} questions**	**L1 influence** **EN-ES^{EU} questions**
I) Intragroup homogeneity	Are the frequencies of the demonstrative pronoun “this” as a determiner in the EN-PT^{EU} / EN-ES^{EU} OSRAs uniformly distributed?	
II) Intergroup heterogeneity	Are the frequencies of the demonstrative pronoun “this” as a determiner in the EN-PT^{EU} and EN-ES^{EU} OSRAs statistically significantly different?	
III) Cross-language congruity	Are the frequencies of the demonstrative pronoun “this” as a determiner in the EN-PT^{EU} and equivalents in the PT-EU OSRAs statistically similar?	Are the frequencies of the demonstrative pronoun “this” as a determiner in the EN-ES^{EU} and equivalents in the ES-EU OSRAs statistically similar?
IV) Intralingual contrast	Are the frequencies of the demonstrative pronoun “this” as a determiner in the EN-PT^{EU} and EN-GB OSRAs statistically significantly different?	Are the frequencies of the demonstrative pronoun “this” as a determiner in the EN-ES^{EU} and EN-GB OSRAs statistically significantly different?

The Mann-Whitney test is used to assess for a mean difference between the groups given that the data is not normally distributed and has a couple of outliers. Questions in relation to effects I and II are answered together. The level of significance used is $p < .05$ for questions in relation to effects II and IV. Because questions in relation to effects I and III look for uniformity and congruity, respectively, a result of $p > .05$ is associated with an L1 effect. Table 79 below shows the results and mean ranks of all comparisons.

	L1 Influence Effects								
	I) Intra-L1 homogeneity (Levene's test)	II) Inter-L1 heterogeneity (Mann-Whitney Test)	EN-PT^{EU} and EN-ES^{EU} similar in variance but different in means?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-PT^{EU}?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-ES^{EU}?
	Corpora Examined			Corpora Examined			Corpora Examined		
<dem>	EN-PT^{EU} vs. EN-ES^{EU}	EN-PT^{EU} vs. EN-ES^{EU}		EN-PT^{EU} vs. PT-EU	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. ES-EU	EN-ES^{EU} vs. EN-GB	
	Reference *p* values			Reference *p* values			Reference *p* values		
	p > .05 **AND**	*p* < .05?		*p* > .05?	*p* < .05?		*p* > .05?	*p* < .05?	
este/esta este/esta this	*F* = 8.802 *p* = .004	*Z* = -2.962 *p* = .003 Mean ranks: EN-PT^{EU}= 55.73 EN-ES^{EU}= 75.27	*no*	*Z* = -3.312 *p* = .001 Mean ranks: EN-PT^{EU}= 54.58 PT-EU=76.42	*Z* = -.259 *p* = .796 Mean ranks: EN-PT^{EU}= 64.65 EN-GB= 66.35	*no*	*Z* = -7.240 *p* = .001 Mean ranks: EN-ES^{EU}= 61.82 ES-EU= 18.05	*Z* = -2.781 *p* = .005 Mean ranks: EN-ES^{EU}= 74.67 EN-GB= 56.33	*no*

Table 79 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups concerning the frequency of the demonstrative pronoun "this" used as a determiner and its equivalents in Portuguese/Spanish to verify for the potential to function as NLID markers.

Based on the results obtained, no strong effect of L1 influence can be argued concerning the use of the demonstrative pronoun "this" <dem> DET S @>N by the Spanish authors writing OSRAs in English. However, some signs of a characteristic use could be useful in detecting L1 if aggregated with other variables. The Spanish authors significantly increase the frequency of use of the demonstrative pronoun "this" as a determiner followed by a noun in relation to the frequency with which they use the same syntactic structure in Spanish, and this increase is high enough as to be significantly different from both the L1 English authors (EN-GB) and the non-L1 English authors who are L1 Portuguese (EN-PT^{EU}) since they both use the demonstrative pronoun "this" as a determiner followed by a noun less frequently.

Based on these results, the concordances of the demonstrative pronoun "this" with the function <dem> DET S @>N are extracted using WordSmisth (Scott 2018b). As can be seen in Table 80 below, both groups of non-L1 English authors use certain expressions more frequently than the L1 English authors, but the Spanish authors writing in English stand out, as can be seen by the number of occurrences marked in bold.

N	Bigrams with "this"	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	this study	117	**130**	**120**
2	this group	14	**21**	**10**
3	this analysis	29	6	7
4	this work	5	**12**	**17**
5	this effect	4	**7**	**22**
6	this regard	2	2	**26**
7	this population	8	**11**	**10**
8	this finding	9	9	9
9	this difference	8	8	8
10	this approach	6	**9**	8
11	this case	7	5	**11**
12	this reason	8	3	**12**
13	this article	1	1	**20**
14	this association	5	**9**	5
15	this increase	4	3	**12**
16	this result	2	7	**10**
17	this process	5	**7**	**6**
18	this sense			**18**
19	this issue	4	**6**	**7**
20	this protein	2	**6**	**9**
21	this hypothesis		**8**	**8**
22	this fact		**5**	**6**

Table 80 – Most frequent expressions with "this" <dem> DET S @>N in the three English corpora in the CoRA

Three expressions in Table 80, i.e., 6, 18, and 22 are most frequently used by the Spanish authors writing in English. To verify if these expressions could be found in the ES-EU corpus in similar frequencies and thus indicate possible language transfer from the Spanish into English, a concordance is extracted from all corpora using WordSmith (Scott 2018b) and the following query:

> “in this regard/this fact/this sense/a este respeito/neste sentido/a este respecto/en esta línea/este facto/este hecho/neste sentido/en este sentido/en tal sentido/no sentido de”

The results obtained are shown below in Table 81. As can be seen, the expressions “in this sense”, “in this regard”, and “this fact” are not used or almost not used by the L1 English authors and the non-L1 English authors who are L1 Portuguese. However, the Spanish authors writing OSRAs in English seem to maintain the use they make in their L1 of expressions like “en este sentido”, “a este respecto”, “este hecho”, despite decreasing their frequencies in comparison to their use in Spanish.

N	Bi/trigrams with “este/a”/"this"	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	en este sentido	-	**38**	-	-	-
2	no sentido de/neste sentido	17	-	-	-	-
3	in this sense	-	-	-	-	**18**
4	en esta línea	-	**2**	-	-	-
5	a este respecto	-	**4**	-	-	-
6	in this regard	-	-	2	2	**26**
7	este hecho	-	**8**	-	-	-
8	este facto	**13**	-	-	-	-
9	this fact	-	-	-	**5**	**6**

Table 81 – Frequencies of the expressions “in this sense”, “in this regard”, and “this fact” and their equivalents in Portuguese and Spanish extracted from the CoRA

Therefore, the demonstrative pronoun “this” when used as <dem> DET S @>N may function as a marker of non-nativeness, and if associated with expressions like “in this sense” or “this fact” may also be associated with languages like Portuguese and Spanish.

The second analysis concerns the demonstratives pronoun “those” whose syntactic structures and their distribution across all English corpora are shown below in Table 82.

N	Syntactic structures w/ “those”	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	<dem> INDP P @P	**201**	74	121
2	<dem> DET P @>N	**58**	32	37
3	<dem> INDP P @SUBJ>	**52**	7	14
4	<dem> INDP P @	**19**	6	9
5	<dem> DET P @P	**10**	2	1
6	<dem> INDP P @N	**5**	2	1
7	<dem> DET P @SUBJ>	**4**		

Table 82 – Most frequent syntactic structures with the demonstrative pronoun “those” in the three English corpora of the CoRA.

As can be seen, the L1 English authors use the most frequent syntactic structures with “those” more frequently than the Portuguese/Spanish authors writing OSRAs in English. This high frequency also reflects on the expressions used in the OSRAs. Table 83 below shows the distribution of the most frequent expressions with “those” in the three English corpora, with all expressions being more frequent in the EN-GB corpus than in the EN-PT^{EU}/EN-ES^{EU} corpora.

N	Bigrams with "those"	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	those with	80	22	22
2	those + [V past particle][23]	60	31	39
3	those without	19	4	6
4	those in	18	4	5
5	those patients	9	1	7
6	those from	8	1	7
7	those children/women/men	3		1
8	those at	3	1	
9	those found		3	3
10	those shown	1		

Table 83 – Most frequent expressions with “those” in the English corpora of CoRA

A comparison is carried out to understand if the Portuguese/Spanish authors avoid using this demonstrative pronoun when writing OSRAs in English, but as can be seen in Figure 16 below these authors actually increase the frequency of use of the demonstrative pronoun “those” in comparison with the frequency with which they use the equivalent demonstrative pronouns “aqueles” / “aquelas” / “esses” / “essas” in Portuguese, and “aquellos” / “aquellas” / “esos” / “esas” in Spanish. Notwithstanding such an increase, the

[23] It refers to regular verbs only.

Portuguese and Spanish authors still use the demonstrative pronoun “those” significantly less frequently than the L1 English authors.

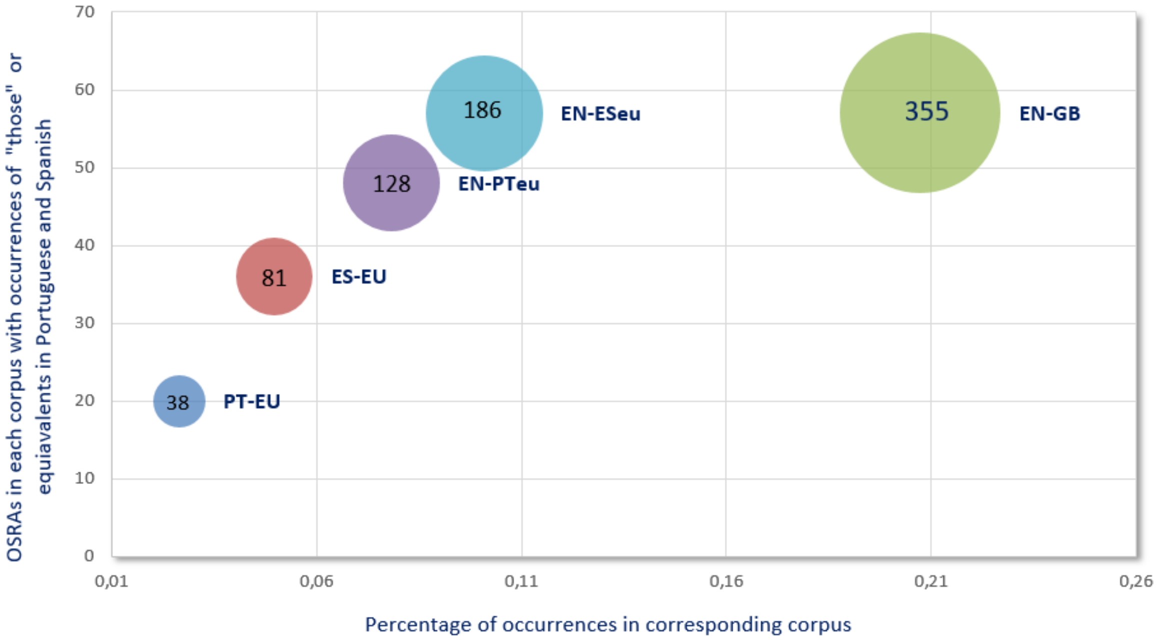


Figure 16 – Distribution of the demonstrative pronoun “those” and its equivalents in the PT-EU and ES-EU corpora.

Hence, the demonstrative pronoun “those” may function as a marker of non-nativeness since the L1 English authors appear to be more comfortable using it than the non-L1 English authors who are L1 Portuguese/Spanish users.

4.5.3. V16: number of adjectives

After examining the samples of V16, the following descriptive statistics are obtained:

V16: frequency of adjectives (all values in words/1000 tokens)							
CORPORA	Mean	Std. Error	95% Confidence Interval for Mean		Median	Variance	Std. Deviation
			Lower Bound	Upper Bound			
PT-EU	113,40	3,03	107,35	119,45	113,00	596,82	24,43
ES-EU	96,11	2,24	91,64	100,59	95,30	326,16	18,06
EN-GB	105,02	2,80	99,44	110,60	99,70	508,05	22,54
EN-PT^{EU}	109,71	2,66	104,40	115,02	107,30	458,82	21,42
EN-ES^{EU}	100,93	2,12	96,69	105,17	98,60	293,44	17,13

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V16 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test, *F* = 2.816, *p* = .096);

II. There are statistically significant differences in the frequency of adjectives between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = 2.582, *p* = 0.11.
The EN-PT^{EU} group uses significantly more adjectives per thousand words than the EN-ES^{EU} group (*MD* = 8.78; *SED* = 3.40; 95% CI = 15.51 to 2.05 adjectives per thousand words);

III. There are no statistically significant differences in frequency of adjectives between the EN-PT^{EU} and PT-EU groups, *t*(128) = .915, *p* = .362.
Both groups use a similar number of adjectives per thousand words (*MD* = 3.69; *SED* = 4.03; 95% CI = to 11.66 to -4.29);

Also, no statistically significant differences were found in the frequency of adjectives between the EN-ES^{EU} and ES-EU groups, *t*(128) = 1.560, *p* = .121.
Both groups have similar means of adjectives per thousand words (*MD* = 4.82; *SED* = 3.09; 95% CI = 10.93 to -1.29 adjectives per thousand words);

IV. There are no statistically significant differences in the frequency of adjectives between the EN-PT^{EU} and EN-GB groups, *t*(128) = 1.217, *p* = .226.
Both groups have fairly similar means of adjectives per thousand words (*MD* = 4.69; *SED* = 3.86; 95% CI = 12.32 to -2.94 adjectives per thousand words).

There are also no statistically significant differences in the frequency of adjectives between the EN-ES^{EU} and EN-GB groups, $t(128) = 1.165$, $p = .246$.
Both groups have similar means of adjectives per thousand words (*MD* = 4.09; *SED* = 3.51; 95% CI = 11.04 to -2.86 adjectives per thousand words).

The following table summarizes the effects of L1 influence found in each group:

Effect of L1 Influence	EN-PT^{EU}	EN-ES^{EU}
I. Intra-L1 homogeneity	✓	
II. Inter-L1 heterogeneity	✓	
III. Cross-language congruity	✓	✓
IV. Intralingual contrast	--	--

For V16 two effects of L1 influence are found for both the EN-PT^{EU} and the EN-ES^{EU} groups. The Levene's test shows that the distributions of adjectives in the OSRAs within the EN-PT^{EU} and the EN-ES^{EU} groups are similar. However, the mean values of the frequencies of adjectives are significantly different between those groups, with the Portuguese authors using significantly more adjectives per thousand words than the Spanish authors when writing OSRAs in English.

Both the Portuguese and the Spanish authors writing OSRAs in English do not differ from their respective L1 counterparts writing OSRAs in their L1 in relation to the frequency of adjectives. The Portuguese authors use as many adjectives when writing OSRAS in Portuguese as they use when writing OSRAs in English; likewise, the Spanish authors use adjectives at similar frequencies when writing OSRAS in Spanish and when writing OSRAs in English.

Additionally, the PT-EU and ES-EU groups are compared, and it is verified that these language groups are also significantly different in relation to V16 ($t(117.860) = 4.588$, $p = .001$) with the L1 Portuguese group using significantly more adjectives per thousand words than the L1 Spanish group, and also more than all the other groups.

In the CoRA, the frequency of use of adjectives (V16) is fairly uniform between each of the non-L1 English groups and the L1 English group (EN-GB) writing OSRAs in English (i.e.

(EN-PT^{EU} vs. EN-GB / EN-ES^{EU} vs. EN-GB), but not between the non-L1 English groups (EN-PT^{EU} vs. EN-ES^{EU}) who additionally are similar to their respective L1 counterparts in terms of adjectives frequencies.

Therefore, the detected effects of L1 influence in relation to adjectives are examined linguistically for both the EN-PT^{EU} or the EN-ES^{EU} groups.

First, the most frequent adjectives in each corpus are extracted and ranked according to their number of occurrences in the corresponding corpus. Table 84 below shows the adjectives extracted per corpus. The brackets indicate that the occurrences are counted as lemmas. That is, in Portuguese and Spanish, the occurrences counted as being of the same lemma are those corresponding to forms inflected for number and/or gender (E.g. "restantes"/"restante", "nueva"/"nuevo"). In English, the forms counted as occurrences of the same lemma are those inflected for grade, to form comparative or superlative adjectives (E.g. "higher"/"high", "greater"/"great").

Rank	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Threshold
1	[doente] n=1104	[mayor] n=459	[high] n=312	[high] n=616	[high] n=494	
2	[grande] n=495	[significativo] n=289	[clinical] n=294	[significant] n=261	[significant] n=301	
3	[alto] n=285	[clínico] n=240	[significant] n=249	[low] n=257	[different] n=282	**95%**
4	[significativo] n=254	[superior] n=178	[low] n=191	[clinical] n=212	[low] n=269	
5	[clínico] n=213	[alto] n=173	[likely] n=187	[different] n=205	[clinical] n=227	
6	[médio] n=178	[primer] n=158	[different] n=155	[increased] n=176	[increased] n=187	**90%**
7	[pequeno] n=169	[menor] n=154	[previous] n=143	[similar] n=157	[previous] n=171	**85%**
8	[elevado] n=156	[diferente] n=147	[increased] n=138	[associated] n=154	[present] n=164	**80%**
9	[baixo] n=141	[medio] n=128	[similar] n=136	[present] n=134	[similar] n=162	**75-70%**
10	[primeiro] n=127	[nuevo] n=124	[important] n=129	[previous] n=130	[specific] n=158	
11	[importante] n=122	[posible] n=120	[great] n=127	[important] n=126	[associated] n=144	
12	[bom] n=118	[importante] n=113	[large] n=125	[mean] n=117	[important] n=128	**65%**
13	[variável] n=115	[previo] n=111	[small] n=123	[positive] n=106	[potential] n=101	
14	[presente] n=113	[similar] n=106	[associated] n=119	[good] n=101	[main] n=95	**60%**
15	[possível] n=110	[bajo] n=95	[recent] n=106	[possible] n=99	[small] n=92	
16	[inferior] n=107	[principal] n=86	[present] n=105	[recent] n=87	[possible] n=90	**55%**
17	[frequente] n=96	[presente] n=82	[additional] n=102	[small] n=84	[new] n=85	
18	[último] n=85	[necesario] n=78	[common] n=101	[common] n=83	[large] n=83	
19	[principal] n=85	[específico] n=67	[possible] n=95	[large] n=80	[recent] n=83	
20	[novo] n=81	[último] n=57	[available] n=80	[specific] n=80	[relevant] n=61	
21	[restante] n=72		[current] n=76	[major] n=77		
22	[semelhante] n=64		[good] n=76			
23	[específico] n=61		[early] n=69			
24	[recente] n=43		[new] n=67			**50%**

Table 84 – Most frequent adjectives in the CoRA (present in at least 50% of the OSRAs) ranked by the number of occurrences

As can be seen in Table 84, the number of adjectives within the 50% threshold is very similar for all groups in the CoRA, with two groups, i.e., the PT-EU and the EN-GB having slightly longer lists.

After the initial observations are made, the three corpora of authors writing OSRAs in English are analyzed, looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. After analysis, the adjectives (ADJs) presented in Table 85 below are deemed unlikely to function as NLID markers given their similar ranks and/or percentages in the corresponding corpus, despite differences in the number of occurrences between the groups.

N	Adjective	Corpus	Occurrences in corpus	Rank	% in corpus
1	[associated]	EN-GB	119	24	0,07
		EN-PT^{EU}	154	12	0,09
		EN-ES^{EU}	144	11	0,08
2	[available]	EN-GB	80	16	0,05
		EN-PT^{EU}	59	43	0,04
		EN-ES^{EU}	46	47	0,02
3	[clinical]	EN-GB	294	7	0,17
		EN-PT^{EU}	212	4	0,13
		EN-ES^{EU}	227	5	0,12
4	[common]	EN-GB	101	17	0,06
		EN-PT^{EU}	83	25	0,05
		EN-ES^{EU}	46	33	0,02
5	[current]	EN-GB	76	23	0,04
		EN-PT^{EU}	43	64	0,03
		EN-ES^{EU}	60	34	0,03
6	[different]	EN-GB	155	2	0,09
		EN-PT^{EU}	205	5	0,13
		EN-ES^{EU}	282	3	0,15
7	[early]	EN-GB	69	21	0,04
		EN-PT^{EU}	60	42	0,04
		EN-ES^{EU}	92	25	0,05
8	[good]	EN-GB	76	22	0,04
		EN-PT^{EU}	101	20	0,06
		EN-ES^{EU}	54	29	0,03
9	[high]	EN-GB	312	1	0,18
		EN-PT^{EU}	616	1	0,38
		EN-ES^{EU}	494	1	0,27
10	[important]	EN-GB	129	3	0,08
		EN-PT^{EU}	126	15	0,08
		EN-ES^{EU}	128	12	0,07

11	[increased]	EN-GB	138	18	0,08
		EN-PT^{EU}	176	6	0,11
		EN-ES^{EU}	187	6	0,10
12	[large]	EN-GB	125	11	0,07
		EN-PT^{EU}	80	28	0,05
		EN-ES^{EU}	83	19	0,05
13	[low]	EN-GB	191	9	0,11
		EN-PT^{EU}	257	3	0,16
		EN-ES^{EU}	269	4	0,15
14	[main]	EN-GB	48	59	0,03
		EN-PT^{EU}	53	49	0,03
		EN-ES^{EU}	95	14	0,05
15	[major]	EN-GB	65	43	0,04
		EN-PT^{EU}	77	30	0,05
		EN-ES^{EU}	39	31	0,02
16	[mean]	EN-GB	84	29	0,05
		EN-PT^{EU}	117	17	0,07
		EN-ES^{EU}	50	76	0,03
17	[new]	EN-GB	67	19	0,04
		EN-PT^{EU}	56	47	0,03
		EN-ES^{EU}	85	17	0,05
18	[possible]	EN-GB	95	13	0,06
		EN-PT^{EU}	99	21	0,06
		EN-ES^{EU}	90	16	0,05
19	[present]	EN-GB	105	14	0,06
		EN-PT^{EU}	134	13	0,08
		EN-ES^{EU}	164	8	0,09
20	[previous]	EN-GB	143	8	0,08
		EN-PT^{EU}	130	14	0,08
		EN-ES^{EU}	171	7	0,09
21	[recent]	EN-GB	106	15	0,06
		EN-PT^{EU}	87	23	0,05
		EN-ES^{EU}	83	18	0,05
22	[significant]	EN-GB	249	5	0,15
		EN-PT^{EU}	261	2	0,16
		EN-ES^{EU}	301	2	0,16
23	[similar]	EN-GB	136	6	0,08
		EN-PT^{EU}	157	10	0,10
		EN-ES^{EU}	162	9	0,09
24	[small]	EN-GB	123	12	0,07
		EN-PT^{EU}	84	24	0,05
		EN-ES^{EU}	92	15	0,05

Table 85 – Adjectives found to be unlikely to function as NLID markers given their similar ranks and their percentage in the corresponding corpus.

As can be verified, in Tables 84 and 85 a number of adjectives are common to all the corpora. These adjectives are [significant], [possible], [important], [clinical], [present], and [similar] in English and its equivalents [significativo], [possível], [importante], [clínico], [presente], and [semelhante] in Portuguese, and [significativo], [possible], [importante], [clínico], [presente], and [similar] in Spanish.

Another group of high ranked adjectives "denoting properties in the domain of size" (Pullum and Huddleston 2002: 527) or breath/degree and that are also transversal to all the corpora in CoRA are [high], [low], [large], [small], and [increased] in English and their equivalents [alto], [baixo], [grande], [pequeno], [elevado] and [inferior] in Portuguese and [mayor], [menor], [alto], [superior] and [bajo] in Spanish.

Since these adjectives are common to all the corpora and are all present in at least 50% of the OSRAs at a number of occurrences of 186 on average, they seem to operate rather as part of the scientific register in the field of health sciences, regardless of the language. Their potential to act as NLID markers is very limited, and thus, these are not examined.

Also, adjectives like [different], [previous], and [recent] are less likely to function as NLID markers since they appear in all English corpora at equal or similar frequencies or in frequencies that are higher in the EN-GB corpora. Additionally, the equivalents of these adjectives are found in lower frequencies in both the PT-EU and the ES-EU corpora, meaning that their higher expression in English is not influenced by the frequency of use in the authors' L1, but most likely by a frequent use within the register.

After the lemmas of the adjectives that are less plausible NLID markers are removed, two groups of adjectives are analyzed.

The first group contains adjectives found in the EN-GB corpus more frequently than in the EN-PT^{EU} and EN-ES^{EU} corpora, and therefore, are analyzed to verify their potential to mark possible strategies of avoidance of use by the non-L1 authors. These nouns are shown in Table 86.

N	Adjective	Corpus	Occurrences in corpus	Rank	% in corpus
1	[additional]	EN-GB	102	20	0,06
		EN-PT^{EU}	25	132	0,02
		EN-ES^{EU}	44	28	0,02
2	[great]	EN-GB	127	10	0,07
		EN-PT^{EU}	51	52	0,03
		EN-ES^{EU}	42	51	0,02
3	[likely]	EN-GB	187	4	0,11
		EN-PT^{EU}	29	106	0,02
		EN-ES^{EU}	41	66	0,02

Table 86 – Adjectives analyzed to verify their potential to mark possible strategies of avoidance of use by non-L1 authors.

The frequencies of the adjectives [additional], [great], and [likely] are compared to see if there are significant differences between the groups. The fourth L1 effect of the unified framework (Jarvis 2010, 2000) is tested for both the EN-PT^{EU} and the EN-ES^{EU} OSRAs, stated as follows:

Effect of L1 Influence	L1 influence EN-PT^{EU} question	L1 influence EN-ES^{EU} question
IV) Intralingual contrast	Are the frequencies of the adjective [additiona]/[great]/[likely] in the EN-PT^{EU}/ EN-ES^{EU} and the EN-GB corpora statistically significantly different?	

Since the data is not normally distributed and has some outliers, the Mann-Whitney test is used to assess for a mean difference between the groups. The level of significance used is $p < .05$. Two tests are performed. The numbers of occurrences are normalized by 100. Table 87 shows the results obtained.

	IV - Intralingual contrast (Mann-Whitney Test)	Effect of L1 influence IV found for the EN-PT^{EU}?	IV - Intralingual contrast (Mann-Whitney Test)	Effect of L1 influence IV found for the EN-ES^{EU}?
	corpora examined			
Adjective	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. EN-GB	
	***p* reference value < .05**			
[additional]	*Z* = -1.968 *p* = .049 *M* rank EN-GB = 29.17 *M* rank EN-PT^{EU} = 21.00	*yes*	*Z* = -2.491 *p* = .013 *M* rank EN-GB = 36.63 *M* rank EN-ES^{EU} = 26.21	*yes*
[great]	*Z* = -2.004 *p* = .045 *M* rank EN-GB = 40.65 *M* rank EN-PT^{EU} = 30.78	*yes*	*Z* = -1.808 *p* = .071 *M* rank EN-GB = 36.79 *M* rank EN-ES^{EU} = 27.88	*no*
[likely]	*Z* = -3.038 *p* = .002 *M* rank EN-GB = 39.23 *M* rank EN-PT^{EU} = 23.03	*yes*	*Z* = -1.903 *p* = .057 *M* rank EN-GB = 38.25 *M* rank EN-ES^{EU} = 28.11	*no*

Table 87 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups in relation to the adjectives [additional], [great] and [likely] deemed as likely to mark strategies of avoidance

As can be seen in Table 87, the Mann Whitney tests indicate statistically significant differences in the ranked frequencies of the adjective [additional] between the L1 and both of the non-L1 English groups writing in English. Also, the adjectives [great] and [likely] are significantly more frequently used by the L1 (i.e., EN-GB) than by the non-L1 authors who are Portuguese L1 users (EN-PT^{EU}) writing OSRAs in English. However, there are no significant differences in the frequency of these two adjectives between the EN-GB and the EN-ES^{EU} groups.

Based on the significance of the results, the adjective [additional] is further examined for both groups (the EN-PT^{EU} and the EN-ES^{EU}). The examinations are based on the concordances of the parsed files. The concordances are obtained with WordSmith 7.0 (Scott 2018b), from which the syntactic tags containing ADJ are extracted. As can be seen below in Table 88, the syntactic structure associated with the adjective [additional] in the EN-GB corpus that most contributes to the significant differences in the number of occurrences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups is [ADJ POS @>N], i.e., prenominal adjective modifying a noun.

N	Followed by a token with the syntactic function	Example	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	ADJ POS @>N	[additional] model/analysis/samples	98	25	43
2	ADJ POS @SUBJ>	whereas an [additional] six are	2	-	-
3	ADJ POS @N	[additional] inflammation	1	-	-
4	ADJ POS @P	collagen centres in [additional] to a metabolism cluster	1	-	-
5	ADJ POS @	information that is [additional] to the clinical findings	-	-	1

Table 88 – Syntactic tags of the adjective [additional] showing the significant differences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups

When the concordances of the adjective [additional] in the English corpora are analyzed, 130 unique combinations of words are found. Most of these combinations are strings of [ADJ + N] (as shown in Table 88, nº 1) or [ADJ + ADJ + N] (e.g. "additional prospective studies") or [ADJ + N + N] (e.g. "additional section membership"). However, the string contributing most to the significant difference between the EN-GB and the non-L1 English corpora is the combination [ADJ + N], specifically the phrase [additional file] (see Table 89), whose occurrences are mostly found in one OSRA in the EN-GB corpus in reference to material that the reader can consult for more information on the research being described in the article.

N	Bigrams w/[additional]	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	additional file	17	5	4
2	additional evidence		2	2
3	additional studies			4
4	additional analysis	3		
5	additional factors		1	2
6	additional research	1		2
7	additional training	3		
8	additional genes			2
9	additional mutations			2
10	additional prospective studies			2
11	additional samples	2		

Table 89 – Most frequent combinations of words with [additional] according to the frequency in the English corpora of the CoRA

Therefore, the difference in the frequency of the adjective “additional” in the L1 and the non-L1 English corpora is not because the non-L1 authors avoid using that word but because there is one combination of words with “additional” that is more frequently used in a specific context of one specific file within the EN-GB corpus.

The concordances of the lemma [great] show that most of the differences between the L1 (i.e., EN-GB) and the non-L1 English authors (i.e., EN-PT^{EU}) are found in the use of the comparative and superlative forms of [great], as shown in Table 90 below.

N	**Followed by a token with the syntactic function**	**Example**	**EN-GB**	**EN-PT^{EU}**
1	ADJ COM @>N	**greater** propensity/diversity/stability	70	24
2	ADJ COM @	mean distance **greater** or equal to that observed	19	4
3	ADJ SUP @>N	**greatest** improvements/benefits/increases	12	3
4	ADJ POS @>N	the **great** majority/a great advantage	8	15
5	ADJ P COM @	4.7 times **greater** than in black women	7	
6	ADJ COM @N	almost 50% **greater** than those in London	5	3
7	ADJ COM @P	genes with **greater** than 50 percent of the nucleotide sequence of the array present in mature mRNAs	3	1
8	ADJ SUP @	was **greatest** when abuse was admitted	2	1
9	ADJ COM @SC>	That is, the less deprived the area, the **greater** the prevalence of CD	1	

Table 90 – Syntactic concurrences of the lemma [great] with frequencies

The specific phrases containing the lemma [great] that justified the significant differences between the EN-GB and the EN-PT^{EU} groups are shown in Table 91. As can be seen, the comparative and superlative forms are more frequently used by the L1 English authors than by the Portuguese authors writing in English.

N	Word combinations w/[great]/[greater]/[greatest]	EN-GB	EN-PT^{EU}
1	greater [N] (e.g. greater chance)	51	8
2	greater than …	25	5
3	greatest [N] (e.g. greatest scope)	12	3
4	greater in …	11	15
5	great [N] (e.g. great deal, great variability)	5	5
6	greater/greatest risk of/for/than	5	3
7	the greater the …	4	1
8	greater when …	2	
9	greater among …	1	
10	greater [N + N] (e.g. greater mentorship quality)	2	2
11	greater [ADJ+N] (e.g. greater tensile strength)	2	

Table 91 – Most frequent concordances of [great] that justify the significant differences between the EN-GB and the EN-PT^{EU} groups

The comparative and superlative uses of [great] may be due to the need to report comparisons between samples of specific types of trials or studies reported in the EN-GB OSRAs. While these forms are also found in the EN-PT^{EU} OSRAs reporting trials or studies implying comparisons, their lower frequency may be due to a lower number of OSRAs reporting comparative research needing to resort to comparative/superlative forms of [great]. To understand if the comparative/superlative of [great] are avoided by non-L1 authors writing OSRAs in English more OSRAs are needed to increase the number of tokens in all English corpora.

In relation to the adjective [likely], the concordances of the syntactic tags show that the significant differences in the frequency of that adjective between the L1 (i.e., EN-GB) and the non-L1 English authors (i.e., EN-PT^{EU}) are found in most of the syntactic functions but especially in the uses of [likely] as a modifier of nouns either in their plural (Table 92, number 1) or singular forms (Table 92, number 2).

Although the statistical comparisons in relation to the frequency of [likely] in the English corpora take into account only the occurrences of [likely] as a simple form, Table 92 shows also the frequencies of the forms [most=likely] and [unlikely] which are also more frequent in OSRAs written by the L1 than by the non-L1 English authors writing in that language.

Nº	Syntactic tags	Examples	EN-GB	EN-PT^{EU}
	[likely]		**187**	**26**
1	ADJ P POS @	Patients with bvFTD were more **likely** to exhibit…	94	7
2	ADJ POS @	It seems **likely** therefore that there is…	69	16
3	ADJ POS @>N	… providing further support for the likely role of…	10	1
4	ADJ POS @AS	Elevated concentrations of epithelial cells were twice as likely in the Peezy group compared with the controls (OR 2.1 (95% CI 1.2 to 3.7)) when controlled for significant variables in the univariate analysis (eGFR and underlying diagnosis).	4	--
5	ADJ POS @N	For HIV clinic appointments, people more likely to be in class 1 most favoured seeing an HIV consultant of all the HCP options…	9	1
6	ADJ POS @PRED>	Those more likely to be in class 2 were indifferent between this and only having access to their non-HIV records.	1	--
	[most=likely]		**10**	**3**
7	ADJ P POS @	those who are most likely to view primary care as an alternative have disclosed their HIV status	6	1
8	ADJ POS @	.. which is most likely due to arrival by car..	2	1
9	ADJ POS @>N	That failure to access medical services is the most likely reason for lower diagnosis rates in…	2	1
	[unlikely]		**28**	
10	ADJ POS @	Much like the diagnosis of cancer, the new diagnosis of coronary heart disease, […], is unlikely to improve quality of life.	24	--
11	ADJ P POS @	which clinicians are unlikely to ignore in practice	4	--

Table 92 – Most frequent syntactic tags that justify the significant differences found between EN-GB and EN-PT^{EU} authors in relation to the frequency of [likely]

The most frequent word combinations containing the adjective [likely] are shown below in Table 93. As can be seen, most of the instances are more frequently found in the EN-GB corpus than in the EN-PT^{EU} corpus.

№	Word combinations w/[likely]	EN-GB	EN-PT^{EU}
1	likely [V INF]	92	12
2	likely to be …	71	6
3	it is likely that …	10	7
4	*other unique phrases with* [likely]	15	2
5	likely [V PPar]	2	1
6	likely due to …	4	1
7	likely [N]	8	
8	less likely	7	
9	most likely	1	3
10	more likely	4	1
11	likely [V PR 3P]	1	3
12	likely because …		2

Table 93 – Most frequent word combination using [likely] in the EN-GB and the EN-PT^{EU} corpora

Taking into consideration that the word combinations containing [likely] do not convey terminological but rather general meanings, it can be argued that the non-L1 English authors who are L1 Portuguese users may be avoiding using [likely] in favor of synonyms like [probable] or even [possible] when writing OSRAs in English. However, upon examination of the EN-GB and the EN-PT^{EU} corpora, it is verified that the occurrences of [possible] are very similar in both groups, i.e., EN-GB = 99 occurrences and EN-PT^{EU} = 95 occurrences; and the occurrences of [probable] although higher in the (EN-GB = 9 occurrences) are very few in either corpus (i.e., EN-PT^{EU} = 2). Therefore, [likely] could be associated with avoidance strategies of the EN-PT^{EU} authors in the CoRA, but the same analyses must be performed in larger corpora of EN-GB and EN-PT^{EU} OSRAs to support this finding.

After examining the lemmas of the adjectives [additional], [great], [likely], the remaining lemmas are examined taking into account that their (a) occurrences, (b) ranks, and (c) percentages of occurrences are higher in both or one of the non-L1 English corpora (EN-PT^{EU} and EN-ES^{EU}) than in the L1 (EN-GB) corpus and therefore, are likely to function as NLID markers of language transfer if similar occurrences, ranks, and percentages are verified in the L1 PT-EU and ES-EU corpora. Since these adjectives are chosen as possible markers of language transfer, the lemmas of their equivalents in the L1 Portuguese/Spanish corpora are extracted, and their frequencies are compared. Table 94 shows the adjectives that, after comparison, are not further analyzed because the frequencies of their equivalents' in the corresponding L1 corpora do not justify the high frequencies found in the non-L1 English corpora (one or both), and therefore, the statistical comparison to test the L1 effect cannot be performed.

Nº	Adjective	Corpus	Occurrences in corpus	Rank	% in corpus
1	[potencial]	PT-EU	8	377	0,01
	[potencial]	ES-EU	31	104	0,02
	[potential]	EN-GB	49	27	0,03
		EN-PT^{EU}	42	68	0,03
		EN-ES^{EU}	101	13	0,05
2	[relevante]	PT-EU	23	151	0,02
	[relevante]	ES-EU	42	68	0,03
	[relevant]	EN-GB	37	67	0,02
		EN-PT^{EU}	50	54	0,03
		EN-ES^{EU}	61	20	0,03
3	[específico]	PT-EU	61	40	0,04
	[específico]	ES-EU	67	29	0,04
	[specific]	EN-GB	70	31	0,04
		EN-PT^{EU}	80	29	0,05
		EN-ES^{EU}	158	10	0,09

Table 94 – Adjectives that upon analysis are disregarded as NLID markers of language transfer

Only one adjective, i.e., [positive] is deemed likely to mark language transfer effects in OSRAs written in English by the L1 Portuguese/Spanish authors since the frequencies of their equivalents in the corresponding L1 corpora (i.e., the PT-EU and the ES-EU) are equal or higher than the frequencies of [positive] in the English corpora.

Nº	Adjectives	Corpus	Occurrences in corpus	Rank	% in corpus
1	[positivo]	PT-EU	107	18	0,07
	[positivo]	ES-EU	62	35	0,04
	[positive]	EN-GB	56	45	0,03
		EN-PT^{EU}	106	19	0,06
		EN-ES^{EU}	62	32	0,03

Table 95 – Analysis of the lemma [positive] to test for L1 transfer effects

The frequencies of the lemma [positive], and its equivalents in Portuguese and Spanish [positivo] and [positivo], are compared to examine significant differences between the groups. For all adjectives, the following questions are asked for both the EN-PT^{EU} and the EN-ES^{EU} corpora.

Effect of L1 Influence	**L1 influence** **EN-PT^{EU} questions**	**L1 influence** **EN-ES^{EU} questions**
I) Intragroup homogeneity	Are the frequencies of the adjective [positive] in the EN-PT^{EU} / EN-ES^{EU} OSRAs uniformly distributed?	
II) Intergroup heterogeneity	Are the frequencies of the adjective [positive] in the EN-PT^{EU} and EN-ES^{EU} OSRAs statistically significantly different?	
III) Cross-language congruity	Are the frequencies of the adjective [positive] in the EN-PT^{EU} OSRAs and the equivalent [positivo] in the PT-EU OSRAs statistically similar?	Are the frequencies of the adjective [positive] in the EN-ES^{EU} OSRAs and the equivalent [positivo] in the PT-EU OSRAs statistically similar?
IV) Intralingual contrast	Are the frequencies of the adjective [positive] in the EN-PT^{EU} and EN-GB OSRAs statistically significantly different?	Are the frequencies of the adjective [positive] in the EN-ES^{EU} and EN-GB OSRAs statistically significantly different?

The Mann-Whitney's test is used to assess for a mean difference between the groups given that the data is not normally distributed and has some outliers. Questions in relation to the effects I and II are answered together. The level of significance used for questions in relation to the effects II and IV is $p < .05$. Because questions in relations to the effects I and III look for uniformity and congruity, respectively, a result of $p > .05$ is associated with an L1 effect. Table 96 below shows the results and mean ranks of all comparisons.

	L1 Influence Effects								
	V) Intra-L1 homogeneity (Levene's test)	VI) Inter-L1 heterogeneity (Mann-Whitney Test)	EN-PT^{EU} and EN-ES^{EU} similar in variance but different in means?	VII) Cross-language congruity (Mann-Whitney Test)	VIII) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-PT^{EU}?	V) Cross-language congruity (Mann-Whitney Test)	VI) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-ES^{EU}?
	Corpora Examined			**Corpora Examined**			**Corpora Examined**		
Adjective	EN-PT^{EU} vs. EN-ES^{EU}	EN-PT^{EU} vs. EN-ES^{EU}		EN-PT^{EU} vs. PT-EU	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. ES-EU	EN-ES^{EU} vs. EN-GB	
	Reference *p* values			**Reference *p* values**			**Reference *p* values**		
	p > .05 **AND**	*p* < .05?		*p* > .05?	*p* < .05?		*p* > .05?	*p* < .05?	
[positivo]/[positivo]/[positive]	*F* = 2.269 *p* = .137	*Z* = -.974 *p* = .330 Mean ranks: EN-PT^{EU}= 32.36 EN-ES^{EU}= 28.22	*no*	*Z* = -.307 *p* = .759 Mean ranks: EN-PT^{EU}= 32.32 PT-EU= 33.70	*Z* = -1.061 *p* = .289 Mean ranks: EN-PT^{EU}= 31.41 EN-GB= 26.98	*no*	*Z* = -.400 *p* = .689 Mean ranks: EN-ES^{EU}= 26.70 ES-EU= 28.30	*Z* = -.153 *p* = .878 Mean ranks: EN-ES^{EU}= 26.78 EN-GB= 26.20	*no*

Table 96 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups in relation to the frequency of the adjective [positive] and its equivalents [positivo] and [positivo] in Portuguese and Spanish

As shown in Table 96, no statistically significant differences are found between the L1 (EN-GB) and the non-L1 (EN-PT^{EU}) English authors writing in English in relation to the ranked frequencies of the adjective lemma [positive]. Therefore, no L1 influence effects can be argued.

Overall, the analyses of the adjectives in the five corpora comprising the CoRA show that there is at least one instance of possible avoidance of use in relation to the adjective [likely], but no adjective could be associated with any effect of L1 influence.

4.5.4. V19: number of verbs

After examining the samples of V19, the following descriptive statistics are obtained:

V19: frequency of verbs (all values in words/1000 tokens)							
CORPORA	Mean	Std. Error	95% Confidence Interval for Mean		Median	Variance	Std. Deviation
			Lower Bound	Upper Bound			
PT-EU	115,26	1,42	112,42	118,11	115,30	131,79	11,48
ES-EU	105,56	1,52	102,53	108,59	104,90	149,57	12,23
EN-GB	164,28	3,07	158,14	170,41	164,30	613,55	24,77
EN-PT^{EU}	153,16	2,20	148,75	157,56	149,60	315,42	17,76
EN-ES^{EU}	160,19	2,27	155,64	164,73	161,40	336,36	18,34

The results of the independent sample *t*-tests indicate that:

I. The variances of the means of V19 of the EN-PT^{EU} and EN-ES^{EU} groups are not significantly different (Levene's test, *F* = .272, *p* = .603);

II. However, there are statistically significant differences in the frequency of verbs between the EN-PT^{EU} and EN-ES^{EU} groups, *t*(128) = 2.221, *p* = .028.
Authors in the EN-ES^{EU} use significantly more verbs per thousand words than the EN-PT^{EU} group (*MD* = 7.03; *SED* = 3.17; 95% CI = 132.96 to .766 verbs per thousand words);

III. There are statistically significant differences in frequency of verbs between the EN-PT^{EU} and PT-EU groups, *t*(109.526) = 14.447, *p* = .001.
The EN-PT^{EU} OSRAs have significantly more verbs per thousand words than the PT-EU OSRAs (*MD* = 37.89; *SED* = 2.62; 95% CI = 43.09 to 32.69);

There also are statistically significant differences in the frequency of verbs between the EN-ES^{EU} and ES-EU groups, *t*(111.530) = 19.983, *p* = .001.
The EN-ES^{EU} OSRAs have significantly more verbs per thousand words than the ES-EU OSRAs (*MD* = 54.63; *SED* = 2.73; 95% CI = 60.04 to 49.21 verbs per thousand words);

IV. There are statistically significant differences in the frequency of verbs between the EN-PT^{EU} and EN-GB groups, *t*(128) = 2.972, *p* = .004.
The EN-GB sample has significantly more verbs per thousand words than the EN-PT^{EU} sample (*MD* = 11.12; *SED* = 3.78; 95% CI = 18.60 to 3.64 verbs per thousand words).

There are no statistically significant differences in the frequency of verbs between the EN-ES^{EU} and EN-GB groups, *t*(117.951) = 1.070, *p* = .287.
Both samples have similar means of verbs per thousand words (*MD* = 4.09; *SED* = 3.82; 95% CI = 11.66 to -3.47 verbs per thousand words).

The following table summarizes the effects of L1 influence found:

<table>
<tr><th>Effect of L1 Influence</th><th>EN-PT^{EU}</th><th>EN-ES^{EU}</th></tr>
<tr><td>I. Intra-L1 homogeneity</td><td colspan="2">✓</td></tr>
<tr><td>II. Inter-L1 heterogeneity</td><td colspan="2">✓</td></tr>
<tr><td>III. Cross-language congruity</td><td>--</td><td>--</td></tr>
<tr><td>IV. Intralingual contrast</td><td>✓</td><td>--</td></tr>
</table>

Effects of L1 influence are found for the EN-PT^{EU} and the EN-ES^{EU} groups in relation to the frequency of verbs per thousand words. For the EN-PT^{EU}, three effects are found, and for the EN-ES^{EU}, only two. The results of the Levene's test performed to determine homogeneity of variances in the EN-PT^{EU} and the EN-ES^{EU} groups indicate that the OSRAs written in English by the PT-EU and ES-EU authors are similar in their distribution of verbs. At the same time, these groups are significantly different in relation to the frequency at which they use verbs. The EN-ES^{EU} uses significantly more verbs per thousand words than the EN-PT^{EU} group. The Spanish authors writing OSRAs in English use as many verbs as the Spanish authors writing OSRAs in their L1 and as many verbs as the L1 English authors writing in English. The Portuguese authors writing OSRAs in English also use as many verbs as the Portuguese authors writing OSRAs in their L1. However, compared to the L1 English authors writing OSRAs in English, the Portuguese authors writing OSRAs in English use significantly fewer verbs per thousand words.

Following these results, a linguistic analysis is performed to examine the verbs that may function as L1 influence markers. Table 97 below shows the lemmas of the most frequent verbs present in at least 50% of the corpora.

Rank	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Threshold
1	[ser] n=3307	[ser] n=2324	[be] n=6190	[be] n=5611	[be] n=5811	**95%**
2	[ter] n=859	[haber] n=934	[have] n=1659	[have] n=1194	[have] n=1313	
3	[poder] n=565	[poder] n=597	[may] n=477	[show] n=934	[can] n=601	
4	[apresentar] n=535	[estar] n=334	[use] n=400	[report] n=516	[show] n=1409	
5	[estar] n=393	[tener] n=434	[do] n=533	[may] n=544	[do] n=629	**90%**
6	[verificar] n=323	[realizar] n=318	[can] n=398	[can] n=344	[may] n=468	
7	[realizar] n=282	[incluir] n=175	[will] n=305	[observe] n=570	[find] n=384	
8	[associar] n=206	[mostrar] n=312	[suggest] n=437	[do] n=384	[compare] n=254	**85%**
9	[existir] n=179	[observar] n=298	[show] n=763	[consider] n=335	[use] n=355	
10	[dever] n=191	[obtener] n=175	[compare] n=272	[suggest] n=376	[observe] n=698	**80%**
11	[considerar] n=160	[presentar] n=293	[include] n=175	[find] n=344	[increase] n=242	
12	[incluir] n=121	[ir] n=152	[report] n=638	[use] n=284	[suggest] n=487	
13	[encontrar] n=212	[encontrar] n=241	[identify] n=287	[compare] n=288	[report] n=436	
14	[avaliar] n=169	[relacionar] n=164	[find] n=170	[describe] n=208	[relate] n=154	
15	[permitir] n=130	[utilizar] n=163	[see] n=416	[increase] n=199	[associate] n=165	**75%**
16	[haver] n=165	[demostrar] n=121	[provide] n=159	[reduce] n=178	[reduce] n=400	
17	[utilizar] n=141	[asociar] n=153	[reduce] n=231	[present] n=190	[will] n=154	**70%**
18	[demonstrar] n=122	[deber] n=139	[consider] n=188	[involve] n=176	[demonstrate] n=151	
19	[obter] n=183	[hacer] n=123	[increase] n=170	[decrease] n=145	[involve] n=244	
20	[descrever] n=133	[considerar] n=152	[assess] n=90	[reveal] n=142	[indicate] n=329	**65%**
21	[analisar] n=111	[existir] n=151	[make] n=266	[associate] n=193	[induce] n=247	
22	[relacionar] n=93	[describir] n=130	[demonstrate] n=138	[take] n=150	[describe] n=237	
23	[referir] n=141	[permitir] n=119	[lead] n=88	[identify] n=97	[consider] n=218	
24	[ocorrer] n=96	[resultar] n=91	[remain] n=89	[detect] n=127	[perform] n=106	**60%**
25	[observar] n=119	[dar] n=93	[observe] n=315	[evaluate] n=126	[lead] n=102	
26	[constituir] n=79	[tratar] n=127	[take] n=252	[occur] n=124	[affect] n=111	
27	[identificar] n=129	[seguir] n=94	[require] n=143	[explain] n=122	[include] n=105	
28	[comparar] n=68	[conocer] n=85	[associate] n=112	[include] n=113	[determine] n=220	
29	[aumentar] n=77	[comparar] n=72	[follow] n=106	[demonstrate] n=111	[confirm] n=161	
30	[variar] n=76	[aumentar] n=113	[describe] n=185	[see] n=148	[support] n=143	

31	[reduzir] n=65	[estudiar] n=71	[improve] n=113	[express] n=124	[shall] n=105	**55%**
32	[fazer] n=98	[destacar] n=71	[present] n=113	[relate] n=104	[seem] n=87	
33	[revelar] n=84	[analizar] n=91	[define] n=70	[confirm] n=133	[evaluate] n=85	
34	[elevar] n=77	[producir] n=89	[give] n=135	[lead] n=111	[detect] n=132	**50%**
35	[determinar] n=61	[evaluar] n=100	[shall] n=72	[seem] n=97	[cause] n=97	
36	[representar] n=74	[indicar] n=75	[represent] n=67	[will] n=116	[assess] n=89	
37	[tratar] n=68	[basar] n=86	[indicate] n=166	[perform] n=106	[provide] n=84	
38			[support] n=115	[follow] n=91	[explain] n=168	
39			[occur] n=94	[affect] n=86	[analyze] n=117	
40			[cause] n=79	[require] n=84	[obtain] n=103	
41			[base] n=74	[assess] n=93	[study] n=81	
42				[correlate] n=92	[follow] n=61	
43				[release] n=106	[remain] n=60	
44				[support] n=102	[present] n=117	
45				[shall] n=96	[play] n=87	
46				[induce] n=87	[result] n=79	
47				[carry] n=80	[need] n=62	
48				[maintain] n=78	[occur] n=70	
49				[determine] n=124	[take] n=228	
50				[make] n=117	[decrease] n=82	
51				[develop] n=100		
52				[remain] n=86		
53				[give] n=82		
54				[allow] n=78		
55				[know] n=132		
56				[indicate] n=94		
57				[bind] n=62		
58				[control] n=58		
59				[contribute] n=94		
60				[cause] n=65		
61				[obtain] n=66		
62				[exclude] n=74		

63				[improve] n=68		
64				[provide] n=67		**50%**
65				[treat] n=61		
66				[receive] n=74		
67				[represent] n=58		
68				[result] n=69		
69				[range] n=62		
70				[propose] n=60		
71				[reflect] n=60		
72				[produce] n=59		
73				[apply] n=56		
74				[promote] n=59		
75				[analyse] n=46		
76				[predict] n=47		
77				[establish] n=56		
78				[reach] n=44		
79				[achieve] n=43		
80				[modulate] n=41		

Table 97 – Most frequent verbs in the CoRA (present in at least 50% of the OSRAs) ranked by the number of occurrences with the threshold

As can be seen in Table 97, the 50% threshold is very broad for the grammatical category verb [V] since all groups have more than thirty-five different verbs distributed in half of the corresponding corpus. Since the verb is the category that expresses action within the phrase, the distribution of verbs is expected to be high compared, for example, to adjectives, since using verbs is unavoidable. Despite the amount of data extracted being very extensive, the same threshold is used on the grounds of methodological consistency in relation to the analysis carried with the other parts of speech.

The first observations are made after the extraction of verbs-related data from the CoRA. As can be seen, the corpus of the OSRAs written in English by the authors who are L1 Portuguese has the longest, therefore most diverse list of verbs distributed in half of the corresponding corpus, followed by the L1 Spanish authors writing in English, and lastly by the L1 English authors writing in their L1. The L1 Portuguese and Spanish corpora have exactly the same number of verbs distributed in at least half the OSRAs within the corresponding corpus. The difference between the Spanish authors writing OSRAs in their L1 and the Spanish authors writing in English is almost none. However, the Portuguese authors writing OSRAs in English use twice the number of verbs used by the Portuguese authors writing OSRAs in their L1.

After completing the initial observations, the three corpora of authors writing OSRAs in English are analyzed, looking at (a) the number of occurrences, (b) the rank, and (c) the percentage of occurrences in the corresponding corpus. After analysis, the verbs (V) presented in Table 98 below are deemed unlikely to function as NLID markers given their similar ranks and/or percentages in the corresponding corpus, despite differences in the number of occurrences between the groups.

N	Verb	Corpus	Occurrences in corpus	Rank	% in corpus
1	[achieve]	EN-GB	58	76	0,03
		EN-PT^{EU}	43	89	0,03
		EN-ES^{EU}	42	113	0,02
2	[affect]	EN-GB	62	72	0,04
		EN-PT^{EU}	86	51	0,05
		EN-ES^{EU}	111	44	0,06
3	[allow]	EN-GB	56	82	0,03
		EN-PT^{EU}	78	57	0,05
		EN-ES^{EU}	46	105	0,02

4	[analy(s/z)e]	EN-GB	59	79	0,03
		EN-PT^{EU}	46	85	0,03
		EN-ES^{EU}	117	38	0,06
5	[apply]	EN-GB	50	96	0,03
		EN-PT^{EU}	56	77	0,03
		EN-ES^{EU}	30	137	0,02
6	[assess]	EN-GB	90	46	0,05
		EN-PT^{EU}	93	47	0,06
		EN-ES^{EU}	89	59	0,05
7	[base]	EN-GB	74	55	0,04
		EN-PT^{EU}	48	82	0,03
		EN-ES^{EU}	63	79	0,03
8	[be]	EN-GB	6190	1	3,62
		EN-PT^{EU}	5611	1	3,43
		EN-ES^{EU}	5811	1	3,15
9	[carry]	EN-GB	47	101	0,03
		EN-PT^{EU}	80	55	0,05
		EN-ES^{EU}	118	37	0,06
10	[cause]	EN-GB	79	53	0,05
		EN-PT^{EU}	65	64	0,04
		EN-ES^{EU}	97	52	0,05
11	[compare]	EN-GB	272	14	0,16
		EN-PT^{EU}	288	12	0,18
		EN-ES^{EU}	254	14	0,14
12	[confirm]	EN-GB	129	32	0,08
		EN-PT^{EU}	133	24	0,08
		EN-ES^{EU}	161	27	0,09
13	[contribute]	EN-GB	53	86	0,03
		EN-PT^{EU}	94	46	0,06
		EN-ES^{EU}	90	56	0,05
14	[define]	EN-GB	70	61	0,04
		EN-PT^{EU}	35	109	0,02
		EN-ES^{EU}	38	118	0,02
15	[demonstrate]	EN-GB	138	28	0,08
		EN-PT^{EU}	111	35	0,07
		EN-ES^{EU}	151	30	0,08
16	[describe]	EN-GB	185	19	0,11
		EN-PT^{EU}	208	14	0,13
		EN-ES^{EU}	237	18	0,13
17	[detect]	EN-GB	71	60	0,04
		EN-PT^{EU}	127	26	0,08
		EN-ES^{EU}	132	33	0,07
18	[determine]	EN-GB	139	27	0,08
		EN-PT^{EU}	124	30	0,08
		EN-ES^{EU}	220	20	0,12
19	[develop]	EN-GB	108	38	0,06
		EN-PT^{EU}	100	41	0,06
		EN-ES^{EU}	118	36	0,06

20	[establish]	EN-GB	64	71	0,04
		EN-PT^{EU}	56	78	0,03
		EN-ES^{EU}	100	50	0,05
21	[evaluate]	EN-GB	80	52	0,05
		EN-PT^{EU}	126	27	0,08
		EN-ES^{EU}	85	63	0,05
22	[exclude]	EN-GB	108	37	0,06
		EN-PT^{EU}	74	58	0,05
		EN-ES^{EU}	38	120	0,02
23	[explain]	EN-GB	87	51	0,05
		EN-PT^{EU}	122	31	0,07
		EN-ES^{EU}	168	25	0,09
24	[follow]	EN-GB	106	39	0,06
		EN-PT^{EU}	91	49	0,06
		EN-ES^{EU}	61	82	0,03
25	[include]	EN-GB	175	21	0,10
		EN-PT^{EU}	113	34	0,07
		EN-ES^{EU}	105	46	0,06
26	[increase]	EN-GB	170	23	0,10
		EN-PT^{EU}	199	15	0,12
		EN-ES^{EU}	242	17	0,13
27	[know]	EN-GB	136	29	0,08
		EN-PT^{EU}	132	25	0,08
		EN-ES^{EU}	198	23	0,11
28	[lead]	EN-GB	88	48	0,05
		EN-PT^{EU}	111	36	0,07
		EN-ES^{EU}	102	49	0,06
29	[maintain]	EN-GB	52	92	0,03
		EN-PT^{EU}	78	56	0,05
		EN-ES^{EU}	68	74	0,04
30	[need]	EN-GB	57	78	0,03
		EN-PT^{EU}	40	95	0,02
		EN-ES^{EU}	62	80	0,03
31	[obtain]	EN-GB	42	110	0,02
		EN-PT^{EU}	66	63	0,04
		EN-ES^{EU}	103	48	0,06
32	[occur]	EN-GB	94	44	0,05
		EN-PT^{EU}	124	28	0,08
		EN-ES^{EU}	70	72	0,04
33	[perform]	EN-GB	62	73	0,04
		EN-PT^{EU}	106	37	0,06
		EN-ES^{EU}	106	45	0,06
34	[play]	EN-GB	48	98	0,03
		EN-PT^{EU}	43	90	0,03
		EN-ES^{EU}	87	62	0,05
35	[present]	EN-GB	113	35	0,07
		EN-PT^{EU}	190	17	0,12
		EN-ES^{EU}	117	39	0,06

36	[produce]	EN-GB	31	142	0,02
		EN-PT^{EU}	59	71	0,04
		EN-ES^{EU}	113	42	0,06
37	[propose]	EN-GB	40	115	0,02
		EN-PT^{EU}	60	69	0,04
		EN-ES^{EU}	78	71	0,04
38	[reflect]	EN-GB	104	40	0,06
		EN-PT^{EU}	60	70	0,04
		EN-ES^{EU}	68	75	0,04
39	[remain]	EN-GB	89	47	0,05
		EN-PT^{EU}	86	52	0,05
		EN-ES^{EU}	60	83	0,03
40	[represent]	EN-GB	67	66	0,04
		EN-PT^{EU}	58	74	0,04
		EN-ES^{EU}	48	99	0,03
41	[require]	EN-GB	143	26	0,08
		EN-PT^{EU}	84	53	0,05
		EN-ES^{EU}	100	51	0,05
42	[result]	EN-GB	68	63	0,04
		EN-PT^{EU}	69	60	0,04
		EN-ES^{EU}	79	69	0,04
43	[reveal]	EN-GB	100	42	0,06
		EN-PT^{EU}	142	23	0,09
		EN-ES^{EU}	126	35	0,07
44	[study]	EN-GB	28	153	0,02
		EN-PT^{EU}	40	94	0,02
		EN-ES^{EU}	81	67	0,04
45	[suggest]	EN-GB	437	7	0,26
		EN-PT^{EU}	376	8	0,23
		EN-ES^{EU}	487	7	0,26
46	[support]	EN-GB	115	33	0,07
		EN-PT^{EU}	102	40	0,06
		EN-ES^{EU}	143	32	0,08
47	[use]	EN-GB	400	9	0,23
		EN-PT^{EU}	284	13	0,17
		EN-ES^{EU}	355	12	0,19

Table 98 – Verbs unlikely to function as NLID markers given their similar ranks and percentage in the corresponding corpus

A second group of verbs that is also not contemplated as likely to mark L1 influence of the Portuguese/Spanish authors writing in English comprises verbs of a terminological nature associated mostly with methods, techniques, test types, instruments, and others used in health research. Since these verbs are associated with terms, their higher or lower frequency in the corpora is not likely to be a consequence of the authors' choice but rather the result of them describing the methods used in the studies with field-specific terms. These verbs are shown below in Table 99.

N	Verb	Corpus	Occurrences in corpus	Rank	% in corpus
1	[associate]	EN-GB	112	36	0,07
		EN-PT^{EU}	193	16	0,12
		EN-ES^{EU}	165	26	0,09
2	[bind]	EN-GB	34	132	0,02
		EN-PT^{EU}	62	65	0,04
		EN-ES^{EU}	132	34	0,07
3	[control]	EN-GB	48	100	0,03
		EN-PT^{EU}	58	73	0,04
		EN-ES^{EU}	94	55	0,05
4	[correlate]	EN-GB	18	218	0,01
		EN-PT^{EU}	92	48	0,06
		EN-ES^{EU}	46	106	0,02
5	[decrease]	EN-GB	37	127	0,02
		EN-PT^{EU}	145	22	0,09
		EN-ES^{EU}	82	65	0,04
6	[express]	EN-GB	65	69	0,04
		EN-PT^{EU}	124	29	0,08
		EN-ES^{EU}	186	24	0,10
7	[improve]	EN-GB	113	34	0,07
		EN-PT^{EU}	68	61	0,04
		EN-ES^{EU}	78	70	0,04
8	[indicate]	EN-GB	166	24	0,10
		EN-PT^{EU}	94	45	0,06
		EN-ES^{EU}	329	13	0,18
9	[induce]	EN-GB	41	114	0,02
		EN-PT^{EU}	87	50	0,05
		EN-ES^{EU}	247	15	0,13
10	[modulate]	EN-GB	25	166	0,01
		EN-PT^{EU}	41	93	0,03
		EN-ES^{EU}	45	110	0,02
11	[predict]	EN-GB	53	87	0,03
		EN-PT^{EU}	47	83	0,03
		EN-ES^{EU}	22	169	0,01
12	[promote]	EN-GB	20	197	0,01
		EN-PT^{EU}	59	72	0,04
		EN-ES^{EU}	64	77	0,03
13	[provide]	EN-GB	159	25	0,09
		EN-PT^{EU}	67	62	0,04
		EN-ES^{EU}	84	64	0,05
14	[range]	EN-GB	46	103	0,03
		EN-PT^{EU}	62	66	0,04
		EN-ES^{EU}	40	114	0,02

15	[reach]	EN-GB	32	139	0,02
		EN-PT^{EU}	44	88	0,03
		EN-ES^{EU}	90	58	0,05
16	[receive]	EN-GB	185	20	0,11
		EN-PT^{EU}	74	59	0,05
		EN-ES^{EU}	116	40	0,06
17	[reduce]	EN-GB	231	17	0,13
		EN-PT^{EU}	178	18	0,11
		EN-ES^{EU}	400	10	0,22
18	[release]	EN-GB	3	640	0,00
		EN-PT^{EU}	106	38	0,06
		EN-ES^{EU}	52	90	0,03
19	[treat]	EN-GB	52	91	0,03
		EN-PT^{EU}	61	68	0,04
		EN-ES^{EU}	95	53	0,05

Table 99 – Verbs unlikely to function as NLID marker given their terminological nature mostly associated with research methods and techniques

Finally, as shown in Table 100 below, a third group of verbs is excluded from the list of possible candidates to mark L1 influence in the Portuguese/Spanish authors writing OSRAs in English. These are modal verbs (i.e., [can], [may], [shall], [will]), verbs that act as auxiliary verbs (i.e., [do], [have]), and verbs that may appear alone or in combination with adverbs (e.g. *out*) or particles (e.g. *up*) forming phrasal verbs with different meanings and therefore, different translations into Portuguese/Spanish (i.e., [give][make][take]). Albeit the numbers of occurrences, ranks, or percentages in the corresponding corpus are sometimes similar, these verbs are excluded based on the many complexities associated with their usage from the rhetorical point of view which would need another study in order to address them.

N	Verb	Corpus	Occurrences in corpus	Rank	% in corpus
1	[can]	EN-GB	398	10	0,23
		EN-PT^{EU}	344	9	0,21
		EN-ES^{EU}	601	6	0,33
2	[do]	EN-GB	533	5	0,31
		EN-PT^{EU}	384	7	0,23
		EN-ES^{EU}	629	5	0,34
3	[give]	EN-GB	135	30	0,08
		EN-PT^{EU}	82	54	0,05
		EN-ES^{EU}	112	43	0,06
4	[have]	EN-GB	1659	2	0,97
		EN-PT^{EU}	1194	2	0,73
		EN-ES^{EU}	1313	3	0,71

5	[make]	EN-GB EN-PT^{EU} EN-ES^{EU}	266 117 144	15 32 31	0,16 0,07 0,08
6	[may]	EN-GB EN-PT^{EU} EN-ES^{EU}	477 544 468	6 5 8	0,28 0,33 0,25
7	[shall]	EN-GB EN-PT^{EU} EN-ES^{EU}	72 96 105	58 44 47	0,04 0,06 0,06
8	[take]	EN-GB EN-PT^{EU} EN-ES^{EU}	252 150 228	16 20 19	0,15 0,09 0,12
9	[will]	EN-GB EN-PT^{EU} EN-ES^{EU}	305 116 154	12 33 29	0,18 0,07 0,08

Table 100 – Modal verbs excluded from the list of possible markers of L1 influence in OSRAs written in English by the L1 Portuguese/Spanish authors

After the verbs that are less plausible to function as NLID markers are excluded, two groups of verbs are analyzed.

The first group contains verbs found in the EN-GB corpus more frequently than in the EN-PT^{EU} and EN-ES^{EU} corpora, and therefore are analyzed to verify their potential to mark possible strategies of avoidance of use by the non-L1 authors. These verbs are shown in Table 101.

N	Verb	Corpus	Occurrences in corpus	Rank	% in corpus
1	[identify]	EN-GB EN-PT^{EU} EN-ES^{EU}	287 97 114	13 42 41	0,17 0,06 0,06
2	[report]	EN-GB EN-PT^{EU} EN-ES^{EU}	638 516 436	4 6 9	0,37 0,32 0,24
3	[see]	EN-GB EN-PT^{EU} EN-ES^{EU}	416 148 210	8 21 22	0,24 0,09 0,11

Table 101 – Verbs analyzed to verify their potential to mark possible strategies of avoidance of use by the non-L1 authors who are L1 users of PT-EU/ES-EI in the CoRA

The frequencies of the verbs with the lemmas [identify], [report], and [see] are compared to examine if there are significant differences between the groups. The fourth L1 effect of the unified framework (Jarvis 2010, 2000) is tested for both EN-PT^{EU} and EN-ES^{EU} OSRAs, stated as follows:

Effect of L1 Influence	**L1 influence** **EN-PT^{EU} question**	**L1 influence** **EN-ES^{EU} question**
IV) Intralingual contrast	Are the frequencies of the verb [identify]/[report]/[see] in the EN-PT^{EU}/ EN-ES^{EU} and the EN-GB corpora statistically significantly different?	

The data is not normally distributed and has some outliers. Therefore, the Mann-Whitney's test is used to assess for a mean difference between the groups. The level of significance used is *p* < .05. Six tests are performed. The numbers of occurrences are normalized by 100. Table 102 shows the results obtained.

	IV - Intralingual contrast (Mann-Whitney Test)	Effect of L1 influence IV found for the EN-PT^{EU}?	IV - Intralingual contrast (Mann-Whitney Test)	Effect of L1 influence IV found for the EN-ES^{EU}?
	corpora examined			
Lemma of the verb	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. EN-GB	
	***p* reference value < .05**			
[identify]	*Z* = -4.810 *p* = .001 *M* rank EN-GB = 67.45 *M* rank EN-PT^{EU} = 39.55	*yes*	*Z* = -2.178 *p* = .029 *M* rank EN-GB = 47.46 *M* rank EN-ES^{EU} = 35.61	*yes*
[report]	*Z* = .000 *p* = 1.000 *M* rank EN-GB = 60.00 *M* rank EN-PT^{EU} = 60.00	*no*	*Z* = -.368 *p* = .713 *M* rank EN-GB = 53.54 *M* rank EN-ES^{EU} = 51.38	*no*
[see]	*Z* = -4.988 *p* = .001 *M* rank EN-GB = 63.61 *M* rank EN-PT^{EU} = 35.39	*yes*	*Z* = -1.156 *p* = .248 *M* rank EN-GB = 42.85 *M* rank EN-ES^{EU} = 36.79	*no*

Table 102 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups in relation to the verbs [identify], [report], and [see] deemed as likely to mark strategies of avoidance

As can be seen in Table 102, the Mann Whitney's tests indicate statistically significant differences in the ranked frequencies of the verb [identify] between the L1 and both of the non-L1 English groups writing in English, with the L1 authors using the verb [identify] more frequently. Also, the verb [see] is significantly more frequently used by the L1 (i.e., EN-GB) than by the non-L1 authors who are Portuguese L1 users (EN-PT^{EU}) writing OSRAs in English. However, there are no significant differences in the frequency of the verb [see] between the EN-GB and the EN-ES^{EU} groups. Finally, the verb [report] is as frequently used by the L1 as by the non-L1 English authors writing OSRAs in English since no significant differences are found between the groups.

Based on the significance of the results, the verb [identify] is further examined for both groups (the EN-PT^{EU} and the EN-ES^{EU}). As described above, the examinations are based on the concordances of the parsed files. The concordances are obtained with WordSmith 7.0 (Scott 2018b), from which the syntactic tags containing the verb [identify] are extracted. Table 103 below shows the syntactic structures associated with the verb [identify] in the EN-GB corpus that most contribute to the significant differences in the number of occurrences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups.

Tags_[identify] with examples	**EN-GB**	**EN-PT^{EU}**	**EN-ES^{EU}**
1. V PCP2 PAS @ICL-AUX	**80**	**37**	**38**
E.g., “However, two small motifs N-terminal to the Nanognb homeodomain were *identified* as similar to motifs in Nanog proteins.”			
2. V IMPF @FS-STA	**65**	**14**	**21**
E.g., “We *identified* three missense mutations in our series.”			
3. V INF @ICL-	**25**	**16**	**10**
E.g., “Since 2013, investigation of MRSA bacteremia requires a locally administered postinfection review (PIR), which aims to *identify* how the case occurred and preventive actions to avoid recurrence.”			
4. V PCP2 PAS @ICL-N	**23**	**4**	**9**
E.g., “As the rectum and urethra are physiologically distinct from the nasopharynx, factors such as the polysaccharide capsule, a well-known meningococcal virulence determinant, may be important for colonisation and persistence (table 1), consistent with the prevalence of encapsulated strains *identified* here and previously.”			
5. V PCP2 AKT @ICL-AUX	**20**	**6**	**6**
E.g., “Furthermore, we have recently *identified* a p53-MYC dual hub responsible for many of the BCR/ABL-induced changes in CML.”			

6. V INF @ICL-AUX	**10**	**4**	**4**
E.g., “First, screening using acuity measurement would only *identify* amblyopia and refractive error.”			
7. V PCP1 @ICL-P	**5**	**2**	**6**
E.g., “Our findings reconfirm milk as the major EoE-related food in Spanish adult patients, after *identifying* EoE recrudescence after milk challenge in 50% of responder cases.”			
8. V INF @ICL-A	**8**	**2**	**-**
E.g., “It has been suggested that reduced decay of EHH of haplotypes that are both rare and extended is informative to *identify* signatures of natural selection.”			

Table 103 – Syntactic tags of [identify] that show the significant differences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups in the CoRA

After examination of the sentences containing the verb [identify] of the most frequent tags found in the English corpora of the CoRA, it can be said that the uses of [identify] are not of a terminological nature but may have a collocational character since it frequently appears associated with descriptions and discussion of findings related to genetics and biochemistry. It could be that the high frequency of that verb in the EN-GB corpus is just the result of having more OSRAs dealing with the fields of genetics, biochemistry, or related methods and techniques of those fields.

In the case of the verb [see], Table 104 shows the most frequent syntactic tags that explain where the differences reside between the EN-GB and the EN-PT^{EU} authors. As can be seen, many uses of [see] concern the indication to the reader of the availability of further information on the topic that is being discussed (i.e., 1 and 5). In contrast, others are general uses that could actually be substituted with synonyms such as [observe]/[identify]/[verify] and therefore may be indicative of choice. Since the non-L1 English authors who are L1 Portuguese do not resort to this verb as frequently as the L1 English authors, it can be argued that its much less frequency in the non-L1 English OSRAs may signal that the person is not an L1 English user, despite being an advanced user of the language. However, it does not mark L1 influence.

Tags_[see] with Examples	EN-GB	EN-PT^{EU}
1. <v.contact> V IMP @FS-	**153**	**72**
E.g., "Examining associations in the stress aware and unaware groups separately did not significantly influence these results (*see* Supplemental materials)."		
2. <v.contact> V PCP2 PAS @ICL-AUX	**94**	**28**
E.g., "The protein carbonyl content of sciatic nerves was analysed by western blotting as a marker of oxidative damage, but no significant differences were *seen* between the intensity of bands obtained from nerves of adult and old mice at rest (Fig. 2C)."		
3. <v.contact> V PCP2 PAS @ICL-N	**52**	**10**
E.g., "There was little improvement in physical limitation or treatment satisfaction, perhaps reflecting the mild physical limitation and excellent treatment satisfaction *seen* at baseline."		
4. <v.contact> V INF @ICL-	**23**	**6**
E.g., "However, the pilot study indicated that for more general symptoms, participants were much more willing to *see* GPs suggesting PLWHIVs preferences for using HIV clinic."		
5. <v.contact> V IMP @FS-COM	**18**	**6**
E.g., "There were no effects of stress awareness on AAAQ scores or SSRT (*see* Supplemental online materials for details)"		

Table 104 – Most frequent syntactic structures of the verb [see] marking the significant difference between the EN-GB and the EN-PTEU groups in the CoRA

After examining the verbs [identify], [report], and [see], the remaining verbs are examined, taking into account that their (a) occurrences, (b) ranks, and (c) percentages of occurrences are higher in both or one of the non-L1 English corpora (EN-PT^{EU} and EN-ES^{EU}) than in the L1 (EN-GB) corpus and therefore, are likely to function as NLID markers of language transfer if similar occurrences, ranks, and percentages are verified in the L1 PT-EU and ES-EU corpora. Since these verbs are chosen as possible markers of language transfer, their equivalents in the L1 Portuguese/Spanish corpora are extracted, and their frequencies are compared. Table 105 below shows the verbs that after comparison are not further analyzed because the frequencies of their equivalents in the corresponding L1 corpora do not justify the high frequencies found in the non-L1 English corpora (one or both), and therefore, the statistical comparison to test for an L1 effect cannot be performed.

N	Verbs	Corpus	Occurrences in corpus	Rank	% in corpus
1	[considerar]	PT-EU	160	15	0,11
	[considerar]	ES-EU	150	17	0,09
	[consider]	EN-GB	188	18	0,11
		EN-PT^{EU}	335	11	0,20
		EN-ES^{EU}	218	21	0,12
2	[envolver]	PT-EU	44	61	0,03
	[implicar]	ES-EU	34	92	0,02
	[involve]	EN-GB	102	41	0,06
		EN-PT^{EU}	176	19	0,11
		EN-ES^{EU}	244	16	0,13
3	[observar]	PT-EU	119	23	0,08
	[observar]	ES-EU	298	8	0,18
	[observe]	EN-GB	315	11	0,18
		EN-PT^{EU}	570	4	0,35
		EN-ES^{EU}	698	4	0,38
4	[mostrar]	PT-EU	62	45	0,04
	[mostrar]	ES-EU	312	7	0,19
	[show]	EN-GB	763	3	0,45
		EN-PT^{EU}	934	3	0,57
		EN-ES^{EU}	1409	2	0,76

Table 105 – Verbs that after analysis are excluded as NLID markers of language transfer

Therefore, three verbs, i.e., [find], [relate], [seem] are deemed likely to mark language transfer effects in OSRAs written in English by the L1 Portuguese/Spanish authors since the frequencies of their equivalents in the corresponding L1 corpora (i.e., the PT-EU and the ES-EU) are equal or higher than the frequencies of [find], [relate], [seem] in the L1 English corpus. Table 106 shows these verbs.

N	Verb	Corpus	Occurrences in corpus	Rank	% in corpus
1	[encontrar]	PT-EU	212	8	0,15
	[encontrar]	ES-EU	241	10	0,15
	[find]	EN-GB	170	22	0,10
		EN-PT^{EU}	344	10	0,21
		EN-ES^{EU}	384	11	0,21
2	[relacionar]	PT-EU	93	27	0,06
	[relacionar]	ES-EU	164	13	0,10
	[relate]	EN-GB	56	81	0,03
		EN-PT^{EU}	104	39	0,06
		EN-ES^{EU}	154	28	0,08
3	[parecer]	PT-EU	70	39	0,05
	[parecer]	ES-EU	65	44	0,04
	[seem]	EN-GB	21	190	0,01
		EN-PT^{EU}	97	43	0,06
		EN-ES^{EU}	87	61	0,05

Table 106 – Verbs analyzed to test for L1 transfer effects

The frequencies of the verbs [find], [relate], [seem], and equivalents [encontrar]/[encontrar], [relacionar]/[relacionar] and [parecer]/[parecer], are compared to examine significant differences between the groups. For all verbs, the following questions are asked for both the EN-PT^{EU} and the EN-ES^{EU} corpora.

Effect of L1 Influence	**L1 influence** **EN-PT^{EU} questions**	**L1 influence** **EN-ES^{EU} questions**
I) Intragroup homogeneity	Are the frequencies of the verb [find]/[relate]/[seem] in the EN-PT^{EU} / EN-ES^{EU} OSRAs uniformly distributed?	
II) Intergroup heterogeneity	Are the frequencies of the verb [find]/[relate]/[seem] in the EN-PT^{EU} and EN-ES^{EU} OSRAs statistically significantly different?	
III) Cross-language congruity	Are the frequencies of the verb [find]/[relate]/[seem] in the EN-PT^{EU} OSRAs and the equivalent [encontrar]/[relacionar]/[parecer] in the PT-EU OSRAs statistically similar?	Are the frequencies of the verb [find]/[relate]/[seem] in the EN-ES^{EU} OSRAs and the equivalent [encontrar]/[relacionar]/[parecer] in the ES-EU OSRAs statistically similar?
IV) Intralingual contrast	Are the frequencies of the verb [find]/[relate]/[seem] in the EN-PT^{EU} and the EN-GB OSRAs statistically significantly different?	Are the frequencies of the verb [find]/[relate]/[seem] in the EN-ES^{EU} and the EN-GB OSRAs statistically significantly different?

The Mann-Whitney's test is used to assess for a mean difference between the groups given that the data is not normally distributed and has some outliers. Questions in relation to the effects I and II are answered together. The level of significance used for questions in relation to the effects II and IV is $p < .05$. Because questions in relation to the effects I and III look for uniformity and congruity, respectively, a result of $p > .05$ is associated with an L1 effect. Table 107 below shows the results and mean ranks of all comparisons.

	L1 Influence Effects								
	I) Intra-L1 homogeneity (Levene's test)	II) Inter-L1 heterogeneity (Mann-Whitney Test)	EN-PT^{EU} and EN-ES^{EU} similar in variance but different in means?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-PT^{EU}?	III) Cross-language congruity (Mann-Whitney Test)	IV) Intralingual contrast (Mann-Whitney Test)	At least two effects of L1 influence found for the EN-ES^{EU}?
	Corpora Examined			**Corpora Examined**			**Corpora Examined**		
Verb	EN-PT^{EU} vs. EN-ES^{EU}	EN-PT^{EU} vs. EN-ES^{EU}		EN-PT^{EU} vs. PT-EU	EN-PT^{EU} vs. EN-GB		EN-ES^{EU} vs. ES-EU	EN-ES^{EU} vs. EN-GB	
	Reference *p* values			**Reference *p* values**			**Reference *p* values**		
	$p > .05$ **AND**	$p < .05$?		$p > .05$?	$p < .05$?		$p > .05$?	$p < .05$?	
[encontrar] [encontrar] [find]	*F* = 2.827 *p* = .095	*Z* = -.275 *p* = .783 Mean ranks: EN-PT^{EU} = 60.62 EN-ES^{EU} = 62.38	*no*	*Z* = -2.357 *p* = .018 Mean ranks: EN-PT^{EU} = 62.46 PT-EU = 48.12	*Z* = -3.805 *p* = .001 Mean ranks: EN-PT^{EU} = 68.34 EN-GB = 45.03	*no*	*Z* = -1.516 *p* = .129 Mean ranks: EN-ES^{EU} = 58.48 ES-EU = 49.33	*Z* = -3.700 *p* = .001 Mean ranks: EN-ES^{EU} = 68.03 EN-GB = 48.38	*yes*
[relacionar] [relacionar] [relate]	*F* = 6.103 *p* = .015	*Z* = -1.709 *p* = .088 Mean ranks: EN-PT^{EU} = 44.88 EN-ES^{EU} = 54.12	*no*	*Z* = -.102 *p* = .919 Mean ranks: EN-PT^{EU} = 45.73 PT-EU = 45.22	*Z* = -.363 *p* = .717 Mean ranks: EN-PT^{EU} = 41.18 EN-GB = 39.42	*no*	*Z* = -.990 *p* = .322 Mean ranks: EN-ES^{EU}= 44.91 ES-EU= 50.32	*Z* = -1.889 *p* = .059 Mean ranks: EN-ES^{EU} = 44.18 EN-GB = 34.68	*no*
[parecer] [parecer] [seem]	*F* = .527 *p* = .470	*Z* = -.203 *p* = .839 Mean ranks: EN-PT^{EU} = 44.03 EN-ES^{EU} = 45.06	*no*	*Z* = -1.494 *p* = .135 Mean ranks: EN-PT^{EU} = 36.20 PT-EU = 43.64	*Z* = -.421 *p* = .674 Mean ranks: EN-PT^{EU} = 30.94 EN-GB = 28.75	*no*	*Z* = -.957 *p* = .339 Mean ranks: EN-ES^{EU} = 33.14 ES-EU = 37.57	*Z* = -.562 *p* = .574 Mean ranks: EN-ES^{EU} = 27.10 EN-GB = 24.50	*no*

Table 107 – Results of the Mann-Whitney's tests performed to assess for mean differences between the groups in relation to the frequency of the verbs [find], [relate], [seem] and equivalents [encontrar]/[encontrar], [relacionar]/[relacionar], and [parecer]/[parecer] in Portuguese and Spanish

According to the results obtained, no overall effects of L1 influence can be associated with OSRAs produced by the Portuguese authors writing in English since only one L1 effect, i.e., intralingual contrast is found between the L1 English authors (EN-GB) and the Portuguese authors writing OSRAs in English (EN-PT^{EU}), and such effect concerns only one verb, i.e., [find]. No cross-language congruity is found between the Portuguese authors writing OSRAs in English (EN-PT^{EU}) and the Portuguese authors writing OSRAs in their L1. No other significant statistical differences exist between the L1 English authors (EN-GB) and the Portuguese authors writing OSRAs in English (EN-PT^{EU}) in relation to the verbs [relate] and [seem]. Likewise, the L1 English authors (EN-GB) and the Spanish authors writing OSRAs in English (EN-ES^{EU}) do not differ significantly in relation to the frequency of use of the verbs [relate] and [seem]. However, these authors, i.e. the EN-GB and the EN-ES^{EU}, differ significantly in relation to the frequency of the verb [find], and at the same time, the EN-ES^{EU} authors do not differ from the Spanish authors writing OSRAs in their L1.

Although both the Portuguese and the Spanish authors differ from the L1 English authors writing OSRAs in English, an overall L1 effect can only be argued for OSRAs written in English by the Spanish authors. Nonetheless, based on the significance of the results, the verb [find] and its equivalents in the Portuguese and Spanish corpora [encontrar] and [encontrar] are further analyzed for both groups (the EN-PT^{EU} and the EN-ES^{EU}). The analyses are based on the concordances of the parsed files. The concordances are obtained with WordSmith 7.0 (Scott 2018b), from which the syntactic structures are extracted. The significant differences in the number of occurrences of the verb [find] between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups are more evident in the syntactic structures shown in Table 108, which are more frequent in the EN-PT^{EU}/EN-ES^{EU} groups than in the EN-GB group.

N	Tags of [find]	Example	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1	V PCP2 PAS @ICL-AUX	*"The maximum inhibition was not **found** when L-732,138 (100 mM) was added…"*	64	160	155
2	V IMPF @FS-STA	*"We **found** marked regional down-regulations of the main glutaminase…"*	58	92	90
3	V PCP2 PAS @ICL-N	*"In all of those studies the positive predictive value of the test was much higher than the value **found** here…"*	11	41	50

4	V PCP2 AKT @ICL-AUX	*"…a recent report has analyzed transcriptome profiling of purified human and mouse astrocytes, and they have **found** species-related differences for several genes…"*	6	5	25
5	V INF @ICL-AUX	*"Nevertheless, we did not **find** changes in mRNA levels of these genes in our group of ICM patients."*	1	9	19
6	V IMPF @FS-N	*"Most factors were at borderline significance or had low statistical power and are difficult to interpret, so we will discuss those we **found** more clinically relevant or interpretable."*	2	6	8

Table 108 – Syntactic Structures showing the significant differences between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} groups in relation to the verb [find]

The distributions of the most frequent syntactic structures of equivalents of the verb [find] in the Portuguese and Spanish corpora of the CoRA, i.e., [encontrar] and [encontrar] are shown below in Table 109.

N	Tags of [encontrar]/[encontrar]		ES-EU	PT-EU
1	V PR 3P IND VFIN @FMV	*"Otros síndromes clínicos variados en los gatos de este estudio no se relacionaron con la presencia del ADN de la bacteria en la sangre o en la boca, al igual que otros autores que tampoco **encuentran** resultados concluyentes"*	30	28
2	V PR 3S IND VFIN @FMV	*"Para medir el EPC se han desarrollados distintos instrumentos entre los que destaca el Practice Environment Scale Nursing Work Index (PES-NWI) por su solidez metodológica y que se **encuentra** validado para el entorno español en AP"*	16	26
3	V PR/PS 1P IND VFIN @FMV	*"**Encontramos** además asociación entre AV menores de 3 meses de vida y flujos AV menores de 500 ml/min."*	40	-
4	V PS 3S IND VFIN @FMV	*"En el anexo 1 se muestra el contenido completo de la guía a excepción de 4*	25	15

		asuntos no incluidos en el trabajo, dado que no se ***encontró*** *evidencia en la literatura analizada en relación a...”*		
5	V PCP M S @IMV @#ICL-AUX<	*“un estudio aleatorizado no ha* ***encontrado*** *diferencias significativas en el resultado final entre pacientes monitorizados y no monitorizados”*	30	5
6	V PS 3P IND VFIN @FMV	*“Además, estos miRNAs se* ***encontraron*** *sobreexpresados en las muestras osteoporóticas del array…”*	26	-
7	V IMPF 3P IND VFIN @FMV	*“En cambio, los pacientes tratados con FL en todos los intervalos de distancia, excepto en uno, las medianas se* ***encontraban*** *por debajo de los 30 Minutos…”*	11	9
8	V PCP F S @IMV @#ICL-AUX<	*“Também não se* ***encontrou*** *qualquer associação entre macrossomia e síndrome de aspiração meconial”*	-	17
9	V INF @IMV @#ICL-P<	*“El presente estudio se diseñó para determinar la probabilidad de* ***encontrar*** *lesiones coronarias significativas…”*	13	2
10	V PCP F S @IMV @#ICL-N<	*“A diferença* ***encontrada*** *em relação à idade (p<0,001) justifica-se pela diferente fase de formação.”*	-	15
11	V PS/MQP 3P IND VFIN @FMV	*“Vários estudos* ***encontraram*** *valores populacionais médios de TSH próximos do limite inferior do intervalo considerado normal…”*	-	15
12	V PCP F P @IMV @#ICL-AUX<	*“Por fim, mais uma vez, foram* ***encontradas*** *taxas muito baixas de hipocoagulação oral, tal como foram encontradas noutros estudos”*	-	10

Table 109 – Distribution of the syntactic structures of the verb [find] and its equivalents [encontrar] and [encontrar] in the CoRA

No further obvious distinctive aspects can be found between the syntactic structures of the verb [find] and its equivalents in Portuguese and Spanish [encontrar] and [encontrar]. Figure 17 below shows the distributions of the verb [find] in the three English corpora and the verbs [encontrar] and [encontrar] in the Portuguese and Spanish corpora, respectively, of the CoRA.

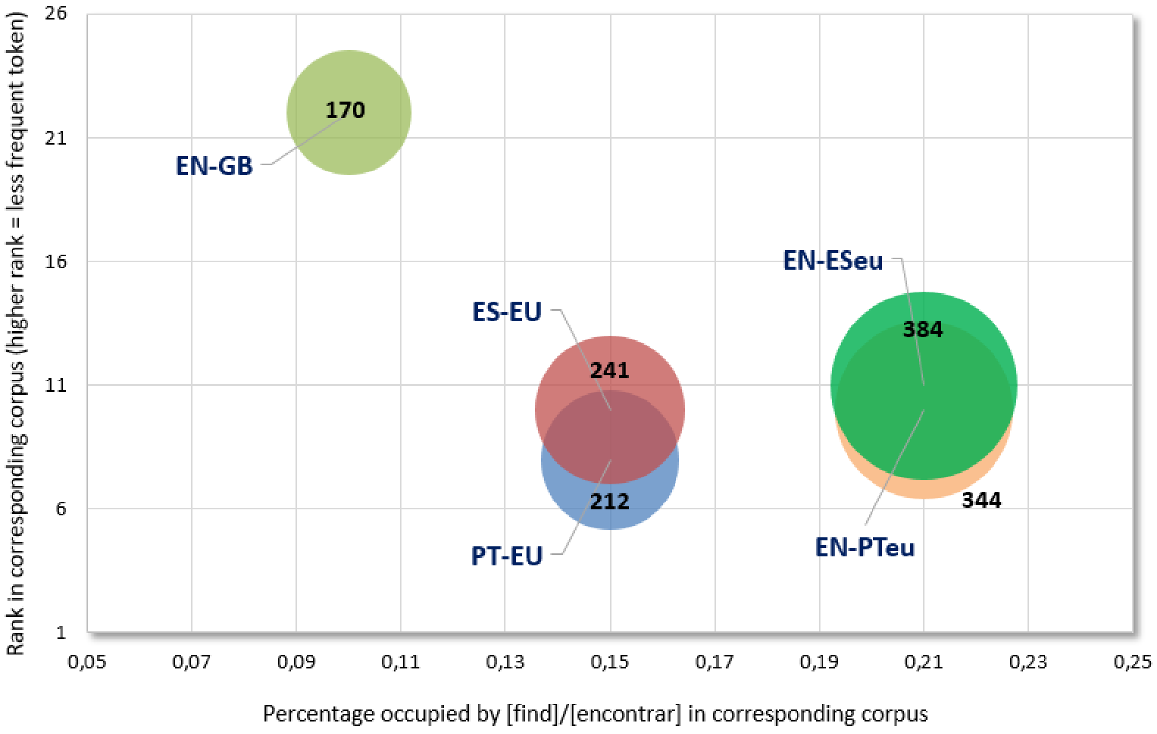


Figure 17 – Distributions of the verbs [find]/ [encontrar]/[encontrar] in the CoRA.

The data shows that overall the Portuguese and the Spanish authors writing OSRAs in English have a preference for the verb [find] since the distribution of this verb in the EN-PT^{EU} and the EN-ES^{EU} corpora are more similar to the distribution of the verbs [encontrar] and [encontrar] in the PT-EU and the ES-EU corpora than the distribution of [find] in the L1 English corpus.

4.6. Discussion of the findings

Investigating the patterns of linguistic variation in original scientific research articles (OSRAs) written in English by L1 and non-L1 authors was the purpose that guided this study. With that purpose in mind, one main research question was asked in relation to 1) the existence of variables associated with L1 influence in OSRAs written in English by L1 authors of European Portuguese (PT-EU) and L1 authors of European Spanish (ES-EU) in the field of health sciences, and what those variables are. Two additional research questions were also posed in relation to 2) the possible explanations and, 3) the possible implications of the absence/presence of the referred variables. The main purpose of the study implied going beyond intuition and testing different variables to find their potential to mark the influence of the authors' L1 in OSRAs they wrote in English and identifying linguistic variables that can be used as markers of authorship influenced by the L1.

The review of the literature on L1 influence indicated that the influence of the L1 is likely to happen even in advanced users of a foreign language, which is the case of the non-L1 authors in the CoRA who are L1 Portuguese/Spanish users. Therefore, the empirical work was undertaken based on the assumption that there were going to be results indicating the influence of the L1 in the Portuguese/Spanish authors of the CoRA writing OSRAs in English. Nineteen variables were studied, first by comparing their frequencies in the five corpora within the CoRA, and then by examining parts-of-speech, looking for specific words or word combinations that can function as markers of the influence of the Portuguese and Spanish authors L1s' when writing OSRAs in English.

The results obtained from the empirical research show that there are content-independent and content-dependent variables that can indicate the influence of the Portuguese and Spanish authors' L1 in the OSRAs they produced in English. That is, there are variables that can be associated with L1 transfer. Moreover, there are also content-independent and content-dependent variables associated with possible strategies of avoidance of use by the non-L1 English authors who are L1 users of European Portuguese/Spanish, i.e., the EN-PT^{EU} and the EN-ES^{EU}.

The content-independent variables considered the frequencies of the nineteen variables in the OSRAs within the CoRA. Overall, for 52.63% of the variables, the frequency was found to play no role or a weak role in possible effects of the Portuguese/Spanish authors' L1 in the OSRAs they wrote in English (Table 110 below, groups 1 and 2), whereas 26.32% of the variables were associated with moderate (Table 110, group 3) and 21.05% with strong (Table 110, group 4) possible effects of the influence of these authors' L1 in the OSRAs they wrote in English.

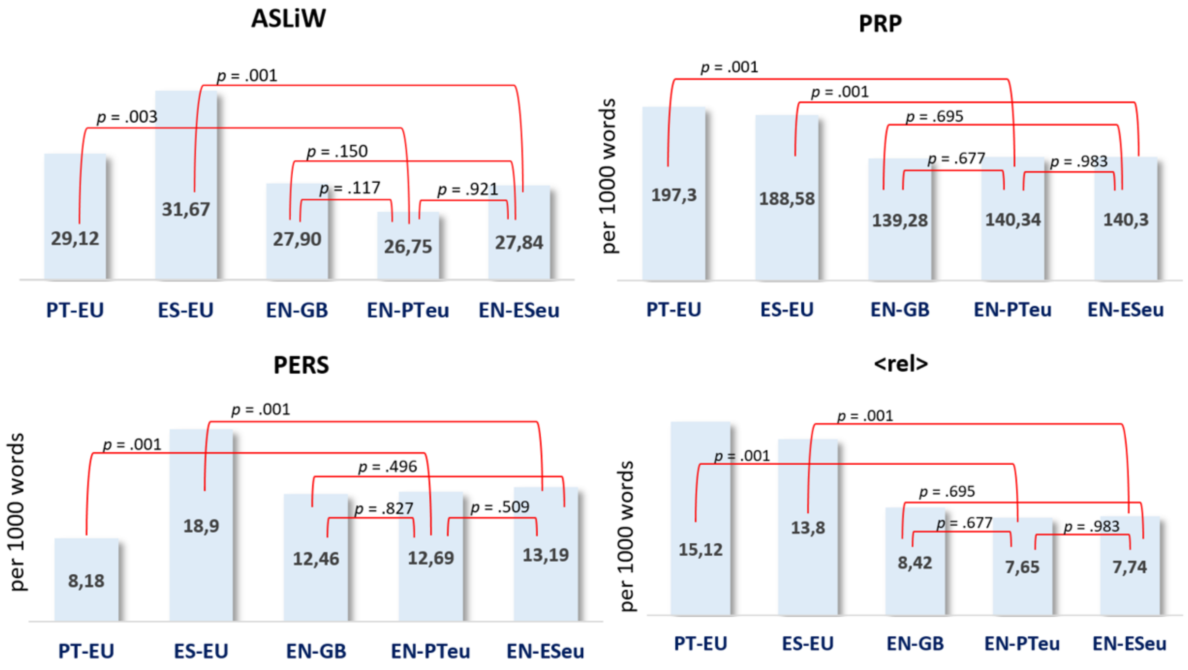
N	Finding	Variables
1.	No effects of L1 influence	**V2**: average sentence length in words **V12**: number of prepositions **V14**: number of relative pronouns **V15**: number of personal pronouns
2.	One effect of L1 influence – Cross-Language congruity	**V3**: number of paragraphs **V4**: standardized type/token ratio (STTR) **V5**: number of 1 to 5-letter words **V6**: number of 6-to-10-letter words **V10**: number of coordinating conjunctions **V11**: number of subordinating conjunctions
3.	One effect of L1 influence – Intralingual Contrast	**V1**: number of commas **V7**: number of 11-to-15-letter words **V8**: number of definite articles **V17**: number of adverbs **V18**: number of nouns
4.	Two effects of L1 influence	**V9**: number of indefinite articles **V13**: number of demonstrative pronouns **V16**: number of adjectives **V19**: number of verbs

Table 110 – Summary of findings according to the effects of L1 influence in relation to the content-independent variables, i.e., frequencies

As shown in Table 110 above, four variables (group 1) did not reveal any potential to function as possible markers of L1 influence in OSRAs written in English by the Portuguese/Spanish authors in the CoRA in relation to their frequency since no effects of L1 influence were found. Briefly, there are no significant difference between the groups of authors writing in English, whether they are L1 or non-L1 users of the language. However, there are significant differences between the L1 Portuguese authors writing OSRAs in Portuguese and the L1 Portuguese authors writing OSRAs in English. Likewise, there are significant differences between the L1 Spanish authors writing OSRAs in Spanish and the L1 Spanish authors writing OSRAs in English. An additional comparison was made of the L1

Portuguese authors writing in Portuguese (PT-EU) and the L1 Spanish authors writing in Spanish (ES-EU) that is not part of the L1 transfer model used for the comparisons of the corpora. This additional comparison and the comparisons carried out according to the model showed that three of the variables with no effects of L1 influence associated (**V2**: average sentence length in words, **V12**: number of prepositions, and **V15**: number of personal pronouns) behave significantly differently in the three main language groups, i.e., the L1 Portuguese authors writing OSRAs in Portuguese, the L1 Spanish authors writing OSRAs in Spanish, and both the L1 and the non-L1 (who are Portuguese or Spanish) English authors writing in English. The mean frequencies of the L1 groups (PT-EU, ES-EU, and EN-GB) in relation to the variables mentioned above are significantly different, while all groups writing in English (EN-GB, EN-PT^{EU}, and EN-ES^{EU}) are similar. Only the relative pronouns (V14) behave slightly differently.

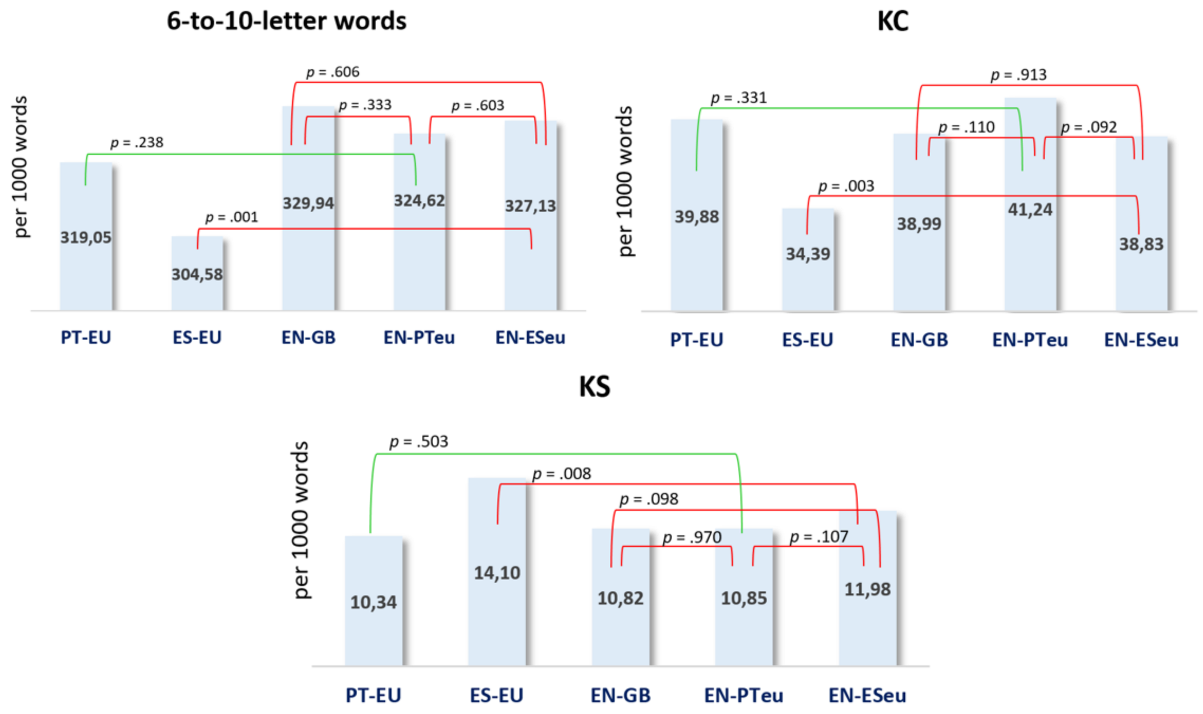
Within this first group in Table 110, the number of relative pronouns (V14) is also very similar between all groups writing OSRAs in English, whether they are L1 or the non-L1 English authors. As with in the other three variables, the number of relative pronouns (V14) is also significantly different between the Portuguese authors writing OSRAs in their L1 (PT-EU) and the Portuguese authors writing OSRAs in English; and between the Spanish authors writing OSRAs in their L1 (ES-EU) and the Spanish authors writing OSRAs in English. However, the frequency of V14 is not significantly different between PT-EU and the ES-EU authors writing OSRAs in their respective L1s. Figure 18 below summarizes the findings obtained for the variables in group 1 in Table 110. The additional comparison between the PT-EU and the ES-EU is not depicted in the figure because it is not part of the L1 transfer model used to compare the OSRAs within the CoRA.



For all results *p* value < .05 = significant differences. The red lines refer to results that do not indicate the presence of an effect of L1 influence. Legend: **V2**: Average sentence length in words – ASLiW; **V12**: Prepositions – PRP; **V14**: Relative pronouns - <rel>; **V15**: Personal pronouns – PRP.

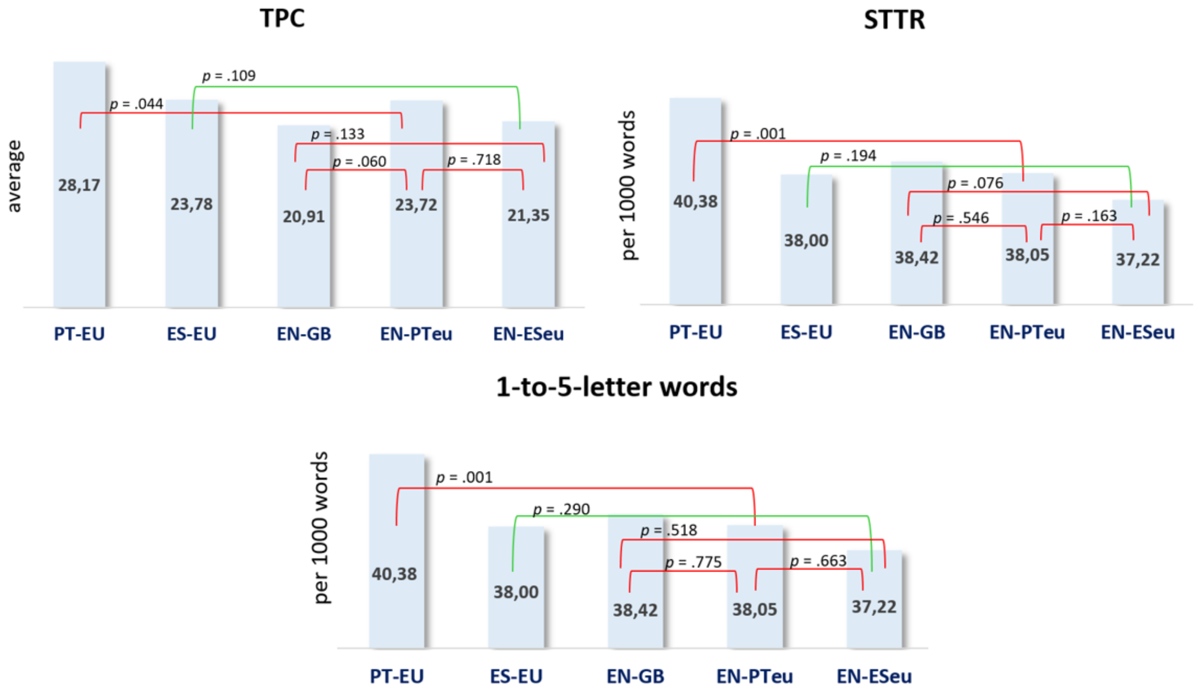
Figure 18 – Variables with no associated effects of L1 influence in relation to their frequencies in the CoRA

For the six variables in group 2 in Table 110, only one effect of L1 influence was found. This effect is associated with cross-language congruity, i.e., there are no significant differences between the Portuguese/Spanish authors writing in their respective L1 and the Portuguese/Spanish authors writing in English in relation to the frequencies of the variables tested. However, all the groups of OSRAs written in English are still not significantly different from each other in relation to the frequency of those variables. As can be seen in Figure 19 below, for the Portuguese group, the variables with only one effect of L1 influence, i.e., cross-language congruity, in relation to the frequency are **V6** (number of 6-to-10-letter words), **V10** (number of coordinating conjunctions), and **V11** (number of subordinating conjunctions). For the Spanish group, in Figure 20 below, the variables with only one effect of L1 influence (cross-language congruity) in relation to the frequency are **V3** (number of paragraphs), **V4** (type/token ratio), and **V5** (number of 1 to 5-letter words).



For all results *p* value < .05 = significant differences. The red lines refer to results that do not indicate an effect of L1 influence. On the contrary, the green lines indicate an effect of L1 influence. Legend: **V6**: Number of 6-to-10-letter words; **V10**: Number of coordinating conjunctions – KC; **V11:** Number of subordinating conjunctions – KS.

Figure 19 – Variables with one effect of L1 influence – Cross-language congruity in the EN-PT^{EU} OSRAs



For all results *p* value < .05 = significant differences. The red lines refer to results that do not indicate an effect of L1 influence. On the contrary, the green lines indicate an effect of L1 influence. Legend: **V3**: number of paragraphs – TPC; **V4**: Standardized type/token ratio – STTR; **V5:** Number of 1-to-5-letter words.

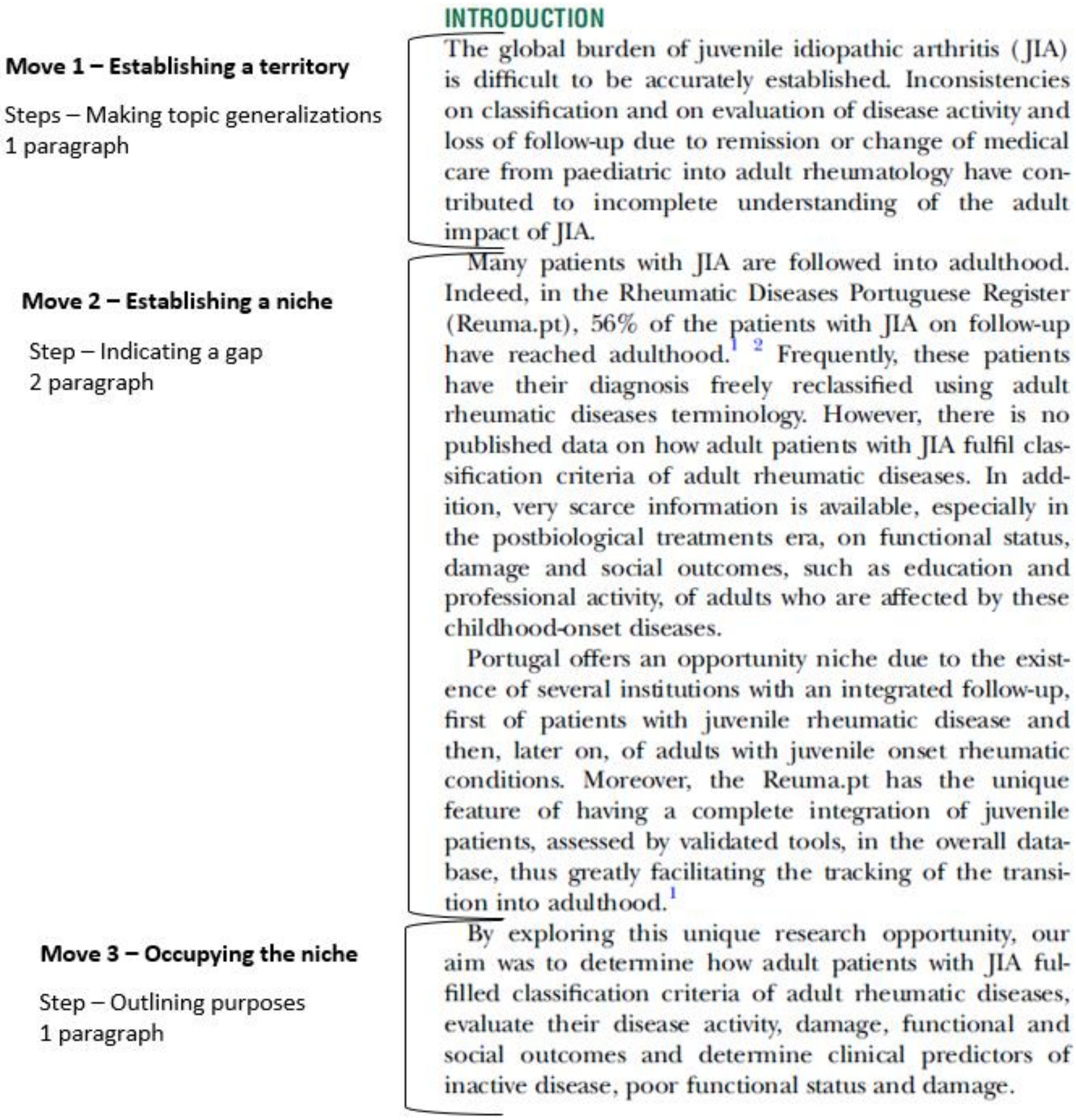
Figure 20 – Variables with only one effect of L1 influence – Cross-language congruity in the EN-ES^{EU} OSRAs

Since no other effects are found in relation to the frequencies of the six variables of group 2, Table 110, these six single effects are considered very weak indicators of the possible influence of the Portuguese or Spanish authors' L1 when they write OSRAs in English.

The 52.63% of null or weak effects associated with the frequencies of these variables can be explained by the constraints imposed by the scientific genre and the register. Since all the texts in the CoRA are OSRAs from the health sciences published within a relatively short time span, one might expect the homogeneity of certain variables due to the usual recommended clarity of scientific writing (Day, Sakaduski, and Day 2011) is expected.

In relation to the ASLiW (**V2**), in OSRAs of experimental physics in English, the sentence length has been reported "to remain fairly stable", from the 19th century to the 1980's, with 27.6 to 23.7 words per sentence on average (Bazerman 1984: 175). A decrease in sentence length has been previously described, for example, for genres in the Corpus of Historical American English (COHA), in relation to magazines and newspapers, which used to have average sentence lengths of 26.02 and 21.57 and by 2000 contained sentences with 17.14 and 16.70, respectively (Rudnicka 2018). However, in scientific writing in general, the common average sentence length has been reported to be approximately 20 to 28 words per sentence (Piqué-Angordans and Aguilar 1999). The results obtained in this study are in line with what is expected from scientific authors according to the recommendations of the style guidelines of scientific journals, which is to write about 25 words per sentence maximum (Iskander et al. 2018). The OSRAs in the CoRA follow the same pattern. Though not surprising, the results related to the ASLiW are different from what was intuitively expected based on my professional experience, which was that the non-L1 English authors who are L1 Portuguese/Spanish would produce significantly longer sentences.

In relation to V3 (number of paragraphs) the lack of effects of L1 influence may be justified by the rhetorical requirements of OSRAs. Research articles, and specifically OSRAs, follow the format IMRAD, in the *Introduction*, *Methodology*, *Results* and *Discussion* sections (with versions considering the section *Conclusions*) and within each section authors are expected to make specific rhetorical moves, which in turn consist of steps, as part of the communicative function of the text (Swales 2004; Swales 1990; Moreno and Swales 2018). In the health sciences, this format and rhetorical organization are followed, regardless of the language of publication of the OSRA. Therefore, paragraph division in OSRAs is likely to coincide with rhetorical moves, which could explain the homogeneity of this variable in the CoRA. Nonetheless, the variable was assessed because there are no specific prescriptions about the number of paragraphs that the authors can use and therefore, it was considered that the paragraph division could work as a possible style variable. Figure 21 below shows an example of a paragraph division according to the rhetorical moves and steps usually found in OSRAs in the health sciences field.



Move 1 – Establishing a territory

Steps – Making topic generalizations
1 paragraph

Move 2 – Establishing a niche

Step – Indicating a gap
2 paragraph

Move 3 – Occupying the niche

Step – Outlining purposes
1 paragraph

INTRODUCTION

The global burden of juvenile idiopathic arthritis (JIA) is difficult to be accurately established. Inconsistencies on classification and on evaluation of disease activity and loss of follow-up due to remission or change of medical care from paediatric into adult rheumatology have contributed to incomplete understanding of the adult impact of JIA.

Many patients with JIA are followed into adulthood. Indeed, in the Rheumatic Diseases Portuguese Register (Reuma.pt), 56% of the patients with JIA on follow-up have reached adulthood.[1] [2] Frequently, these patients have their diagnosis freely reclassified using adult rheumatic diseases terminology. However, there is no published data on how adult patients with JIA fulfil classification criteria of adult rheumatic diseases. In addition, very scarce information is available, especially in the postbiological treatments era, on functional status, damage and social outcomes, such as education and professional activity, of adults who are affected by these childhood-onset diseases.

Portugal offers an opportunity niche due to the existence of several institutions with an integrated follow-up, first of patients with juvenile rheumatic disease and then, later on, of adults with juvenile onset rheumatic conditions. Moreover, the Reuma.pt has the unique feature of having a complete integration of juvenile patients, assessed by validated tools, in the overall database, thus greatly facilitating the tracking of the transition into adulthood.[1]

By exploring this unique research opportunity, our aim was to determine how adult patients with JIA fulfilled classification criteria of adult rheumatic diseases, evaluate their disease activity, damage, functional and social outcomes and determine clinical predictors of inactive disease, poor functional status and damage.

Figure 21 – An example of paragraph division following rhetorical moves of research articles, based on the introduction of the OSRA EN-PTeu_OSRA_016 of the CoRA.(Swales 1990: 141)

In relation to **V4**, as was explained in the methodological chapter, the standardized type/token ratio (STTR) was calculated every 1000 tokens. This variable was assessed as an attempt to verify possible differences between the groups of OSRAs in relation to the lexical density. The results obtained are expected and in agreement with the controlled conditions of the samples. That is, all the texts in the CoRA are of the same genre, and therefore follow the same structure. Also, all the texts are of very similar lengths in words. Lastly, all texts discuss health topics which is why the vocabulary is similar in all texts.

The weak results obtained for **V5** (number of 1-to-5-letter words) and **V6** (number of 6-to-10-letter words) in relation to their potential to mark L1 influence can be interpreted by looking at the words that comprise these two groups.

The **V5** contains mostly function words in the higher ranks of the lists, i.e., the most frequent words, which make up at least 30% of these words. Other POS such as nouns (data, cell, renal, value, acids, heart, drugs, state, class, point, ratio, spine, gland), verbal forms (shows, avoid, carry, plays, act, noted, occur, must, exert), adjectives (obese, male, whole, wide, broad, false, usual, every, such), acronyms (RNA, VAT, PCR, MDA, P1, RR, TNF, RCC, GLS), and adverbs (never, often, ever) are also within this group.

The **V6** (number of 6-to-10-letter words) also contains some function words such as conjunctions (however, although, whether, because) but the group comprises mostly nouns (activity, tissue, receptors, conditions, amygdala, diagnosis, surgery, adherence, apoptosis, background) formal verbs (observed, associated, induced, reported, compared, treated, involved, described, expressed, performed, decreased, identified, suggesting), adjectives (different, specific, present, healthy, genetic, negative, recent, metabolic), and adverbs (finally, highly, recently, especially, strongly, probably, currently, usually, clearly, completely, partially, directly, slightly).

Together these two groups of words made up about 90% of the texts in the CoRA, as can be seen below in Table 111. These two groups contained mostly words related to topics of health sciences, and therefore many are recurrent in the five corpora of the CoRA, which may explain the lack of significant differences in relation to their frequency.

Word type	PT-EU	ES-EU	EN-GB	EN-PT^{EU}	EN-ES^{EU}
1-to-5-letter words	58.67%	60.68%	60.57%	60.46%	60.23%
6-to-10-letter words	31.59%	30.48%	33.09%	32.40%	32.65%
Total	**90.26%**	**91.16%**	**93.66%**	**92.86%**	**92.88%**

Table 111 – Percentages of 1-to-5-letter and 6-to10-letter words in the CoRA

Since the results obtained for **V5** (number of 1 to 5-letter words) and **V6** (number of 6-to-10-letter words) do not indicate a possible effect of L1 influence in these variables, it is not surprising that **V12** (number of prepositions), **V14** (number of relative pronouns), **V15** (number of personal pronouns), **V10** (number of coordinating conjunctions), and **V11** (number of subordinating conjunctions), all function words of mostly 1 to 10 characters of length, also do not have any effect of L1 influence.

Despite the lack of effects of L1 influence in relation to the frequencies of the variables in group 1 and 2 (Table 110), specific prepositions (**V12**) were shown to be possible markers of avoidance in OSRAs written in English by the non-L1 authors. These prepositions are "for" and "within" for both the L1 Portuguese authors and the L1 Spanish authors writing in English, and "across" and "over" for the L1 Spanish authors.

Prepositions are known to be a difficult part-of-speech to master in any foreign language (Ferrando 2006). Therefore, it is not unexpected that some prepositions are apparently avoided by the EN-PT^{EU} and the EN-ES^{EU} authors in the CoRA. Prepositions have the syntactical function of expressing the relation between two entities and most prepositions have several of these functions. (Quirk et al. 1985: 657). The prepositions "for", "across", and "over" have at least four different functions (Quirk et al. 1985: 678; 82; 96).

The results for the preposition "for" can be related to some of its communicative functions expressed through syntactic constructions that may be not very obvious options to the non-L1 users in the CoRA. I evaluated two of these functions. The first function is used to express purpose and the second to refer to an intended recipient (Quirk et al. 1985).

To verify the occurrence of the first syntactic function, I searched the English corpora of the CoRA, using the query [for * of][24], which brought results such as those shown below:

[24] An asterisk is used in to indicate that any word can occupy that space (Scott 2018a)

N	Example of "for [token] of)
1	It has been suggested that, rather than nicotine, it is polycyclic aromatic hydrocarbons that are responsible **for induction of** CYP1A2 isoenzymatic activity.
2	For example, this study has demonstrated that a satisfactory proportion of patients proceeded to randomisation (95%) and completed 30-day follow-up **for evaluation of** the primary outcome (100%)
3	Therefore, we conclude that mp-MRI on a 1.5-T magnet without ERC is highly specific and sensible, and may be used **for assessment of** tumor aggressiveness.
4	Descriptive statistics **for measures of** bone health, covariates and PA-by-intensity variables are reported in Table 1 for pre-menopausal and post-menopausal women separately.
5	Like other authors, we found few studies regarding this issue, with no comprehensive audit standards **for outbreaks of** VPD or communicable diseases being available.*
6	PA was confirmed in 69.2% of children under the age of 6 years who were referred **for assessment of** suspected PA.

Of all the concordances obtained from the three corpora (n=150), 50% are from the EN-GB corpus, whereas 21% are from the EN-ES^{EU} corpus and 29% are from EN-PT^{EU}.

To verify the occurrence of the second syntactic function (intended recipient), I searched the English corpora of the CoRA and obtained the distribution of the expressions "for women/for men/for children/for patients/for individuals". Results such as those shown below were obtained:

N	Examples of the expressions "for women/for men/for patients/for individuals"
1	It is concerning that manufacturers recommend GMP **for patients** over 1 year old even though there is no published evidence supporting its safety in patients o11 years of age and maternal PKU, particularly when GMP contributes to Phe intake and this impact is unknown.
2	**For men** with sexual dysfunction, patient education and appropriate prescribing of phosphodiesterase type 5 inhibitors, or where that fails, use of medicated urethral system for erections (MUSE), vacuum pumps or intracavernosal injections, may be helpful for enhancing sexual function.
3	Overall, this multistage empiric dietary approach may be recommended as a successful alternative to simplify the dietary management **for patients** with EoE.
4	The Se recommended daily allowance (RDA) **for individuals** aged between 14 and 52 years (excluding the states of pregnancy and lactation) has been set at 0.055 mg.

5	The applicability of the proportion of unemployed as a measure of deprivation **for women** of this age group may be questioned.
6	LNAA products are only recommended by the manufacturers **for patients** 48 years old, excluding maternal PKU patients.

Of all the concordances obtained from the three corpora (n=116), 68% are from the EN-GB corpus, whereas 11% are from the EN-ES^{EU} corpus and 21% are from EN-PT^{EU}.

As can be seen, the EN-GB authors used these two syntactic functions of "for" that were assessed more frequently than EN-PT^{EU} and EN-ES^{EU}. In the PT-EU and the ES-EU corpora, the equivalents of these functions can also be found, and are actually found at higher frequencies than when these authors use "for" in English to communicate purpose. In the PT-EU corpus 70 expressions are found for the query [para a * de/para o * de] with examples like the ones presented below:

N	Examples of the expressions "para a * de / para o * de"
1	Após o estabelecimento de um protocolo, em 2007, com o Hospital Juan Canalejo (A Coruña) e com o Hospital de Santa Marta (Lisboa), os doentes com idade inferior a 65 anos foram considerados **para a realização de** transplante pulmonar.
2	Não existem ainda métodos totalmente eficazes **para a identificação de** macrossomia antes do nascimento.
3	A ecografia do 3ºtrimestre (30-32 semanas) é um exame **para o diagnóstico de anomalias tardias** e avaliação do desenvolvimento fetal, nomeadamente a deteção da restrição do crescimento fetal, que afeta cerca 15% das gestações e está associada a morbimortalidade fetal tardia e neonatal.
4	O conhecimento e a capacidade crítica são componentes fundamentais **para a mudança** de comportamentos pelo que é tão importante investir numa literacia crítica.

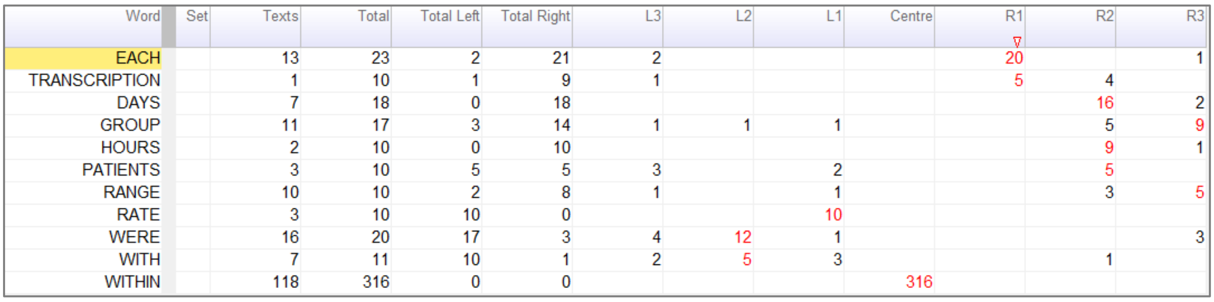
In the ES-EU corpus 150 hits are found for the query [para la * de/para el * de] with examples like the ones presented below:

N	Examples of the expressions "para la * de / para el * de"
1	Además, **para el diagnóstico** de ERC es imprescindible no sólo estimar el FG, sino medir la albuminuria, ya que ésta, aparte de ser un importante factor de riesgo vascular, es el principal marcador de progresión de la propia enfermedad renal.
2	Es importante la determinación del FG, es el criterio adoptado **para la valoración** de la función renal y evita las ERC ocultas que ocurren con la sola utilización de creatinina.

3	Por ejemplo **para la formación** de enlaces de tipo base de Schifft entre un aldehído de un soporte y un grupo amino de una enzima se requiere que el pH de la reacción sea de 10.
4	En los últimos años, varios candidatos a fármacos **para el tratamiento** de la THA han sido estudiados en fases clínicas.

The higher frequency of the preposition “for” in the EN-GB than in the EN-PT^{EU} and the EN-ES^{EU} OSRAs of the CoRA, and specifically the higher frequency of the syntactic functions evaluated (purpose and intended recipient), show that the non-L1 authors seem to avoid this preposition. This avoidance does not have to be necessarily due to lack of proficiency in English. This avoidance could be due to a lack of command of the many syntactic functions of “for”, which could be interpreted as not the most precise linguistic option. The fact is that, as it is a common preposition with many syntactic functions, it is possible that its frequency is not similar in all the English corpora.

Another preposition that was associated with avoidance in both the EN-PT^{EU} and the EN-ES^{EU} OSRAs is “within” whose only syntactic function refers to the indication of space, and which in most cases can be substituted by “in” (Quirk et al. 1985: 674). The avoidance observed could be due to the use of “in” instead of “within” as a preferred choice of non-L1 English authors. The instances of “within” in the English OSRAs of the CoRA do not contain any occurrence of syntactic meanings of “within” that cannot be conveyed using “in”, such as “within reach” or “within” referring to physical limits or boundaries, such “within these four walls” (Quirk et al. 1985: 674). This means that the syntactic uses of “within” are likely to be those that can alternate with “in”. When the groups are compared it is verified that both non-L1 groups (EN-PT^{EU} and EN-ES^{EU}) use “in” more frequently than the L1 English group. However, the differences between the EN-GB and the EN-PT^{EU} authors in relation to the frequency of use of the preposition “in” are not significant, whereas the EN-ES^{EU} authors use “in” significantly more frequently than the EN-GB authors (Z = -2.389; p = .017; Mean ranks: EN-ES^{EU} = 73.39, EN-GB = 57.61). Therefore, the EN-PT^{EU} and the EN-ES^{EU} authors could be using “in” where “within” could be a preferred option for an English native author to express the syntactic meaning of this preposition. Figure 22 shows the collocates of “within” in the CoRA extracted with WordSmith (Scott 2018b).



Word	Set	Texts	Total	Total Left	Total Right	L3	L2	L1	Centre	R1	R2	R3
EACH		13	23	2	21	2				20		1
TRANSCRIPTION		1	10	1	9	1				5	4	
DAYS		7	18	0	18						16	2
GROUP		11	17	3	14	1	1	1			5	9
HOURS		2	10	0	10						9	1
PATIENTS		3	10	5	5	3		2			5	
RANGE		10	10	2	8	1		1			3	5
RATE		3	10	10	0			10				
WERE		16	20	17	3	4	12	1				3
WITH		7	11	10	1	2	5	3			1	
WITHIN		118	316	0	0				316			

Figure 22 – Collocates of “within” in the English corpora of the CoRA

The expressions containing the main collocates to the right of “within” are found more frequently in the EN-GB corpus, as can be seen in below in Table 112.

N	Examples	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Total
1	other	**138**	74	36	248
2	within each	**15**	4	1	20
3	within XX day(s)	**16**		2	18
4	within-subject	**8**	1		9
5	within transcription	**6**			6
6	within X hour(s)	**2**			2

Table 112 – Expression containing the collocates to the right of “within”as distributed in the CoRA

A quick assessment of the frequency of use of the expression “in each”, as an alternative of “within each”, shows that the EN-GB group is also the group that uses it more, with 28 occurrences, whereas the non-L1 groups use that expression less frequently, i.e., 12 occurrences for the EN-ES^{EU} group and 19 occurrences for the EN-PT^{EU} group.

Although, the choice of “in” for “within” in the non-L1 users of the CoRA would need a more complete study which I have not conducted in this research, it can be affirmed that one possibility of the differences between the English groups in relation to use of “within” could be that non-L1 users alternate with “in”.

Two other prepositions were found to be significantly more frequent in the L1 English group (EN-GB) than in one of the non-L1 groups of the CoRA, i.e., in the EN-ES^{EU} group. Those prepositions are “across” and “over”. The occurrences of these two prepositions are shown below in Table 113.

N	Expressions with	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Total
	ACROSS	**102**	**16**	**15**	**133**
1	across different	4	2		6
2	across the X	**44**	6	4	54
3	other	54	8	11	73
OVER		**201**	**65**	**54**	**320**
4	over the course of	4			4
5	over the study period	**11**			11
6	over the XX-year period	2			2
7	over X years	**14**	5	1	20
8	other	170	60	53	283

Table 113 – Occurrences of the PRPs "across" and "over" in the CoRA

The expressions 2, 5 and 7 in Table 113 were analyzed to understand their syntactic function(s). The occurrence described in 2 contains instances such as "the missingness ranged between 0.0–2.9% across the 28 markers", "accounting for 37 billion Euros annually across the 27 countries", "clinical follow-up received by women with breast cancer across the social spectrum", "across all these different populations", or "risk estimates were consistent across the different ages". In my interpretation, these instances indicate "static pervasiveness" (Quirk et al. 1985: 685) in the sense of extending to every entity (marker, country, population, women, age). Within this meaning "across" may alternate with expressions such as "in all the groups" or "for all groups". However, after extracting the concordances of these expressions it was verified that they are very few and although slightly more frequent in the EN-PT^{EU} and the EN-ES^{EU} OSRAs, some of these uses may not be grammatically correct if alternated with "across" (e.g. "Waist circumference was reduced in all the participants"). Therefore, the lower frequency of use of the preposition "across" could be related to the studies in the EN-PT^{EU} and the EN-ES^{EU} OSRAs of the CoRA, which could be describing, for example, results that do not refer to conditions applying to all the individuals being studied.

The instances in 5 and 7 in Table 113 indicate "duration" (Quirk et al. 1985: 689) and in many cases (such as "this rate remained stable over the study period", "There was a decreasing trend in the mean and median hospital stay for [...] over the study period") can alternate with the preposition "during" whose cognate in Portuguese and Spanish ("durante") would make it more familiar to these non-L1 authors. See in Table 114, that the occurrences of the expression "during the study period", despite being only a few, are more used by the non-L1 English authors of the CoRA (EN-PT^{EU} and EN-ES^{EU}).

Examples	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Total
during the study period	5	5	7	17
over the study period	11			11
Total	**16**	**5**	**7**	**28**

Table 114 – Occurrences of the expressions "during the study period" and "over the study period" in the CoRA

In fact, as described in section 4.2.2 of chapter 4, the preposition "during" is more frequently used by the non-L1 English authors (EN-PT^{EU} and EN-ES^{EU}) than by the L1 English authors (EN-GB) of the CoRA, although this preposition was found to be associated with L1 transfer only in the EN-ES^{EU} authors of the CoRA. Nonetheless, the possibility of alternating "during" and "over" could justify the findings in relation to these prepositions.

Within the groups 1 and 2 in Table 110, the coordinating conjunction (**V10**) "as=well=as" was found to be associated with L1 transfer in EN-ES^{EU} authors, and the subordinating conjunction (**V11**) "that" was found to be associated with L1 transfer in both the EN-PT^{EU} and the EN-ES^{EU} OSRAs. The latter is not considered in this section as the reasons for its possible transfer of syntactic function from the EN-PT^{EU} and the EN-ES^{EU} authors' L1 was discussed in the section dedicated to subordinating conjunctions (section 4.3.6).

The frequent use of the coordinating conjunction "as well as" by the EN-ES^{EU} authors of the CoRA seems to mirror the use that L1 users of Spanish make of the equivalent conjunctions "así como" and "así como también". The EN-PT^{EU} authors also use "as well as" more frequently than EN-GB users even though the difference is not significant. When the data of "as well as" and its equivalents in Portuguese and Spanish are plotted in a bubble graph, the similarity in relation to the frequency of use of "as well as" in the EN-PT^{EU} and the EN-ES^{EU} groups can be appreciated (Figure 23). It is also possible to see that both the EN-PT^{EU} and the EN-ES^{EU} groups diminish the use of this coordinating conjunction when writing in English, though in the case of the EN-ES^{EU} authors, the decrease is not enough as to reveal significant differences – or cross-language congruity, and at the same time, there are too many occurrences of "as well as" for the EN-ES^{EU} and the EN-GB and to be similar – or intralingual contrast.

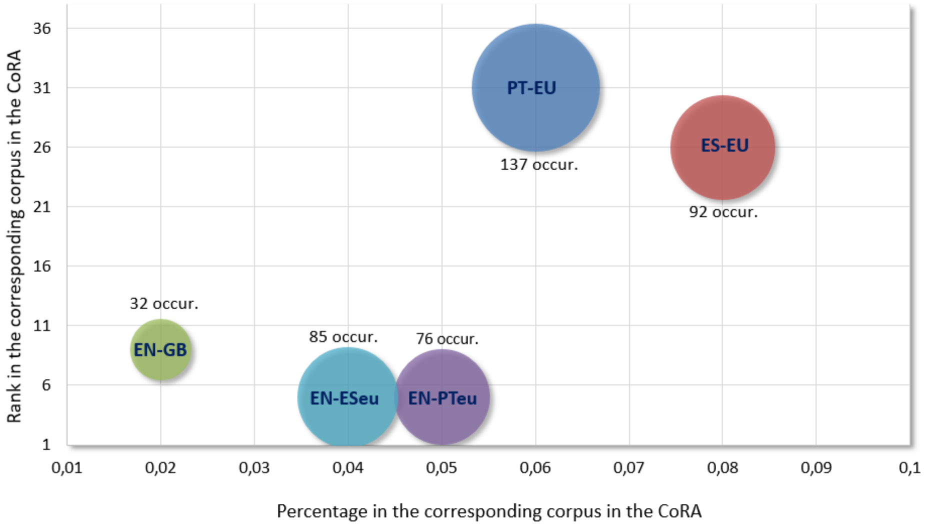


Figure 23 – Distribution of the coordinating conjunction "as well as" and its equivalents in the PT-EU and the ES-EU corpora of the CoRA

The syntactic function of this conjunction is one of coordination. It is used when a repetition of the coordinator "and" is to be avoided because it has been already used, like in sentence a. below:

a. Independent samples *t*-tests confirmed that moderate **and** severe subgroups differed in speech comprehension ability (t(18)=7.77), **as well as** in written comprehension, repetition, naming, reading and writing abilities (all p<0.05).

The coordinating conjunction "as well as" is also use when the element after "as well as" belongs into the same category that something that was said before, like in sentence b. below:

b. Nuestro trabajo describe el **fenotipo asociado a cada tipo de** alteración de BMPR2, **así como de** las formas asociadas a KCNK3 **y** TBX4, mucho menos conocidas.

The coordinating conjunction "as well as" is also used when the element following this conjunction is the argument of the main verb of the sentence but belongs in a different category than another argument mentioned before, such as in sentence c. below:

c. Table 2 **reveals** the mean BMI and BMI *Z*-scores for the three groups, **as well as the prevalence** of obesity in each of the three groups.

However, other uses of "as well as" are simple an alternative to "and", such as the ones in sentences c., d., and e below:

d. Regarding biochemical parameters, insulin **as well as** ALT blood levels were significantly decreased within this group in accordance with previous studies.
e. LPA may modify N-methyl-D-aspartate (NMDA) receptor functions in hippocampal neurons **as well as** calcium intracellular levels

All these syntactic functions also exist in Spanish, and for that matter, also in Portuguese, but I focus on Spanish because of the significant differences between the EN-GB and the ES-EU groups. An inspection to the ES-EU OSRAs shows precisely that, as can be seen below in sentences f., g., h., and i.:

f. Concretamente, se observó una significativa disminución de los niveles plasmáticos de glucosa a los 30 **y** 60 minutos, **así como** del AUC de glucosa durante el OGTT (Figura 8c), respecto a las hembras SR no tratadas.

g. Las técnicas de reacción en cadena de polimerasa (PCR) **han permitido determinar** la etiología de las IRA en niños, **así como** **conocer** que las infecciones virales asintomáticas son frecuentes.

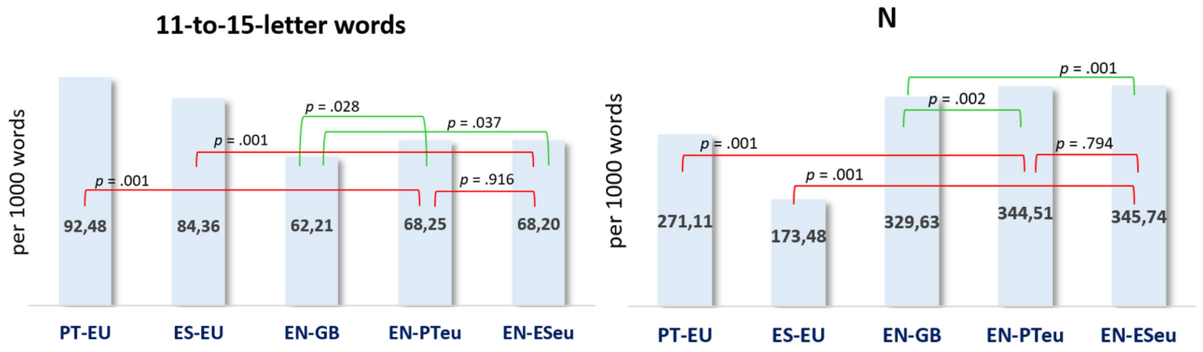
h. La activación de esta vía **da como resultado** una **expansión de las células** osteoprogenitoras, **así como** **una reducción** de la apoptosis de los osteoblastos, lo que conlleva efectos anabólicos sobre el hueso.

i. Por otro lado, el miR-22-3p es un miRNA sérico que **se ha asociado** previamente a la fractura osteoporótica, **así como** **ha sido implicado** en la diferenciación osteogénica.

A thorough analysis would be needed to determine the exact reasons for the significant differences between these groups. However, two possible explanations can be proposed. One would be that the more frequent use of "as well as" in non-L1 users of English when writing OSRAs is related to an overuse as an alternative of "and". Another could be related to the need to save space, writing more information using fewer words.

The third group in Table 110 contains five variables that were associated with only one effect of L1 influence, i.e., intralingual contrast. In brief, there are significant differences between the Portuguese/Spanish authors writing in English and the L1 English authors writing in their L1.

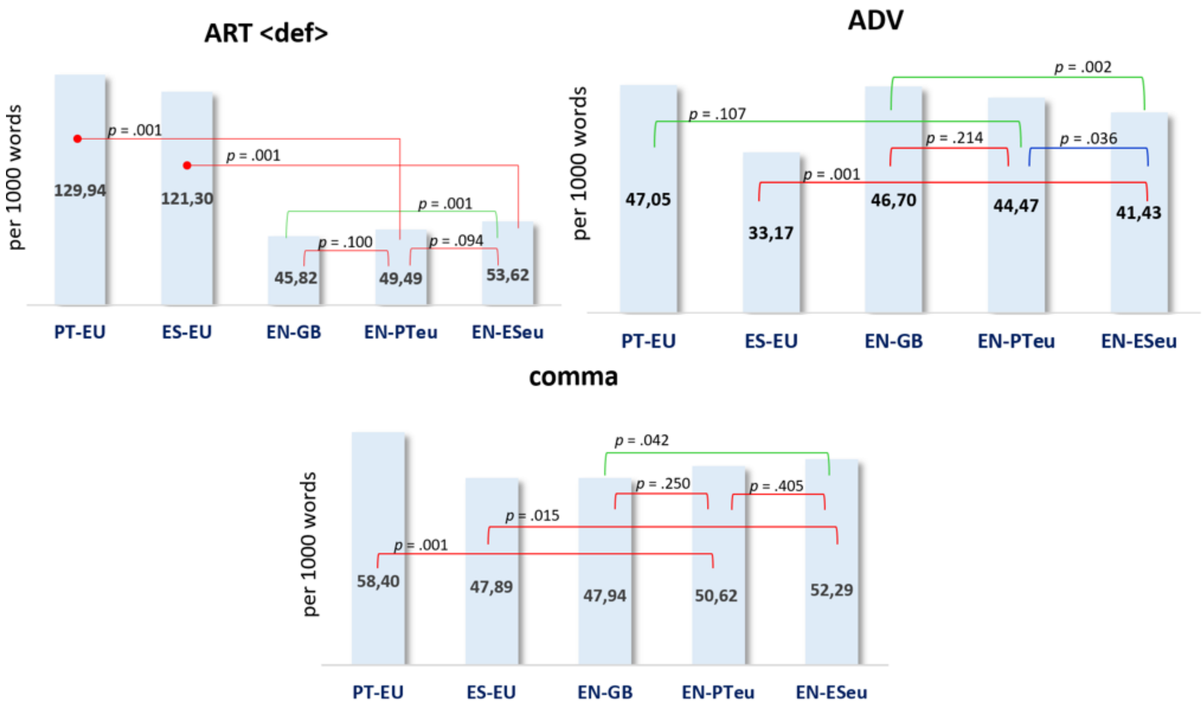
Two of these variables, **V7** (11-to-15-letter words) and **V18** (number of nouns – N), were found to have an effect of intralingual contrast in OSRAs written in English by both groups of non-L1 authors, i.e., the L1 Portuguese and the L1 Spanish authors writing in English (EN-PT^{EU} and EN-ES^{EU}), as shown in Figure 24 below.



For all results, *p* value < .05 = significant differences. The red lines refer to results that do not indicate an effect of L1 influence. The green lines indicate an effect of L1 influence.
Legend: **V7**: 11-to-15-letter words and **V18**: number of nouns – N

Figure 24 – Variables with only one effect of L1 influence – Intralingual contrast in EN-PT^{EU} and EN-ES^{EU} OSRAs

Other three variables, **V1** (number of commas), **V8** (number of definite articles – ART <def>) and **V17** (number of adverbs – ADV), were found to have an effect of intralingual contrast only in OSRAs written in English by the L1 Spanish authors (EN-ES^{EU}) as shown below in Figure 25.



For all the results, *p* value < .05 = significant differences. The red lines refer to results that do not indicate an effect of L1 influence. The green lines indicate an effect of L1 influence. The blue line in the ADV stands for a result that despite being significant (which is what is expected) cannot be associated with the effect of L1 influence for which the groups (EN-PT^{EU} and EN-ES^{EU}) were tested (inter-L1 group heterogeneity). This is because, as explained in the methodology (section 3.2.3), the result of the inter-L1 group heterogeneity test (i.e., independent-samples *t*-test) is interpreted together with the result of the test carried out to verify the intra-L1 group homogeneity (Levene's). Since the Levene's test (*F* = 7.755; *p* = .031) shows that the variances of the number of ADVs in these groups are significantly different, the groups cannot be assumed to be from different populations despite the significant results of the *t*-test (*t* (118.822 = 2.124; *p* = .036). That is, the significant results of the *t*-test could also be due to sampling problems (e.g. not enough cases). Legend: **V8**: number of definite articles – ART <def> , **V17**: number of adverbs – ADV), and **V1**: number of commas.

Figure 25 – Variables with one effect of L1 influence – Intralingual contrast only in EN-ES^{EU} OSRAs

In relation to the results of group 3, I only discuss in this section the POS "there" since the other findings were examined above in section 4.4.

The POS "there", when used in its existential function (Quirk et al. 1985: 1403) has been reported as "a key feature in the academic author's rhetorical toolbox", used to organized the text, despite the style guides of scientific journals recommending its avoidance for being an "empty structure" (Jiang and Hyland 2020b: 2). The results of this study in relation to the frequency of "there" show that non-L1 users of English seem to avoid "there" in general since significant differences were found between the EN-GB authors and the EN-PT^{EU}/EN-ES^{EU} authors in relation to the frequency of "there" in the CoRA.

To verify if the same could be the case with the existential "there", I counted the occurrences of the expressions "there is", "there are", "there was", and "there were". The results obtained are shown below in Table 115.

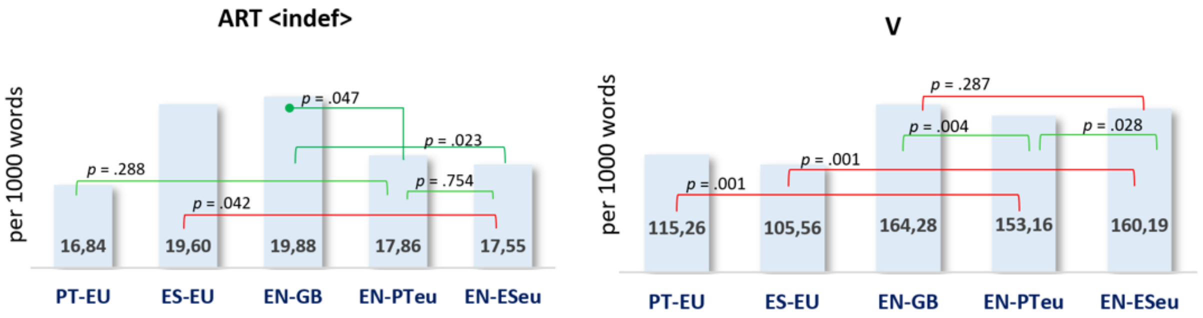
Existential "there"	EN-GB	EN-PT^{EU}	EN-ES^{EU}	Total
there are	54	32	39	**125**
there is	105	69	61	**235**
there was	150	76	38	**264**
there were	94	42	45	**181**
Total	**403**	**219**	**183**	**805**

Table 115 – Occurrences of existential "there" in the CoRA

As can be seen, L1 English authors use the existential "there" more frequently than the non-L1 authors in the CoRA. To assess the differences between these groups, I ran a statistical test (non-parametric Mann-Whitney test since the distribution of the samples is not normal). The results indicate that there is a statistically significant difference in the ranked frequencies of existential "there" between the EN-GB and the EN-PT^{EU} authors (Z = -3.752; p = .001; mean ranks: 72.85 and 48.97, respectively); and between the EN-GB and the EN-ES^{EU} authors (Z = -3.059; p = .002; mean ranks: 65.20 and 46.30, respectively). Therefore, as a rhetorical feature, the existential "there" is more frequent in the L1 (EN-GB) than in the non-L1 (EN-PT^{EU} and EN-ES^{EU}) authors in the CoRA. This difference may be the reason for the overall discrepancy between the groups in relation to "there". In fact, it can be hypothesized that non-L1 English authors follow style guidelines more strictly than L1 English authors do, perhaps to pass the editorial process more easily.

Lastly, four variables (i.e., **V9**: number of indefinite articles, **V13**: number of demonstrative pronouns, **V16**: number of adjectives, **V19**: number of verbs) have high potential in relation to their overall frequency to mark L1 influence in OSRAs written in English by the L1 Portuguese/Spanish authors since they accumulate two effects of L1 influence.

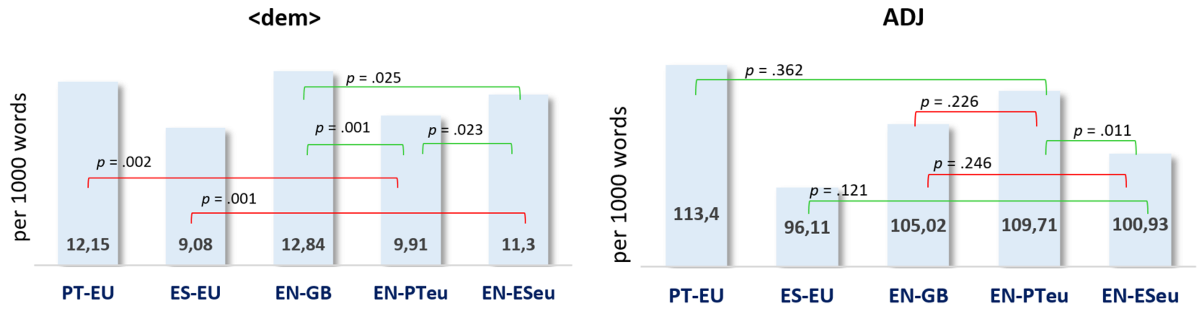
The variables **V9**: number of indefinite articles and **V19**: number of verbs have the potential to mark the influence of the authors' L1 only in the EN-PT^{EU} OSRAs, as shown in Figure 26 below.



For all the results, *p* value < .05 = significant differences. The red lines refer to results that do not indicate an effect of L1 influence. The green lines indicate an effect of L1 influence. Legend: **V9:** indefinite articles ART <indef> and **V19**: number of verbs V.

Figure 26 – Variables with two effects of L1 influence only in the EN-PT^{EU} OSRAs

The variables **V13** (number of demonstrative pronouns) and **V16** (number of adjectives) have the potential to mark the influence of the authors' L1 in both the EN-PT^{EU} and the EN-ES^{EU} OSRAs, as can be seen in Figure 27 below.



For all the results, p value < .05 = significant differences. The red lines refer to results that do not indicate an effect of L1 influence. The green lines indicate an effect of L1 influence. Legend: **V13**: number of demonstrative pronouns – <dem> and **V16**: number of adjectives – ADJ.

Figure 27 – Variables with two effects of L1 influence in both EN-PT^{EU} and the EN-ES^{EU} OSRAs

In relation to the findings in this group (4 in table 110), I discuss only the demonstrative pronoun "this" since the other findings were examined in the corresponding section (4.5).

In the scientific text, the demonstrative pronoun "this" is mostly used as an anaphoric element, i.e., it functions with a co-reference that is mentioned before in the text (Quirk et al. 1985). This use contributes to textual cohesion. The general recommendation for the use of "this" is not to leave the pronoun "unattended", i.e., to accompany "this" with a summary word so that the reference to the antecedent is clear (Swales 2005: 2). As a form of ensuring the clarity of the text style guides frequently recommend the authors to avoid the use of “this” without a clear antecedent (Wulff, Römer, and Swales 2012). Recently, an increase in the use of the demonstrative pronoun “this” was shown to have taken place in Applied Linguistics, Sociology and Biology, with “an increase in the proportional use of unattended ‘this’” in the applied fields (Jiang and Hyland 2020a: 27). Based on this information, I aimed at verifying if the general significant differences that exist between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} OSRAs in the CoRA could be related to differences in the use of the unattended “this” as indicated by Jiang and Hyland (2020a). For this, the occurrences of “this” in the English corpora of the CoRA were extracted using WordSmith (Scott 2018b). The extraction was carried out using a query with the syntax [this Verb], i.e., occurrences of “this” that are not followed by a noun. The results obtained are shown below in Table 116.

Unattended "this"	**EN-GB**	**EN-PT^{EU}**	**EN-ES^{EU}**	**Total**
this could …	9	**12**	**16**	37
this suggests …	**10**	6	11	27
this would …	**17**	3	7	27
this might …	4	4	6	14
this indicates …	1	1	7	9
this means …	1	3	5	9
this included …	**8**			8
this does …	1	2	4	7
this increased …	**4**	1	2	7
this should …	**4**	1	2	7
this supports …	**5**	1	1	7
this will …	**6**		1	7
this seems …		3	2	5
this showed …	**4**		1	5
this suggested …	**2**			2
this demonstrated …	**1**			1
this found …	**1**			1
this identified …	**1**			1
Total	**79**	**37**	**65**	**181**

Table 116 – Occurrences of unattended “this” in the CoRA

As can be seen, the L1 English authors use the unattended “this” more frequently than the non-L1 authors. To assess for the differences between these groups, I ran a statistical test (non-parametric Mann-Whitney test since the distribution of the samples is not normal). The results indicate that there are no statistically significant differences between the EN-GB and the EN-PT^{EU} OSRAs or between the EN-GB and the EN-ES^{EU} OSRAs in relation to this variable in the CoRA. Therefore, the general significant differences that exist between the EN-GB and the EN-PT^{EU}/EN-ES^{EU} OSRAs in the CoRA are not related to the unattended anaphoric use of “this”, but to other uses which were not evaluated in this research.

In the conclusions, presented in the next section, I provide a table with a summary of the most relevant findings in relation to the variables associated with L1 influence in OSRAs written in English by Portuguese/Spanish authors, and make the final remarks in relation to the whole research project.

5. Conclusions

This study shows that it is possible to find effects of the L1 influence in highly specialized genres like the original scientific research article in the health sciences produced by advanced non-L1 English authors who are L1 Portuguese/Spanish users. These L1 effects comprise both content-independent and content-dependent variables.

The findings of the empirical work indicate that the variables with the higher potential to mark the influence of the L1 in the Portuguese/Spanish writing OSRAs in English, are prepositions, coordinating conjunctions, adverbs, nouns, demonstrative pronouns, adjective and verbs, as described in Table 117 below.

		Content-independent		Content-dependent	
Groups	Variables	Analysis of frequencies		POS Analysis	
		EN-PT^{EU}	EN-ES^{EU}	EN-PT^{EU}	EN-ES^{EU}
No effects of L1 influence	**V2**: average sentence length in words	--	--		
	V12: number of prepositions	--	--	“for” and “within” -> avoidance	“for”, “across”, “over”, and “within” -> avoidance “during” -> L1 transfer potential with effects III and IV
	V14: number of relative pronouns	--	--	--	--
	V15: number of personal pronouns	--	--	--	--
One effect of L1 influence – Cross-Language	**V3**: number of paragraphs	--	--		
	V4: standardized type/token ratio (STTR)	--	--		
	V5: number of 1-to-5-letter words	--	--		
	V6: number of 6-to-10-letter words	--	--		
	V10: number of coordinating conjunctions	--	--	--	L1 transfer potential “as=well=as” effects III and IV
One effect of L1 influence – Intralingual Contrast	**V1**: number of commas	--	L1 transfer potential		
	V7: number of 11-to-15-letter words	L1 transfer potential	L1 transfer potential		
	V8: number of definite articles	L1 transfer potential	L1 transfer potential		
	V17: number of adverbs	L1 transfer potential	--	“where”, “there”, “out” -> avoidance “namely” -> L1 transfer potential. Effects I, II, and III	“where” and “there” -> avoidance
	V18: number of nouns	L1 transfer potential	L1 transfer potential	“presence/existence” -> L1 transfer potential with effects III and IV	“study” -> L1 transfer potential with effects III and IV
Two effects of L1 influence	**V9**: number of indefinite articles	avoidance effects I, II, III	avoidance Effects I and II		
	V13: number of demonstrative pronouns	avoidance effects I, II, e IV	avoidance effects I, II, e IV	“this” and “those” -> avoidance	“those” -> avoidance
	V16: number of adjectives	L1 transfer effects I, II and II	L1 transfer effects I, II and II	“likely” -> avoidance	--
	V19: number of verbs	Avoidance effects I, II e IV	--	“to see” and “to identify” -> avoidance	“to find” -> L1 transfer effects III and IV

Table 117 –Summary of the results obtained from the quantitative and linguistic analyses of the CoRA

In the case of content-dependent variables L1 influence comprises transfer of syntactic function of specific prepositions, coordinating conjunctions, adverbs and verbs which hold the greatest potential to mark the influence of the Portuguese and Spanish authors' L1 in OSRAs.

Three of the variables with more relevant results in relation to the number of effects of L1 influence (<dem>, V, ART <indef>, with two L1 effects of L1 influence) show that the non-L1 English authors who are L1 Portuguese or Spanish users writing OSRAs in English in the CoRA deploy strategies of avoidance. This finding agrees with the literature, which refers avoidance as more frequent in advanced users of English than in learners.

The strategies of L1 transfer are more like to occur with commas, 11-to-15 letter words, the use of the definite article, nouns, adverbs, and adjectives. Also, Non-L1 authors seem to be more prone to nominalization than the native authors who use verbs significantly more frequently than non-L1 authors. Also, the Portuguese authors writing OSRAs in English are more prone than the EN-GB authors and EN-ES^{EU} authors to using adjectives.

Due to the methodology followed during this research other markers were detected that are associated with a more frequent use by the L1 than by the non-L1 English authors who are L1 Portuguese/Spanish. The demonstrative pronoun "this" when used as a determiner was shown to have the potential to function as a marker of non-nativeness, and if associated with expressions like "in this sense" or "this fact" may also be associated with the L1 Portuguese/Spanish authors. Similarly, the demonstrative pronoun "those" was shown to be likely to function as a marker of non-nativeness, since the L1 English authors appear to be more comfortable using it than the non-L1 English authors who are L1 Portuguese/Spanish users.

The findings of this research can be applied or tested with other scientific genres. Other scientific genres and non-scientific genres may also find in these results variables that can indicate the influence of the L1 in non-L1 English writing. Likewise, the findings may be extended to non-L1 English authors who are L1 users of other Romance languages with similar lexical choices and syntactic constructions.

This study has also showed the usefulness of the unified framework for investigating L1 influence (Jarvis 2010, 2000) in scientific writing, specifically in the original scientific research article. The research also showed the utility of distribution thresholds to obtain lists of POS for comparison of frequencies, ranks, and percentages of lexical units in corpora, which may lead to specific L1 influence markers, in this case, at the lexical level.

Probably, the most enriching experience of conducting this research was the construction of the Comparative Corpora of Research Articles – CoRA. Building corpora of my own allowed for the customization of the research material and the learning of many details related to this process.

Finally, despite the challenge that the mixed-methods represented, the approach is one that works well and provides empirical data to direct future research.

5.1. Limitations of the present study

Limitations to the present study have been provided above in section 3.1.4 in relation to the compilation of the CoRA. Other limitations faced in this research concern the L1 of the OSRAs authors, which are assumed instead of objectively known or informed by the actual writers of the texts. By the same token, the assumption of the authors being advanced English users is based on principles of the genre OSRA and in discourse community. The criteria used to decide on the L1 and the stage of development of the non-L1 (i.e., English) of OSRAs' authors are perfectly valid and preceded by other researchers' work. Nonetheless, it is acknowledged that informed authors' L1 and knowledge of English would have been ideal and could have resulted in a different research outcome. The collection of data directly from the authors, according to their perceptions of what their native language is and how proficient they are in English, could make it possible to obtain more reliable results.

Also, the theoretical framework used as a reference for the classification of the OSRAs authors in L1 or non-L1 users of English (Kachru 1997, 1992, 1985) may introduce interpretations of superiority of English native speakers in terms of language realization, i.e., OSRA writing. In this respect, approaching the comparative nature of this study from theoretical paradigms that recognize non-L1 English users' realization of the language as nothing less than another variety of English may be more in line with the reality of non-L1 English researchers when using English to communicate their science. However, these theoretical paradigms may contradict the need of comparative studies to use a given group of language users as a reference. Therefore, the problems posed by theories like multilinguism and translinguism, that avoid considering any variety of a language as a model to follow, need to be carefully examined before any form of comparative studies can take place.

5.2. Future research directions

The research developed during this study has high growth potential. Therefore, its continuation may focus on five main aspects.

First, the detailed examination of the syntactic structures found to be at the base of the significant differences between groups concerning certain variables would add insight to the capacity of these variables to mark L1 influence. Additionally, the semantic implications of the use of specific syntactic structures can also contribute to a more complete description of a possible marker. That is, if it is possible to say, for example, that the adverb “namely” functions as a marker of L1 influence in Portuguese authors writing OSRAs in English, and its specific position in the sentence and the meaning deriving from this usage contribute to marking L1 influence, then it can be said that the adverb “namely” would be a stronger marker.

Second, other language varieties could be incorporated in the CoRA to compare the results of this study with results that arise from other comparisons. The most obvious choices for growing the number of language varieties within the CoRA are Brazilian Portuguese, any of the Latin American varieties of Spanish, and any of the inner circle varieties of English, e.g., American, and even outer circle varieties of English, e.g., South African. However, the CoRA may also grow in the number of languages contemplated, including other Romance languages like French.

Third, OSRAs from other scientific areas or other scientific genres from health sciences and/or other areas could be incorporated in the CoRA in the three languages currently available, allowing for comparisons between scientific areas and between genres.

Fourthly, the instances of cross-linguistic influence previously explained would benefit from evaluation in the light of Second Language Acquisition and Language Transfer, which was not planned in the research design implemented.

Finally, the mixed nature of the methodological approach implemented in this exploratory study posed important challenges in terms of quantitative methods. However, the combination of approaches remains a strong interest, and therefore, the next stage would be to test the results in a computational model.

References

[ode]. 2019. 'LiteraryDevices Editors'. https://literarydevices.net/ode/ Retrieved on November 5 , 2019.

Aad, Georges, Tatevik Abajyan, B Abbott, J Abdallah, S Abdel Khalek, et al. 2012. 'Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC', *Physics Letters B*, 716: 1-29.

Aad, Georges, B Abbott, J Abdallah, O Abdinov, Rosemarie Aben, et al. 2015. 'Combined Measurement of the Higgs Boson Mass in p p Collisions at s= 7 and 8 TeV with the ATLAS and CMS Experiments', *Physical review letters*, 114: 191803.

Aberson, Christopher L, Dale E Berger, Michael R Healy, Diana J Kyle, and Victoria L Romero. 2000. 'Evaluation of an interactive tutorial for teaching the central limit theorem', *Teaching of Psychology*, 27: 289-91.

Adler, Joachim. 2014. 'Mapping the Ancient City: Historical Linguistics and Conceptual Clarification', *Philosophy of Language and Linguistics: The Legacy of Frege, Russell, and Wittgenstein*, 53: 11.

Alfajarín, Joan-Rafael Ramos. 2013. 'Norma y variación lingüística: paralelismos y divergencias entre el español y el catalán', *Normas: revista de estudios lingüísticos hispánicos*: 127-60.

Alpaydin, Ethem. 2004. *Introduction to Machine Learning* (MIT press: Cambridge, MA).

Amuchi, Faith, Ameer Al-Nemrat, Mamoun Alazab, and Robert Layton. 2012. "Identifying cyber predators through forensic authorship analysis of chat logs." In *2012 Third Cybercrime and Trustworthy Computing Workshop*, 28-37. IEEE.

Anderson, Carolyn J. . 2010. 'CENTRAL LIMIT THEOREM.' in Irving B Weiner and W Edward Craighead (eds.), *The Corsini Encyclopedia of Psychology* (John Wiley & Sons).

Anderson, Kenneth, and Joan Maclean. 1997. 'A Genre Analysis Study of 80 Medical Abstracts', *Edinburgh working papers in applied linguistics*, 8: 1-23.

Ansell, Emily B, Aidan GC Wright, John C Markowitz, Charles A Sanislow, Christopher J Hopwood, et al. 2015. 'Personality disorder risk factors for suicide attempts over 10 years of follow-up', *Personality Disorders: Theory, Research, and Treatment*, 6: 161.

Argamon, Shlomo, and Moshe Koppel. 2012. 'A systemic functional approach to automated authorship analysis', *JL & Pol'y*, 21: 299.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. 'Gender, genre, and writing style in formal written texts', *Text-The Hague Then Amsterdam Then Berlin-*, 23: 321-46.

Argamon, Shlomo, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. 'Automatically profiling the author of an anonymous text', *Communications of the ACM*, 52: 119-23.

Arostegui, Maitena Etxebarria, and Maitena Etxebarria. 1985. *Sociolingüística urbana: el habla de Bilbao* (Universidad de Salamanca).

Atkins, Sue, Jeremy Clear, and Nicholas Ostler. 1992. 'Corpus design criteria', *Literary and Linguistic Computing*, 7: 1-16.

Baker, Colin. 2001. 'Bilingualism: Definitions and Distinctions.' in, *Foundations of bilingual education and bilingualism* (Multilingual matters).

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2012. 'Gender in twitter: Styles, stances, and social networks', *https://arxiv.org/vc/arxiv/papers/*.

Barlow, Michael. 2010. "Individual usage: a corpus-based study of idiolects." In *34th International LAUD Symposium*. Landau, Germany: .

Barthes, Roland. 1986 [1964]. *Elements of semiology* (HILL and WANG - A DIVISION OF FARRAR, STRAUS AND GIROUX: New York).

Barton, William E. 1926. 'The Truth About the Bixby Letter - Dr. Barton Believes Lincoln Actually Wrote to Mrs. Bixby, But Thinks That Any Existing Copies Are Forgeries', *The Dearborn Independent*.

Bawarshi, Anis S, and Mary Jo Reiff. 2010. *Genre: An introduction to history, theory, research, and pedagogy* (Parlor Press West Lafayette, IN).

Bazerman, Charles. 1984. 'Modern evolution of the experimental report in physics: Spectroscopic articles in Physical Review, 1893-1980', *Social studies of science*, 14: 163-96.

BBCNews. 2013. 'How JK Rowling was unmasked', *BBC News*.

Belcher, Diane. 2009. 'What ESP is and can be: An introduction', *English for specific purposes in theory and practice*: 1-20.

Benfield, John R, and Christine B Feak. 2006. 'How authors can cope with the burden of English as an international language', *Chest*, 129: 1728-30.

Bennett, Karen. 2008. 'English academic discourse : its hegemonic status and implications for translation', Univeristy of Lisbon.

Bevendorff, Janek, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, et al. 2020. "Shared Tasks on Authorship Analysis at PAN 2020." In *European Conference on Information Retrieval*, 508-16. Springer.

Bhatia, Vijay Kumar. 1993. *Analysing genre: Language use in professional settings* (Routledge).

Bhatia, Vijay Kumar. 2014. *Analysing genre: Language use in professional settings* (Routledge).

Biber, Douglas. 1989. 'A typology of English texts', *Linguistics*, 27: 3-44.

Biber, Douglas. 1993. 'Representativeness in corpus design', *Literary and Linguistic Computing*, 8: 243-57.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison* (Cambridge University Press).

Biber, Douglas, and Susan Conrad. 2009. *Register, genre, and style* (Cambridge University Press).

Biber, Douglas, and Susan Conrad. 2019. *Register, genre, and style* (Cambridge University Press).

Biber, Douglas, Biber Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use* (Cambridge University Press).

Bick, Eckhard. 2000. *The Parsing System" Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework* (Aarhus Universitetsforlag).

Bick, Eckhard. 2006. "A constraint grammar parser for spanish." In *Proceedings of TIL 2006–4th Workshop on Information and Human Language Technology (Ribeirão Preto, October 27–28, 2006)*, 3-10.

Bick, Eckhard. 2010. "Degrees of orality in speech-like corpora: Comparative annotation of chat and e-mail corpora." In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 721-29.

Bick, Eckhard. 2012. "Towards a Semantic Annotation of English Television News-Building and Evaluating a Constraint Grammar FrameNet." In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, 60-69.

Bick, Eckhard. 2014. 'PALAVRAS: a Constraint Grammarbased Parsing System for Portuguese.' in Tony Berber Sardinha and Thelma Lurdes São Bento Ferreira (eds.), *Working with Portuguese corpora* (Bloomsbury Publishing Plc: United Kingdom and United States of America).

Bloch, Bernard. 1948. 'A set of postulates for phonemic analysis', *Language*, 24: 3 - 46.

Bloomfield, Leonard. 1933. *Language* (Holt, Rinehart, and Winston: New York).

Bond, Gary D, and Adrienne Y Lee. 2005. 'Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language', *Applied Cognitive Psychology*, 19: 313-29.

Bradford, Richard. 2005. *Stylistics* (Routledge).

Brien, Donna Lee, and Bronwyn Fredericks. 2020. 'Collaborative writing to enhance cross-cultural understanding within the Academy', *Sat*.

Burger, John D, John Henderson, George Kim, and Guido Zarrella. 2011. "Discriminating gender on Twitter." In *Proceedings of the conference on empirical methods in natural language processing*, 1301-09. Association for Computational Linguistics.

Burke, Michael. 2014. 'Introduction Stylistics: From classical rhetoric to cognitive neuroscience.' in, *The Routledge Handbook of Stylistics* (Routledge).

Burlingame, Michael. 1999. 'The Trouble With The Bixby Letter', *American Heritage*.

Busch-Lauer, Ines-A. 1995. 'Abstracts in German medical journals: a linguistic analysis', *Information processing & management*, 31: 769-76.

Calero, Francisco. 2006. 'Luis Vives fue el autor del Lazarillo de Tormes', *Espéculo*, 32: 1-62.

Carney, Terrence. 2014. 'Being (im) polite: A forensic linguistic approach to interpreting a hate speech case', *Language Matters*, 45: 325-41.

Castelvecchi, Davide. 2015. 'Physics paper sets record with more than 5,000 authors', *Nature*, 15.

CE. 2014. "Languages for Democracy and Social Cohesion: Diversity, Equity and Quality. Sixty Years of European Co-Operation." In, edited by Education Department - Language Policy Unit, 37. Strasbourg: Council of Europe, Language Policy Division.

Cenoz, Jasone. 2013. 'Multilingualism.' in Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics* (Blackwell Publishing Ltd.).

Chambers, Jack K. 2013. 'Studying language variation: An informal epistemology.' in J.K. Chambers and Natalie Schilling (eds.), *The handbook of language variation and change* (Wiley-Blackwell).

Christiansen, M. H., and I. Arnon. 2017. 'More Than Words: The Role of Multiword Sequences in Language Learning and Use', *Top Cogn Sci*, 9: 542-51.

Clarke, Isobelle, and Jack Grieve. 2019. 'Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018', *PloS one*, 14: e0222062.

Cohn, Trevor, and Mirella Lapata. 2007. "Machine translation by triangulation: Making effective use of multi-parallel corpora." In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 728-35.

Coseriu, Eugenio. 1978. *Sincronía, Diacronía e Historia. El Problema del Cambio Lingüístico* (Biblioteca Románica Hispánica. Editorial Gredos.: Madrid, España).

Coseriu, Eugenio. 1982. *Sentido y Tareas de la Dialectología* (Centro de Lingüística Hispánica, Instituto de Investigaciones Filológicas, Universidad Nacional Autónoma de México: México, D. F.).

Coulthard, Malcolm. 1992. *Advances in spoken discourse analysis* (Routledge: USA and Canada).

Coulthard, Malcolm. 1994. 'On the use of corpora in the analysis of forensic texts', *International Journal of Speech, Language and the Law*, 1: 27-43.

Coulthard, Malcolm. 2004. 'Author identification, idiolect, and linguistic uniqueness', *Applied linguistics*, 24: 431 - 47.

Coulthard, Malcolm, Alison Johnson, and David Wright. 2017. *An introduction to Forensic Linguistics: Language in Evidence* (Routledge: Abingdon, Oxon; New York, NY.).

Coulthard, Malcolm, and Rui Sousa-Silva. 2016. "Forensic Linguistics." In *What are Forensic Sciences? Concepts, Scope and Future Perspectives*, edited by Ricardo Jorge Dinis-Oliveira and Teresa Magalhães. Lisbon: Pactor.

Cramer, Duncan, and Dennis Laurence Howitt. 2004. *The Sage dictionary of statistics: a practical resource for students in the social sciences* (Sage).

Cronin, Blaise. 2001. 'Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices?', *Journal of the American Society for Information Science and Technology*, 52: 558-69.

Crystal, David, and Pavle Ivić. 2014. "Dialect." In *Encyclopædia Britannica*. Encyclopædia Britannica, inc.

Cuéllar-González, Álvaro, and Germán Vega García-Luengos. 2017. 'ETSO: Estilometría aplicada al Teatro del Siglo de Oro', Accessed August 8. estilometriatso.com.

Currie, Haver C. 1952. 'A projection of socio-linguistics: The relationship of speech to social status', *Southern Journal of Communication*, 18: 28-37.

Davies, Mark. 2002. 'In Corpora List Archive" Legal aspects of corpora compiling', *Online at http://torvald.aksis.uib.no/corpora/2002-4/0016.html*.

Day, Robert A, Nancy Sakaduski, and Nancy Day. 2011. *Scientific English: A guide for scientists and other professionals* (ABC-CLIO).

De Beaugrande, Robert. 1999. 'Linguistics, Sociolinguistics, and Corpus Linguistics: Ideal Language Versus Real Language', *Journal of Sociolinguistics*, 3: 128-39.

DGE. n.d. 'διάλεκτος. In Diccionario Griego–Español DGE en línea - http://dge.cchs.csic.es/xdge/'. http://dge.cchs.csic.es/xdge/%CE%B4%CE%B9%E1%BD%B1%CE%BB%CE%B5%CE%BA%CF%84%CE%BF%CF%82 on August 15, 2020.

Dicciogriego. n.d.-a. '*ἴδιος*. In Diccionario didáctico interactivo griego ↔ español - Dicciogriego.es'. https://www.dicciogriego.es/index.php#formas?lema=513&n=18801&forma=18801 on August 15, 2020.

Dicciogriego. n.d.-b. '*λεκτος*. In Diccionario didáctico interactivo griego ↔ español - Dicciogriego.es'. https://www.dicciogriego.es/index.php#formas?lema=638&n=5729&forma=5729 on August 15, 2020.

Dittmar, Norbert. 1996. 'Explorations in'Idiolects'.' in Robin Sackmann and Monika Budde (eds.), *Theoretical linguistics and grammatical description : papers in honour of Hans-Heinrich Lieb on the occasion of his 60th birthday* (John Benjamins Publishing Company: Amsterdam/Philadelphia).

Dörnyei, Zoltán. 2005. *The psychology of the language learner: Individual differences in second language acquisition* (Lawrence Erlbaum Associates, Inc.: New Jersey).

Eades, Diana. 2008. *Courtroom talk and neocolonial control* (Walter de Gruyter: New York).

Eades, Diana, Helen Fraser, Jeff Siegel, Tim McNamara, and Brett Baker. 2003. 'Linguistic identification in the determination of nationality: A preliminary report', *Language policy*, 2: 179-99.

Eckert, Penelope. 1989. 'The whole woman: Sex and gender differences in variation', *Language Variation and Change*, 1: 245-67.

Eckert, Penelope, and Sally McConnell-Ginet. 2003. *Language and gender* (Cambridge University Press).

Eddington, David. 2016. *Statistics for linguists: A step-by-step guide for novices* (Cambridge Scholars Publishing).

Ede, Lisa. 1990. *Singular Texts/Plural Authors: Perspectives on Collaborative Writing* (Southern Illinois University Press Carbondale & Edwardsville).

El Bouanani, Sara El Manar, and Ismail Kassou. 2014. 'Authorship analysis studies: A survey', *International Journal of Computer Applications*, 86: 22 - 29.

ElMalik, Abdullahi Tambul, and Hilary Nesi. 2008. 'Publishing research in a second language: The case of Sudanese contributors to international medical journals', *Journal of English for Academic Purposes*, 7: 87-96.

Emerson, Jason. 2006. 'America's most famous letter', *American Heritage*.

Erdman, David Vorse, and Ephim Gregory Fogel. 1966. *Evidence for Authorship: Essays on Problems of Attribution, with an Annotated Bibliography of Selected Readings* (Cornell University Press: Ithaca: New York).

Eurostat. 2018. 'What languages are studied the most in the EU? ', European Commission. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20200925-1.

Farrel, Joseph. 2003. 'Classical Genre in Theory and Practice', *New Literary History*, 34: 383-408.

FCT. 2018. "EVALUATION GUIDE R&D UNITS EVALUATION 2017-18." In.: The Portuguese Science and Technology Foundation (FCT)

Feldman, Ronen. 2013. 'Techniques and applications for sentiment analysis', *Communications of the ACM*, 56: 82-89.

Ferguson, Charles A. 1994. 'Dialect, register, and genre: Working assumptions about conventionalization', *Sociolinguistic perspectives on register*: 15-30.

Ferrando, I.N. 2006. 'On the meaning of three English prepositions', *In-roads of Language: Essays in English Studies*, 25: 167.

Ferreira, Manuela Barros, Ernestina Carrilho, Maria Lobo, João Saramago, and Luísa Segura Cruz. 1996. 'Variação linguística: perspectiva dialectológica.' in Isabel Hub Faria, Emília R. Pedro, Inês Duarte and Carlos A. M. Gouveia (eds.), *Introdução à Linguística Geral e Portuguesa* (Caminho: Lisboa).

Ferro, Nicola, and Carol Peters. 2019. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF* (Springer).

Firth, J.R. 1962. 'A Synopsis of Linguistic Theory, 1930-1955.' in, *Studies in Linguistic Analysis* (Oxford: Blackwell).

Fleary, F.G. 1874. 'On metrical tests applied to dramatic poetry, I. Shakespeare.', *New Shakespearean Society Transaction*, Series 1: 1–16, 38–39.

Flowerdew, Lynne. 2004. 'The argument for using English specialized corpora to understand academic and professional language', *Discourse in the professions: Perspectives from corpus linguistics*, 11: 33.

Fornaciari, Tommaso, and Massimo Poesio. 2011. "Lexical vs. surface features in deceptive language analysis." In *Proceedings of the ICAIL 2011 Workshop: Applying Human Language Technology to the Law*, 2-8. University of Pittsburgh School of Law Pittsburgh, PA.

Fornaciari, Tommaso, and Massimo Poesio. 2012. "On the use of homogenous sets of subjects in deceptive language analysis." In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, 39-47.

Foroodi-Nejad, Farzaneh, and Johanne Paradis. 2009. 'Crosslinguistic transfer in the acquisition of compound words in Persian-English bilinguals'.

Foucault, Michel. 1979. 'Authorship: what is an author?', *Screen*, 20: 13-34.

Foulkes, Paul, Peter French, and Kim Wilson. 2019. 'LADO as Forensic Speaker Profiling.' in, *Language Analysis for the Determination of Origin* (Springer).

Frello, Birgitta. 2013. 'Cultural Hybridity.' in Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics* (Blackwell Publishing Ltd.).

Furnival, F.J. 1887. *Introduction to the Leopold Shakespeare* (Cassel & Co.: London).

Galve, Ignacio Guillén. 1998. 'The textual interplay of grammatical metaphor on the nominalizations occurring in written medical English', *Journal of Pragmatics*, 30: 363-85.

Gayle, Alberto Alexander, and Motomu Shimaoka. 2017. 'Evaluating the lexico-grammatical differences in the writing of native and non-native speakers of English in peer-reviewed medical journals in the field of pediatric oncology: Creation of the genuine index scoring system', *PloS one*, 12.

Gayo, Iria Del Río, Marcos Zampieri, and Shervin Malmasi. 2018. "A Portuguese native language identification dataset." In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, 291-96.

'GE - Portuguese Journal of Gastroenterology'. 2019. www.karger.com/PJG, Accessed October 25. https://www.karger.com/Journal/Home/272027.

Ghosh, Dhiren, and Andrew Vogt. 2012. "Outliers: An evaluation of methodologies." In *Joint statistical meetings*.

Gibbons, John, and M Teresa Turell. 2008. *Dimensions of forensic linguistics* (John Benjamins Publishing).

Glance, Natalie, Matthew Hurst, Kamal Nigam, Matthew Siegler, Robert Stockton, and Takashi Tomokiyo. 2005. "Deriving marketing intelligence from online discussion." In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 419-28.

Goffman, Erving. 1981. *Forms of talk* (University of Pennsylvania Press).

Goldin, Gili, Ella Rabinovich, and Shuly Wintner. 2018. "Native language identification with user generated content." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3591-601.

González-Alcaide, Gregorio, Juan Carlos Valderrama-Zurián, and Rafael Aleixandre-Benavent. 2012. 'The impact factor in non-English-speaking countries', *Scientometrics*, 92: 297-311.

Goswami, Sumit, Sudeshna Sarkar, and Mayur Rustagi. 2009. "Stylometric analysis of bloggers' age and gender." In *Third international AAAI conference on weblogs and social media*.

Grant, Tim. 2007. 'Quantifying evidence in forensic authorship analysis', *International Journal of Speech, Language & the Law*, 14.

Grant, Tim. 2008. 'Approaching questions in forensic authorship analysis.' in, *Dimensions of forensic linguistics*.

Grant, Tim. 2010. 'Text messaging forensics Txt 4n6: Idiolect free authorship analysis?' in, *The Routledge handbook of forensic linguistics* (Routledge).

Grant, Tim. 2013. 'TXT 4N6: method, consistency, and distinctiveness in the analysis of SMS text messages', *JL & Pol'y*, 21: 467.

Grant, Tim, and Nicci MacLeod. 2018a. 'Resources and constraints in linguistic identity performance: A theory of authorship', *Language and Law/Linguagem e Direito*, 5: 80 - 96.

Grant, Tim, and Nicola MacLeod. 2018b. 'Resources and constraints in linguistic identity performance–a theory of authorship', *Language and Law/Linguagem e Direito*, 5: 80-96.

Gries, Stefan Th. 2013. '50-something years of work on collocations: What is or should be next …', *International Journal of Corpus Linguistics*, 18: 137-66.

Grieve, Jack, Isobelle Clarke, Emily Chiang, Hannah Gideon, Annina Heini, Andrea Nini, and Emily Waibel. 2018. 'Attributing the Bixby Letter using n-gram tracing', *Digital Scholarship in the Humanities*, 34: 493-512.

Gross, Alan G, Joseph E Harmon, Michael Reidy, and Michael S Reidy. 2002. *Communicating science: The scientific article from the 17th century to the present* (Oxford University Press on Demand).

Gudjonsson, Gisli H. 1993. 'Confession evidence, psychological vulnerability and expert testimony', *Journal of community & applied social psychology*, 3: 117-29.

Halliday, M. A. K., Angus McIntosh, and Peter Strevens. 1964. *The linguistic sciences and language teaching* (Longman: London).

Halliday, M.A.K. 1989. 'PART A.' in F. Christie (ed.), *Language, context, and text: Aspects of language in a social-semiotic perspective* (Oxford University Press: Oxford).

Hamel, Rainer Enrique. 2007. 'The dominance of English in the international scientific periodical literature and the future of language use in science', *Aila Review*, 20: 53-71.

Hazen, Kirk. 2007. 'The study of variation in historical perspective', *Sociolinguistic variation: Theories, methods, and applications*: 70-89.

Henning, Andersen. 1989. 'Markedness: The First 150 Years', *Markedness in Synchrony and Diachrony, Olga M. Tomic (ed.), Mouton de Gruyter, Berlin–Germany*: 11-46.

Hoaglin, David C, and Boris Iglewicz. 1987. 'Fine-tuning some resistant rules for outlier labeling', *Journal of the American Statistical Association*, 82: 1147-49.

Hockett, Charles F. 1958. *A course in modern linguistics* (Macmillan: New York).

Hodson, Thomas C. 1939. 'Sociolinguistics in India', *Man in India*, 19: 94-98.

Hoey, M. 2005. *Lexical priming: A new theory of words and language* (Routledge: London).

Holmes, David I. 1994. 'Authorship attribution', *Computers and the Humanities*, 28: 87-106.

Holmes, David I. 1998. 'The evolution of stylometry in humanities scholarship', *Literary and Linguistic Computing*, 13: 111-17.

Holmes, Frederic L. 1987. 'Scientific writing and scientific discovery', *Isis*, 78: 220-35.

Holtzman, Nicholas S, Allison M Tackman, Angela L Carey, Melanie S Brucks, Albrecht CP Küfner, et al. 2019. 'Linguistic markers of grandiose narcissism: A LIWC analysis of 15 samples', *Journal of Language and Social Psychology*, 38: 773-86.

Horowitz, Harold W, Nicholas H Fiebach, Stuart M Levitz, Jo Seibel, Edwin H Smail, et al. 1996. 'Ode to multiauthorship: a multicentre, prospective random poem', *The Lancet*, 348: 1746.

Hu, Guangwei, and Feng Cao. 2011. 'Hedging and boosting in abstracts of applied linguistics articles: A comparative study of English-and Chinese-medium journals', *Journal of pragmatics*, 43: 2795-809.

Hudson, Richard A. 1996. *Sociolinguistics* (Cambridge university press).

Hyland, Ken. 2004. *Disciplinary Discourses. Social Interactions in Academic Writing* (University of Michigan Press).

Ingram, John K. 1874. 'On the 'Weak Endings' of Shakespere', *Transactions of the New Shakespere Society*, Series 1: 442–46.

Iskander, John K, Sara Beth Wolicki, Rebecca T Leeb, and Paul Z Siegel. 2018. 'Successful Scientific Writing and Publishing: A Step-by-Step Approach', *Preventing chronic disease*, 15.

Isurin, Ludmila. 2005. "Cross linguistic transfer in word order: Evidence from L1 forgetting and L2 acquisition." In *Proceedings of the 4th International Symposium on Bilingualism*, 1115-30.

Jakobson, Roman. 1971. 'Results of a Joint Conference of Anthropologists and Linguists.' in, *SELECTED WRITINGS* (MOUTON & CO.: The Hague, THE NETHERLANDS.).

Jarvis, Scott. 2000. 'Methodological rigor in the study of transfer: Identifying L1 influence in them interlanguage lexicon', *Language Learning*, 50: 245-309.

Jarvis, Scott. 2007. 'Theoretical and methodological issues in the investigation of conceptual transfer', *Vigo International Journal of Applied Linguistics*: 43-71.

Jarvis, Scott. 2010. 'Comparison-based and detection-based approaches to transfer research', *Eurosla yearbook*, 10: 169-92.

Jarvis, Scott. 2012a. 'Crosslinguistic influence and multilingualism', *The encyclopedia of applied linguistics*: 1554-62.

Jarvis, Scott. 2012b. 'The detection-based approach: An overview', *Approaching language transfer through text classification: Explorations in the detection-based approach*: 1-33.

Jarvis, Scott, and Scott A Crossley. 2012. *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach* (Multilingual Matters).

Jarvis, Scott, and Magali Paquot. 2015. 'Learner corpora and native language identification.' in S. Granger, G. Gilquin and F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research (Cambridge Handbooks in Language and Linguistics).* (Cambridge University Press.: Cambridge).

Jarvis, Scott, and Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition* (Routledge).

Jiang, Feng Kevin, and Ken Hyland. 2020a. 'Prescription and reality in advanced academic writing', *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)*: 14-42.

Jiang, Feng Kevin, and Ken Hyland. 2020b. '"There are significant differences…": the secret life of existential there in academic writing', *Lingua*, 233: 102758.

Johns, Ann M, and Tony Dudley-Evans. 1991. 'English for specific purposes: International in scope, specific in purpose', *Tesol Quarterly*, 25: 297-314.

Johnson, A., and D. Wright. 2014. 'Identifying idiolect in forensic authorship attribution: An n-gram textbite approach', *Language and Law/Linguagem e Direito*, 1: 37 - 69.

Johnson, Alison, and Malcolm Coulthard. 2010. 'Introduction Current debates in forensic linguistics.' in Alison Johnson and Malcolm Coulthard (eds.), *The Routledge handbook of forensic linguistics* (Routledge: Abingdon, Oxford. UK).

Johnson, Rob, Anthony Watkinson, and Michael Mabe. 2018. 'The STM Report: An overview of scientific and scholarly publishing', *International Association of Scientific, Technical and Medical Publishers*.

Johnstone, Barbara. 1996. *The linguistic individual: Self-expression in language and linguistics* (Oxford Univ. Press: Oxford).

Juola, Patrick. 2007. "Future trends in authorship attribution." In *IFIP International Conference on Digital Forensics*, 119-32. Springer.

Juola, Patrick. 2008. 'Authorship attribution', *Foundations and Trends® in Information Retrieval*, 1: 233-334.

Juola, Patrick. 2013. "How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling. Author of the Harry Potter books has a distinct linguistic signature." In *Scientific American*.

Juola, Patrick. 2015. 'Industrial uses for authorship analysis', *Mathematics and Computers in Sciences and Industry*: 21-25.

Kachru, B.B. 1985. 'Standards, codification and sociolinguistic realism: The English language in the Outer Circle.' in R. Quirk. and H. Widdowson. (eds.), *English in the world: Teaching and learning the language and literatures* (Cambridge University Press.: Cambridge).

Kachru, B.B. 1992. 'World Englishes: Approaches, issues and resources', *Language Teaching*, 25: 1-14.

Kachru, B.B. 1997. 'World Englishes and English-using communities', *Annual review of applied linguistics*, 17: 66-87.

Kafes, Hüseyin. 2018. 'A Genre Analysis of English and Turkish Research Article Introductions', *Novitas-ROYAL (Research on Youth and Language)*, 12: 66-79.

Kallet, Richard H. 2004. 'How to write the methods section of a research paper', *Respiratory care*, 49: 1229-32.

Kay, Martin. 1997. 'The proper place of men and machines in language translation', *machine translation*, 12: 3-23.

King, Christopher. 2012. 'Multiauthor papers: onward and upward', *Science Focus*, 7: 62-64.

Koerner, Konrad. 1991. 'Toward a history of modern sociolinguistics', *American Speech*, 66: 57-70.

Koester, Almut. 2010. 'Building small specialised corpora', *The Routledge handbook of corpus linguistics*, 1: 66-79.

Kredens, Krzysztof, Ria Perkins, and Tim Grant. 2020. 'Developing a framework for the explanation of interlingual features for native and other language influence detection', *Language and Law= Linguagem e Direito*, 6: 10-23.

Kripke, Saul A. 1982. *Wittgenstein on rules and private language: An elementary exposition* (Harvard University Press).

Kurdi, M Zakaria. 2019. "Content-Dependent Versus Content-Independent Features for Gender and Age Range Identification in Different Types of Texts." In *The Thirty-Second International Flairs Conference*.

Kyle, Kristopher, Scott A Crossley, and You Jin Kim. 2015. 'Native language identification and writing proficiency', *International Journal of Learner Corpus Research*, 1: 187-209.

Labov, William. 1963. 'The social motivation of a sound change', *Word*, 19: 273-309.

Labov, William. 1966 [2006]. *The social stratification of English in New York city* (Cambridge University Press).

Labov, William. 1972. *Sociolinguistic patterns* (Univ. of Pennsylvania: Philadelphia).

Labov, William. 1989. 'The exact description of the speech community: Short 'a' in Philadelphia.' in Ralph W Fasold and Deborah Schiffrin (eds.), *Language change and variation* (John Benjamins Publishing).

Labov, William. 1990. 'The intersection of sex and social class in the course of linguistic change', *Language Variation and Change*, 2: 205 - 54.

Langefeld, Carl D., Hannah C. Ainsworth, Deborah S. Cunninghame Graham, Jennifer A. Kelly, Mary E. Comeau, et al. 2017. 'Transancestral mapping and genetic load in systemic lupus erythematosus', *Nature Communications*, 8: 16021.

Le Page, Robert Brock, and Andrée Tabouret-Keller. 1985. *Acts of identity: Creole-based approaches to language and ethnicity* (Cambridge University Press).

Lee, Jerry Won, and Suresh Canagarajah. 2021. 'Translingualism and World Englishes', *Bloomsbury World Englishes Volume 1: Paradigms*: 99.

Leung, Wilson, Christopher D. Shaffer, Laura K. Reed, Sheryl T. Smith, William Barshop, et al. 2015. '<em>Drosophila</em> Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million Years of Evolution', *G3: Genes|Genomes|Genetics*, 5: 719-40.

Levsky, Marc E, Alex Rosin, Troy P Coon, William L Enslow, and Michael A Miller. 2007. 'A descriptive analysis of authorship within medical journals, 1995-2005', *Southern Medical Journal*, 100: 371-76.

Litvinova, Tatiana, Olga Zagorovskaya, Olga Litvinova, and Pavel Seredin. 2016. "Profiling a set of personality traits of a text's author: a corpus-based approach." In *International Conference on Speech and Computer*, 555-62. Springer.

Liu, Fei, Julien Perez, and Scott Nowson. 2016. 'A language-independent and compositional model for personality trait recognition from short texts', *arXiv preprint arXiv:1610.04345*.

López-Pellisa, Teresa, Neus Rotger, and Fernando Rodríguez-Gallego. 2020. 'Collaborative writing at work: Peer feedback in a blended learning environment', *Education and Information Technologies*: 1-18.

Lord, Robert D. 1958. 'Studies in the history of probability and statistics. VIII. De Morgan and the Statistical study of literary style', *Biometrika*, 45: 282-82.

Love, Harold. 2002. *Attributing authorship: An introduction* (Cambridge University Press).

MacLeod, Nicci, and Tim Grant. 2017. '"go on cam but dnt be dirty": linguistic levels of identity assumption in undercover online operations against child sex abusers', *Language and Law= Linguagem e Direito*, 4: 157-75.

Maia, Belinda Mary Harper Sousa. 1997. "Do-it-yourself corpora... with a little bit of help from your friends!" In *PALC '97 Practical applications in Language Corpora*.

Malmasi, Shervin, and Mark Dras. 2017. 'Multilingual native language identification', *Natural Language Engineering*, 23: 163-215.

Malmasi, Shervin, and Mark Dras. 2018. 'Native language identification with classifier stacking and ensembles', *Computational linguistics*, 44: 403-46.

Malone, Edmond. 1787. *A dissertation on the three parts of King Henry VI. tending to shew that those plays were not written originally by Shakspeare* (from the Press of Henry Baldwin).

Marko, Karoline. 2021. 'Exploring the Distinctiveness of Emoji Use for Digital Authorship Analysis', *Language and Law/Linguagem e Direito*, 7: 36-55.

Marquilhas, Rita. 2013. 'Fenómenos de Mudança na História do Português.' in Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota, Luísa Seguro and Amália Mendes (eds.), *Gramática do Português* (Fundação Calouste Gulbenkian).

Martín, Francisco Miguel Martínez. 1983. *Fonética y sociolingüística en la ciudad de Burgos* (Editorial CSIC-CSIC Press).

Mascol, Conrad. 1888. 'Curves of pauline and pseudo-pauline style i', *Unitarian Review*, 30: 453-60.

Matos, Gabriela, and Eduardo Buzaglo Paiva Raposo. 2013. 'Estructuras de coordenação.' in, *Gramática do Português* (Fundação Calouste Gulbenkian: Lisboa).

Mauranen, Anna, Carmen Pérez-Llantada, and John M Swales. 2010. 'Academic Englishes: A standardized knowledge?' in, *The Routledge handbook of world Englishes* (Routledge).

May, Alison, Rui Sousa-Silva, and Malcolm Coulthard. 2021. 'Introduction.' in Malcolm Coulthard, Alison May and Rui Sousa-Silva (eds.), *The Routledge handbook of forensic linguistics* (Routledge).

McKnight, Larry, and Padmini Srinivasan. 2003. "Categorization of sentence types in medical abstracts." In *AMIA Annual Symposium Proceedings*, 440. American Medical Informatics Association.

McMenamin, Gerald R. 2002. *Forensic linguistics: Advances in forensic stylistics* (CRC press).

McMenamin, Gerald R. 2010. 'Forensic stylistics. Theory and practice of forensic stylistics.' in, *The Routledge handbook of forensic linguistics* (Routledge).

Mendenhall, Thomas Corwin. 1887. 'The characteristic curves of composition', *Science*, 9: 237-49.

Mendenhall, Thomas Corwin. 1901. 'A mechanical solution to a literary problem', *Popular Science Monthly*, 60: 97-105.

Methven, Elyse Patricia. 2017. 'Dirty talk: A critical discourse analysis of offensive language crimes', *University of Technology Sydney*.

Michel, L, and W Ceelen. 2007. 'Twelve Steps to Writing an Effective "Materials and Methods"-section', *Acta chir belg*, 107: 102.

Migueláñez, Daniel. 2019. 'El Holmes de la filología. Germán Vega "devuelve" La monja alférez a su verdadero autor', *CHAMBERÍ - SUPLEMENTO CULTURAL,* July 7, pp. 8.

Miller, Carolyn R. 1984. 'Genre as social action', *Quarterly journal of speech*, 70: 151-67.

Miranda, José Carlos Ribeiro. 2011. 'Será Afonso, o Sábio, o autor anónimo de A 36/A 39?', *Seminário Medieval 2009-2011*.

Miranda, José Carlos Ribeiro. 2016. 'Calheiros, Sandim e Bonaval: uma rapsódia «de amigo»', *Guarecer. Revista Electrónica de Estudos Medievais*, 1: 47-62.

Mohan, Shailendra. 2004. 'Development of Sociolinguistic Studies at The Deccan College', *Bulletin of the Deccan College Research Institute*, 64: 261-69.

Moreno, Ana I, and John M Swales. 2018. 'Strengthening move analysis methodology towards bridging the function-form gap', *English for Specific Purposes*, 50: 40-63.

Mosteller, Frederick, and David L Wallace. 1963. 'Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers', *Journal of the American Statistical Association*, 58: 275-309.

Mosteller, Frederick, and David L. Wallace. 1964. *Inference and disputed authorship: The Federalist* (Addison-Wesley: Reading, MA).

Nayar, P Bhaskaran. 1997. 'ESL/EFL dichotomy today: Language politics or pragmatics?', *Tesol Quarterly*, 31: 9-37.

Neuendorf, Kimberly A, and Anup Kumar. 2015. 'Content analysis', *The international encyclopedia of political communication*: 1-10.

Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "" How old do you think I am?" A study of language and age in Twitter." In *Seventh International AAAI Conference on Weblogs and Social Media*.

Nini, Andrea. 2014. 'Authorship profiling in a forensic context', Aston University.

Nini, Andrea. 2018. 'Developing forensic authorship profiling', *Language and Law/Linguagem e Direito*, 5: 38-58.

Nini, Andrea, and Tim Grant. 2013. 'Bridging the gap between stylistic and cognitive approaches to authorship analysis using Systemic Functional Linguistics and multidimensional analysis', *International Journal of Speech, Language & the Law*, 20.

Noorizadeh-Honami, Leila, and Azizeh Chalak. 2018. 'Comparative Analysis of Architecture Research Article Abstracts Written by Native and Non-native Authors: A Cross-linguistic, Cross-cultural Study', *Theory and Practice in Language Studies*, 8: 325-30.

Oberlander, Jon, and Scott Nowson. 2006. "Whose thumb is it anyway? Classifying author personality from weblog text." In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 627-34.

Odlin, Terence. 1989. *Language transfer* (Cambridge University Press Cambridge).

Odlin, Terence. 2005. 'Crosslinguistic influence and conceptual transfer: What are the concepts?', *Annual review of applied linguistics*, 25: 3.

Olsson, John. 2008. *Forensic Linguistics: Second Edition* (Continuum International Publishing Group).

Pan, Fan, Randi Reppen, and Douglas Biber. 2016. 'Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals', *Journal of English for Academic Purposes*, 21: 60-71.

Paquot, Magali. 2013. 'Lexical bundles and L1 transfer effects', *International Journal of Corpus Linguistics*, 18: 391-417.

Patrick, Peter L. 2019. 'Language Analysis for the Determination of Origin (LADO): An Introduction.' in Peter L Patrick, Monika S Schmid and Karin Zwaan (eds.), *Language Analysis for the Determination of Origin: Current Perspectives and New Directions* (Springer).

Paul, Hermann. 1890. *Principles of the history of language* (Longmans, Green, and Co.: London).

Pavelec, Daniel, Edson Justino, Leonardo V Batista, and Luiz S Oliveira. 2008. "Author identification using writer-dependent and writer-independent strategies." In *Proceedings of the 2008 ACM symposium on Applied computing*, 414-18. ACM.

Pavlenko, Aneta. 2000. 'L2 influence and L1 attrition in adult bilingualism.' in Monika S. Schmid, Barbara Köpke, Merel Keijzer and LinaWeilemar (eds.), *First Language Attrition Interdisciplinary perspectives on methodological issues* (John Benjamins Publishing Company: Amsterdam/Philadelphia).

Pavlenko, Aneta, and Scott Jarvis. 2002. 'Bidirectional transfer', *Applied linguistics*, 23: 190-214.

Peersman, Claudia, Walter Daelemans, and Leona Van Vaerenbergh. 2011. "Predicting age and gender in online social networks." In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 37-44.

Penco, Carlo. 2007. "Idiolect and context." In *Library of Living Philosphers: the Philosophy of Michael Dummett*, edited by L. E. Hahn. Open Court.

Pennebaker, James W. 2013. 'How authors' forgettable words reveal their personality and social behaviors', *Information Design Journal (IDJ)*, 20.

Pennebaker, James W, Matthias R Mehl, and Kate G Niederhoffer. 2003. 'Psychological aspects of natural language use: Our words, our selves', *Annual review of psychology*, 54: 547-77.

Pérez-Llantada, Carmen. 2012. *Scientific discourse and the rhetoric of globalization: The impact of culture and language* (A&C Black).

Perkins, Ria. 2014. 'Linguistic identifiers of L1 Persian speakers writing in English: NLID for authorship analysis', Aston University.

Perkins, Ria. 2015. 'Native language identification (NLID) for forensic authorship analysis of weblogs.' in, *New threats and countermeasures in digital crime and cyber terrorism* (IGI Global).

Perkins, Ria, and Tim Grant. 2018. 'Native language influence detection for forensic authorship analysis: Identifying L1 Persian bloggers', *International Journal of Speech, Language & the Law*, 25.

Petrenko, Anton. 2006. 'Idiolect and common language', University of Ottawa (Canada).

Piqué-Angordans, Jordi, and MJ Coperías Aguilar. 1999. 'Verb tense in essay writing: frequency of use in health sciences', *Lenguas para Fines Específicos-VI. Investigación y enseñanza*: 245-51.

Prince, Michael B. 2003. 'Mauvais genres', *New Literary History*, 34: 453-79.

Pullum, Geoffrey K, and Rodney Huddleston. 2002. 'Adjectives and adverbs', *The Cambridge grammar of the English language*: 525-95.

Qian, Jing, Mai ElSherief, Elizabeth M Belding, and William Yang Wang. 2018. 'Leveraging intra-user and inter-user representation learning for automated hate speech detection', *arXiv preprint arXiv:1804.03124*.

Queralt, Sheila. 2014. 'Acerca de la prueba lingüística en atribución de autoría hoy', *Revista de Llengua i Dret*.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language* (Longman Group Limited: New York).

Raghavan, Sindhu, Adriana Kovashka, and Raymond Mooney. 2010. "Authorship attribution using probabilistic context-free grammars." In *Proceedings of the ACL 2010 conference short papers*, 38-42.

Raidt, Edith H. 1993. 'The role of women in Linguistic Change.' in, *Historical Linguistics 1989* (John Benjamins).

Reddy, T Raghunadha , B Vishnu Vardhan, and P Vijaypal Reddy. 2016. 'A survey on authorship profiling techniques', *International Journal of Applied Engineering Research*, 11: 3092-102.

Rieber, Robert W, and William A Stewart. 1990. 'The Interactions of the Language Sciences and the Law', *Annals of the New York Academy of Sciences*, 606: 1-4.

Rosch, Eleanor. 1978. 'Principles of Categorization.' in Eleanor Rosch and Barbara Bloom Lloyd (eds.), *Cognition and categorization* (Lawrence Erlbaum: Hillsdale, NJ.).

Roudometof, Victor. 2016. 'Theorizing glocalization: Three interpretations1', *European Journal of Social Theory*, 19: 391-408.

Rudnicka, Karolina. 2018. 'Variation of sentence length across time and genre', *Diachronic corpora, genre, and language change*: 220-40.

Salager-Meyer, Françoise. 1990. 'Discoursal flaws in medical English abstracts: A genre analysis per research-and text-type', *Text-Interdisciplinary journal for the study of discourse*, 10: 365-84.

Salager-Meyer, Françoise. 1992. 'A text-type and move analysis study of verb tense and modality distribution in medical English abstracts', *English for Specific Purposes*, 11: 93-113.

Salager-Meyer, Françoise. 1994. 'Hedges and textual communicative function in medical English written discourse', *English for Specific Purposes*, 13: 149-70.

Sanchez-Stockhammer, Christina. 2012. 'Hybridization in language.' in, *Conceptualizing cultural hybridization* (Springer).

Santini, Marina. 2004. 'State-of-the-art on automatic genre identification'.

Sapir, Edward. 1927. 'Speech as a personality trait', *American Journal of Sociology*, 32: 892-905.

Savoy, Jacques. 2020. *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling* (Springer: Neuchatel, Switzerland).

Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. "Effects of age and gender on blogging." In *AAAI spring symposium: Computational approaches to analyzing weblogs*, 199-205.

Scott, Mike. 2018a. 'WordSmith tools help', *Liverpool: Lexical Analysis Software*.

Scott, Mike. 2018b. 'WordSmith tools version 7', *Liverpool: Lexical Analysis Software*, 122.

Sebastiani, Fabrizio. 2002. 'Machine learning in automated text categorization', *ACM computing surveys (CSUR)*, 34: 1-47.

Selinker, Larry. 1972. 'Interlanguage', *IRAL-International Review of Applied Linguistics in Language Teaching*, 10: 209-32.

Selinker, Larry. 2014. 'Interlanguage 40 years on', *2014). Interlanguage: forty years later*: 229-63.

Silva, Rui Sousa, Gustavo Laboreiro, Luís Sarmento, Tim Grant, Eugénio Oliveira, and Belinda Maia. 2011. "'twazn me!!!;('automatic authorship analysis of micro-blogging messages." In *International Conference on Application of Natural Language to Information Systems*, 161-68. Springer.

Sinclair, John. 1991. *Corpus, Concordance and Collocation* (Oxford University Press).

Sinclair, John. 2004. "Corpus and Text — Basic Principles." In *Developing Linguistic Corpora: a Guide to Good Practice*. Tuscan Word Centre.

Sinha, Avanika, Niroj Banerjee, Ambalika Sinha, and Rajesh Kumar Shastri. 2009. 'Interference of first language in the acquisition of second language', *International Journal of Psychology and Counselling*, 1: 117-22.

Solan, Lawrence M. 2013. 'Intuition versus Algorithm: The Case of Forensic Authorship Attribution', *Brooklyn Journal of Law and Policy*, 21.

Sosa-Napolskij, Milaydis. 2021, March 26. 'CoRA - Comparative Corpora of Research Articles', *Zenodo*.

Sousa-Silva, Rui. 2013. 'Detecting plagiarism in the forensic linguistics turn', Aston University.

Sousa-Silva, Rui. 2014. 'Detecting translingual plagiarism and the backlash against translation plagiarists', *Language and Law= Linguagem e Direito*, 1.

Sousa-Silva, Rui. 2019. 'Plagiarism Across Languages and Cultures: A (Forensic) Linguistic Analysis', *Handbook of the Changing World Language Map*: 2325-45.

Sousa-Silva, Rui. Forthcoming 2021. 'LINGUÍSTICA FORENSE NO COMBATE E PREVENÇÃO DO CIBERCRIME.' in I.M.E.S. Guedes and M.A. Melo Gomes (eds.), *CRIMINALIDADE – NOVOS DESAFIOS, OFENSAS E SOLUÇÕES* (PACTOR).

Sousa-Silva, Rui, and Bruna Batista Abreu. 2015. 'Plágio: um problema forense', *Language and Law= Linguagem e Direito*, 2.

Sousa-Silva, Rui, Luıs Sarmento, Tim Grant, Eugénio Oliveira, and Belinda Maia. 2010. "Comparing sentence-level features for authorship analysis in Portuguese." In *9th INTERNATIONAL CONFERENCE PROPOR*.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. 2000. 'Automatic text categorization in terms of genre and author', *Computational linguistics*, 26: 471-95.

Stirling, Lesley, and Rodney Huddleston. 2002. 'Deixis and anaphora.' in Rodney Huddleston and Geoffrey K Pullum (eds.), *The Cambridge grammar of the English language* (Language. Cambridge: Cambridge University Press: Cambridge).

Strevens, Peter. 1977. 'Special-purpose language learning: A perspective', *Language Teaching*, 10: 145-63.

Strevens, Peter. 1982. 'III. WORLD ENGLISH AND THE WORLD'S ENGLISHES-OR, WHOSE LANGUAGE IS IT ANYWAY?', *Journal of the Royal Society of Arts*, 130: 418-31.

Svartvik, Jan. 1968. 'The Evans Statements. A Case for Forensic Linguistics.' in Alvar Ellegård (ed.), *Acta Universitatis Gothoburgensis* (University of Gothenburg).

Swales, John. 1990. *Genre analysis: English in academic and research settings* (Cambridge University Press).

Swales, John M. 2005. 'Attended and unattended “this” in academic writing: A long and unfinished story', *ESP Malaysia*, 11: 1-15.

Swales, John M. 2004. *Research Genres: Explorations and Applications* (Cambridge University Press: Cambridge).

Sweller, John. 1988. 'Cognitive load during problem solving: Effects on learning', *Cognitive science*, 12: 257-85.

Tetreault, Joel, Daniel Blanchard, and Aoife Cahill. 2013. "A report on the first native language identification shared task." In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, 48-57.

Tomokiyo, Laura Mayfield, and Rosie Jones. 2001. "You're not from'round here, are you?: naive Bayes detection of non-native utterance text." In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1-8. Association for Computational Linguistics.

Trudgill, Peter. 1999. 'Standard English: What it isn’t', *Standard English: the widening debate*: 117-28.

Tuldava, Juhan. 2004. 'The development of statistical stylistics (a survey)', *Journal of Quantitative Linguistics*, 11: 141-51.

Turell, M. Teresa. 2010. 'The use of textual, grammatical and sociolinguistic evidence in forensic text comparison', *International Journal of Speech, Language & the Law*, 17.

Turell, M. Teresa. 2013. 'Presidential address.' in Rui Sousa-Silva, Rita Faria, Núria Gavaldà, Belinda Maia and Rui Effe (eds.), *Bridging de Gap(s) between Language and the Law: Proceedings of the 3rd European Conference of the International Association of Forensic Linguists* (Faculdade de Letras da Universidade do Porto: Porto).

Turell, M. Teresa, and Nuria Gavaldà. 2013. 'Towards an Index of Idiolectal Similitude (Or Distance) In Forensic Authorship Analysis', *Journal of Law and Policy*, 21: 10.

'Ulrich's Periodicals Directory - Ulrichsweb'. Accessed October 25. https://www.ulrichsweb.com/ulrichsweb/faqs.asp.

"UNESCO Science Report: Towards 2030." In. 2016. 794pp. France: United Nations Educational, Scientific and Cultural Organization.

UNHCR. 1951. "Convention relating to the status of refugees." In, edited by United Nations High Commissioner for Refugees. Geneva: United Nations High Commissioner for Refugees: Communication and Public Information Service.

Utiyama, Masao, and Hitoshi Isahara. 2007. "A comparison of pivot methods for phrase-based statistical machine translation." In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 484-91.

Veado, Rosa Maria Assis. 1983. 'Redução do ditongo: uma variável sociolinguística', *Cadernos de Linguística e Teoria da Literatura*, 5: 208-29.

Vinciarelli, Alessandro, and Gelareh Mohammadi. 2014. 'A survey of personality computing', *IEEE Transactions on Affective Computing*, 5: 273-91.

Weinreich, Uriel. 1951. 'Research problems in bilingualism, with special regard to Switzerland', *Diss. Columbia U*.

Weinreich, Uriel. 1953. "Languages in Contact: Finding and Problems." In. Netherlands: The Hague: Mounton Publishers.

Weinreich, Uriel. 1957. 'Functional aspects of Indian bilingualism', *Word*, 13: 203-33.

Weinreich, Uriel, William Labov, and Marvin I Herzog. 1968. 'Empirical foundations for a theory of language change', *WP Lehmann-Y. Malkiel (Hrsgg.), Directions for Historical Linguistics, Austin/London*.

Weisser, Martin. 2016. *Practical corpus linguistics: An introduction to corpus-based language analysis* (John Wiley & Sons).

Weren, Edson RD, Anderson U Kauer, Lucas Mizusaki, Viviane P Moreira, J Palazzo M de Oliveira, and Leandro K Wives. 2014. 'Examining multiple features for author profiling', *Journal of information and data management*, 5: 266-66.

Williams, Carrington Bonsor. 1970. *Style and vocabulary: numerical studies* (Griffin).

Williams, Ian A. 1996. 'A contextual study of lexical verbs in two types of medical research report: Clinical and experimental', *English for Specific Purposes*, 15: 175-97.

Wittgenstein, Ludwig. 1986 [1953]. *Philosophical Investigations* (Blackwell Publishers: Oxford).

Wong, Sze-Meng Jojo, and Mark Dras. 2011. "Exploiting parse structures for native language identification." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1600-10. Association for Computational Linguistics.

Wood, Alistair. 2001. 'International scientific English: The language of research scientists around the world', *Research perspectives on English for academic purposes*, 71: 83.

Woolston, Chris. 2015. 'Fruit-fly paper has 1,000 authors', *Nature News*, 521: 263.

Wray, Alison. 2017. 'Formulaic Sequences as a Regulatory Mechanism for Cognitive Perturbations During the Achievement of Social Goals', *Top Cogn Sci*, 9: 569-87.

Wright, David. 2014. 'Stylistics versus Statistics: A corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails', University of Leeds.

Wu, Hua, and Haifeng Wang. 2009. "Revisiting pivot language approach for machine translation." In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, 154-62. Association for Computational Linguistics.

Wulff, Stefanie, Ute Römer, and John Swales. 2012. 'Attended/unattended this in academic student writing: Quantitative and qualitative perspectives', *Corpus Linguistics and Linguistic Theory*, 8: 129-57.

Xiao, Richard, and Yan Cao. 2013. 'Native and non-native English abstracts in contrast: A multidimensional move analysis', *Belgian Journal of Linguistics*, 27: 111-34.

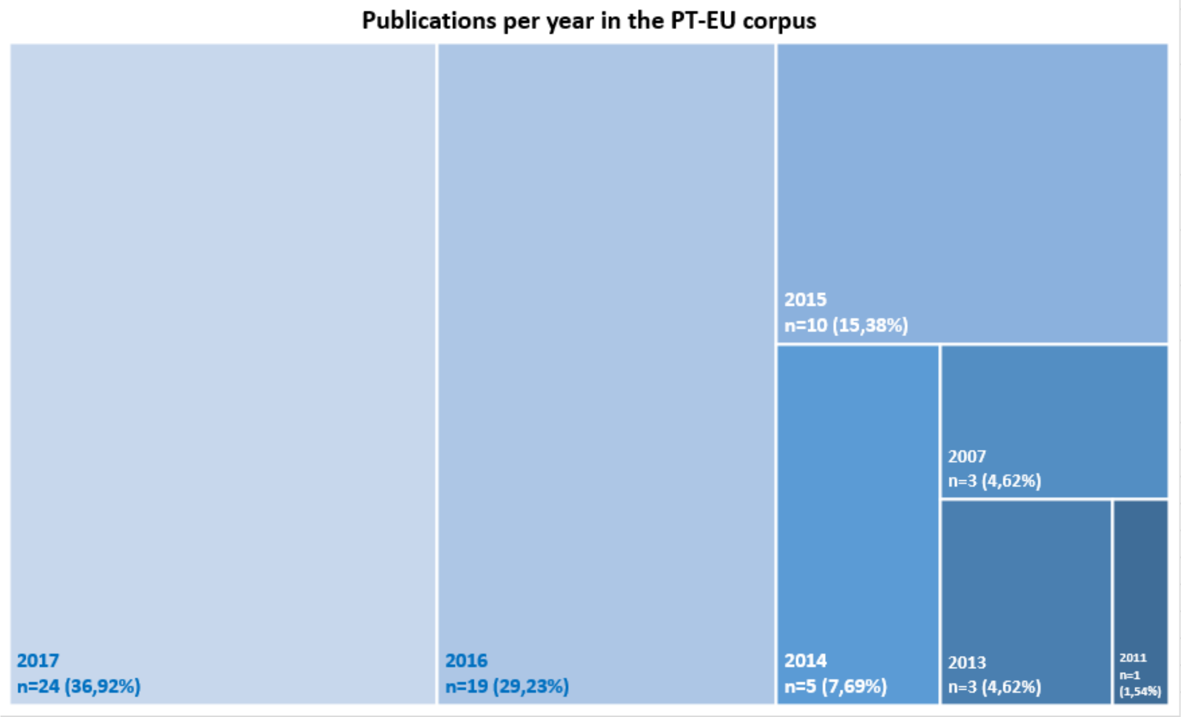
Yakhontova, T. 2006. 'Cultural and disciplinary variation in academic discourse: The issue of influencing factors', *Journal of English for Academic Purposes*, 5: 153-67.

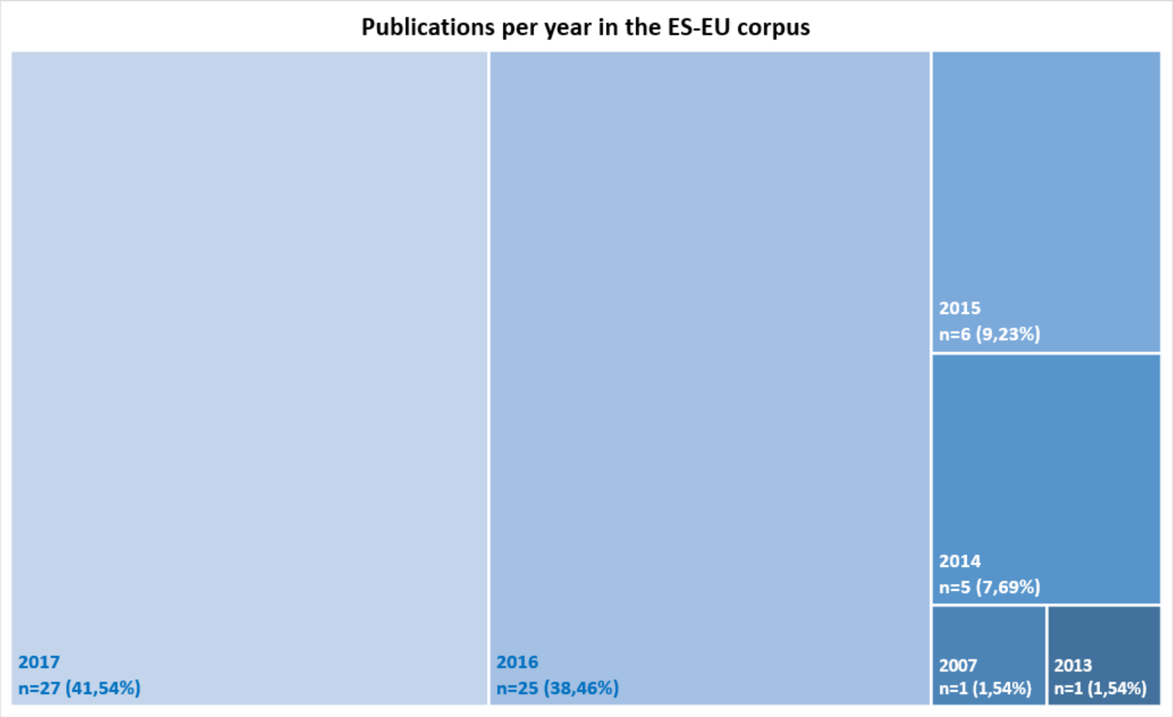
Yllera, Alicia. 1979. *Estilística, Poética e Semiótica Literária* (LIVRARIA ALMEDINA: Coimbra).

Yule, George Udny. 1944. *The Statistical Study of Literary Vocabulary* (Cambridge University Press).

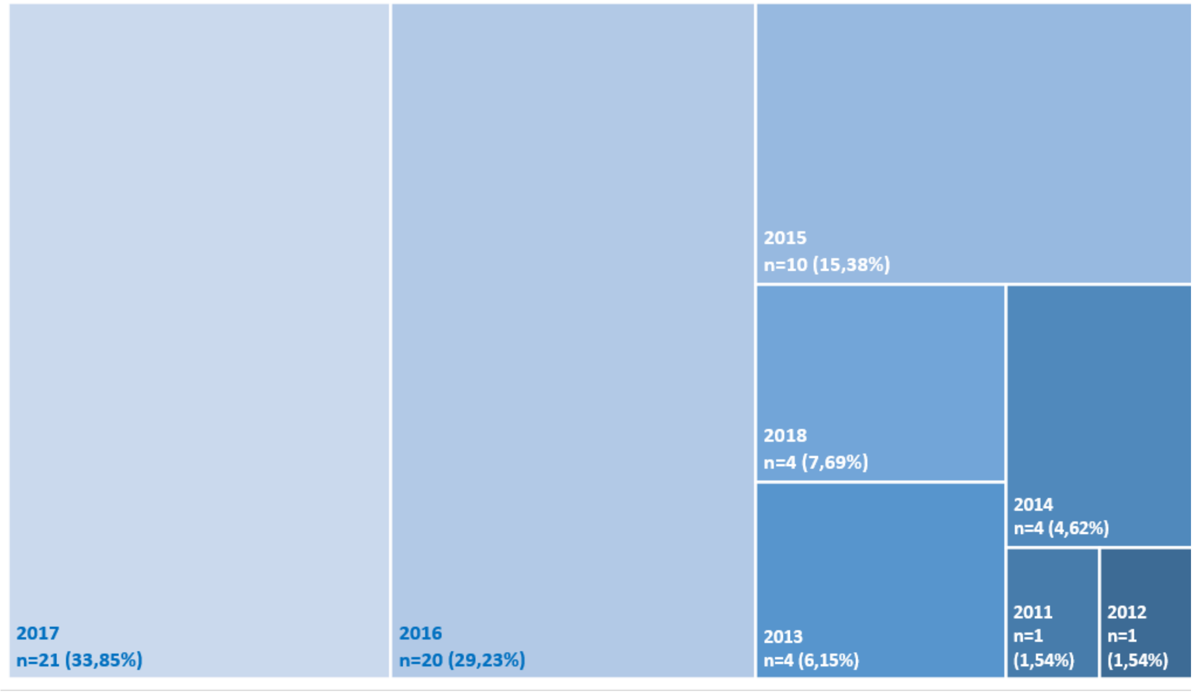
Annexes

Annex 1 – Distribution of publications per year within each CoRA corpus

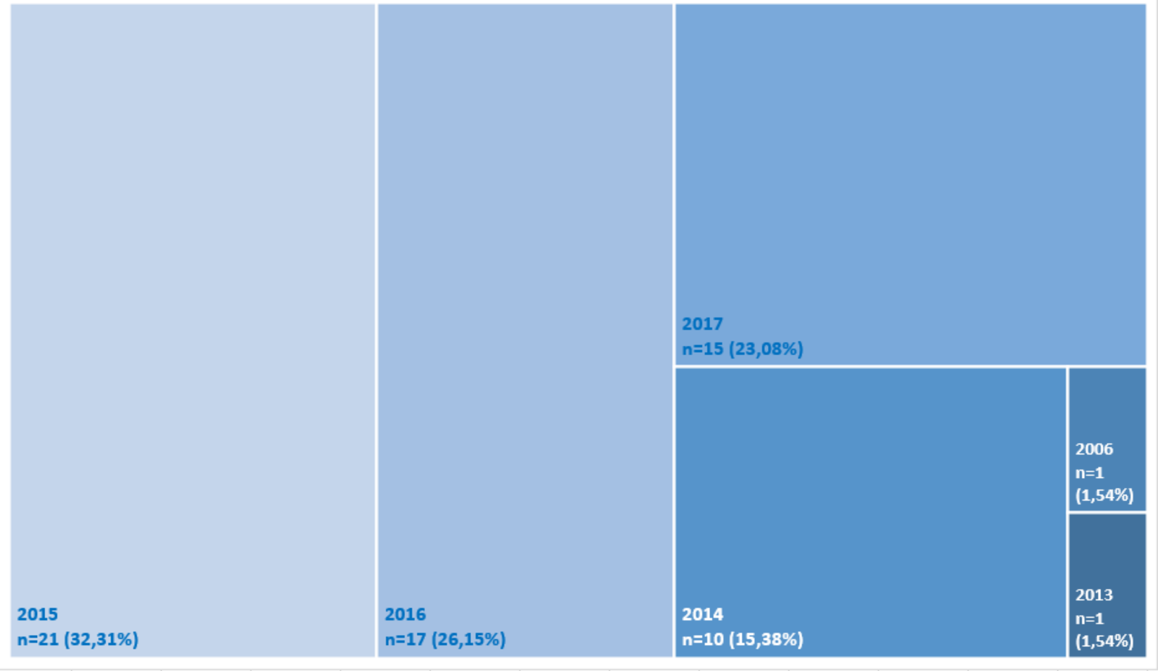

Publications per year in the PT-EU corpus

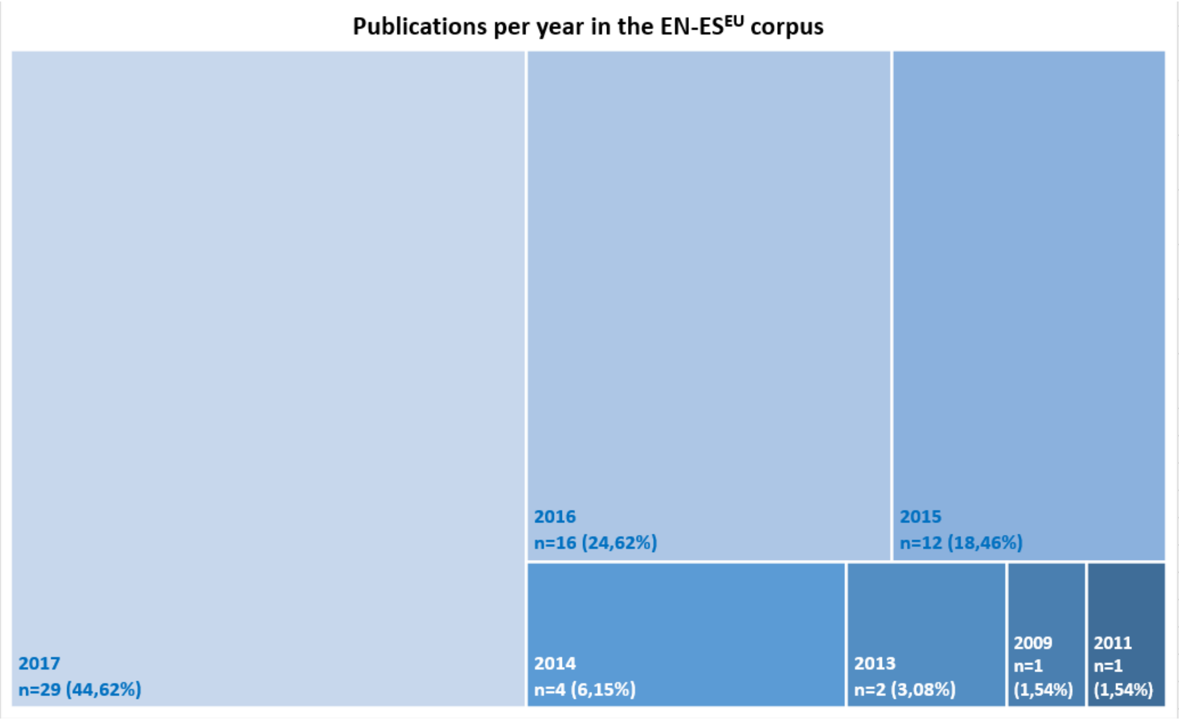

Publications per year in the ES-EU corpus

Publications per year in the EN-GB corpus



Publications per year in the EN-PT^{EU} corpus





Publications per year in the EN-ES^{EU} corpus

Annex 2 – List of journals from which the CoRA OSRAs were extracted

1. Journals of the PT-EU Corpus

Journal	Nº of OSRAs
Acta Médica Portuguesa	3
Acta Obstétrica e Ginecológica Portuguesa	3
Acta Urológica Portuguesa	9
Angiologia e Cirurgia Vascular	2
Arquivos de Medicina	5
GE Jornal Português de Gastrenterologia	2
Medicina Interna	6
Nascer e Crescer	4
Revista Portuguesa de Cardiologia	10
Revista Portuguesa de Endocrinologia, Diabetes e Metabolismo	2
Revista Portuguesa de Imunoalergologia	5
Revista Portuguesa de Medicina Geral e Familiar	5
Revista Portuguesa de Ortopedia e Traumatologia	5
Revista Portuguesa de Pneumologia / Pulmonology Jornal	2
Revista Portuguesa de Saúde Pública	2
TOTAL	**65**

2. Journals of the ES-EU Corpus

Journal	Nº of OSRAs
Anales de La Real Academia Nacional De Farmacia	7
Anales de Pediatría	2
Cirugía Plastica Iberolatinoamericana	1
Emergencias	10
Enfermería Nefrológica	3
Gaceta Sanitaria	3
Gerokomos	2
Gerokomos	3
Neurocirugía	2
Neurología	3
Revista Andaluza de Medicina del Deporte	2
Revista Clínica Médica Familiar	4
Revista de la Sociedad Española del Dolor	1
Revista de Osteoporosis y Metabolismo Mineral	3
Revista Española de Cardiologia	11
Revista Española de Salud Pública	8
TOTAL	**65**

3. Journals of the EN-EU Corpus

Journals	Nº of OSRAs
AIDS	1
Archives of Disease in Childhood	1
Arthritis Research & Therapy	1
BJU International	2
BMC Musculoskeletal Disorders	1
BMJ Disease in childhood	4
BMJ Heart	3
BMJ Open Diabetes Res Care	3
BMJ Open Gastroenterology	3
Bone	1
Clinical Infectious Diseases	1
Emergency Medicine Journal	5
Free Radical Biology and Medicine	1
International Journal of Cancer	1
International Journal of Epidemiology	1
JAMA	1
Journal of Clinical Pathology	6
Journal of Medical Genetics	6
Leukemia	1
Multiple Sclerosis and Related Disorders	1
Musculoskeletal Care	1
Neurology Neurosurgery and Psychiatry Journal	3
Open Biology Journal	4
Parasites & Vectors	1
Pediatric Rheumatology	1
Psychopharmacology	1
Respiratory Research	1
Respiratory Research	1
Sexually Transmitted Infections Journal	3
The Journal of Allergy and Clinical Immunology	1
Thorax Journal	3
Value in Health	1
Grand Total	**65**

4. Journals of the EN-PT^{EU} Corpus

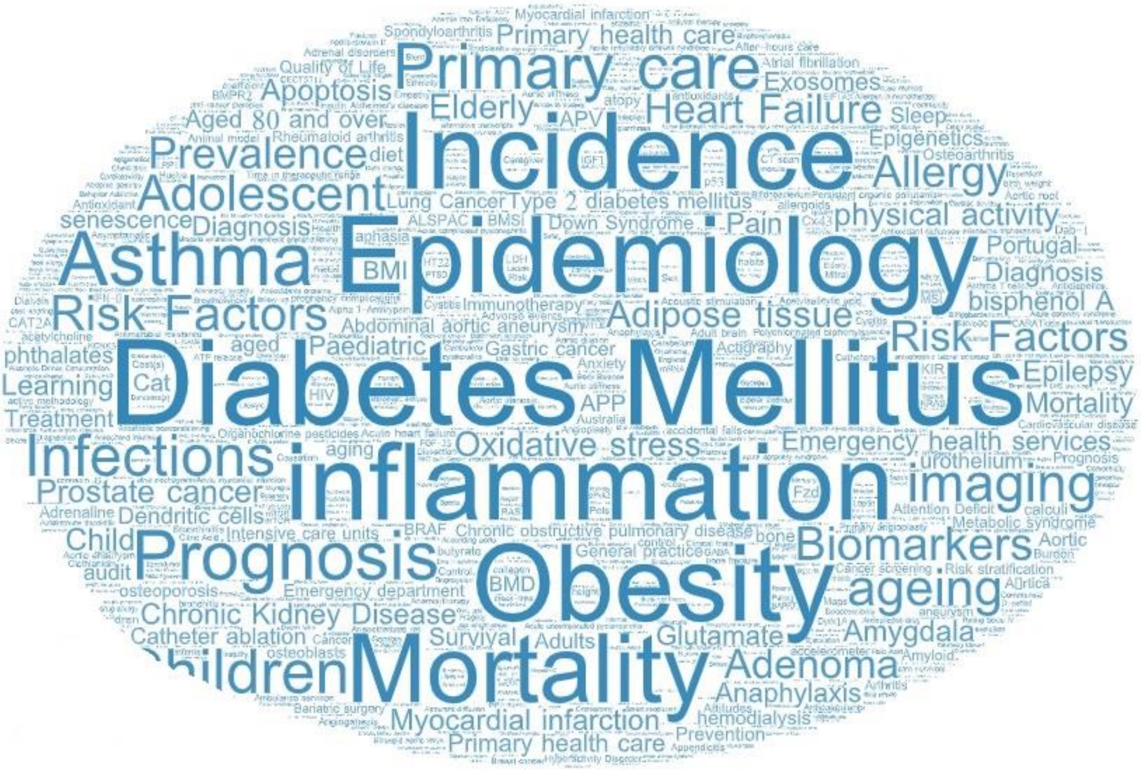
Journals	Nº of OSRAs
Acta Urológica Portuguesa	3
Angiologia e Cirurgia Vascular	1
Annals of Intensive Care	1
Biochemical Pharmacology	1
BMJ Open Diabetes Research & Care	1
Cancer Letters	1
Cardiology	1
Cardiovascular Research	1
Cell and Tissue Research	1
Cellular and Molecular Immunology	1
Clinical Endocrinology	1
DIABETOLOGY & METABOLIC SYNDROME	1
Environmental Toxicology	1
Environmental Research	2
Epidemiology and Infection	1
Epilepsia	1
European Journal of Cancer	1
European Journal of Clinical Nutrition	1
Fertility and Sterility	2
GASTROINTESTINAL ENDOSCOPY	1
Helicobacter	1
International Journal of Endocrinology	1
International Journal of Immunogenetcis	1
International Journal of Surgery	1
Journal of Cardiovascular Computed Tomography	1
Journal of Cardiovascular Pharmacology and Therepeutics	1
Journal of Cellular and Molecular Medicine	1
Journal of Clinical Endocrinology and Metabolism	1
Journal of Cranio-Maxillo-Facial Surgery	1
Journal of Diabetes and Its Complications	1
Journal of Infection and Public Health	1
Journal of Medical Genetics	1
Journal of Neuroimmunology	1
Life Sciences	1
Nature Communications	1
Oncotarget	1
Pediatric Research	2
PLOS One	4
Portuguese Journal of Cardiology	2
Portuguese Journal of Gastroenterology	3
Portuguese Journal of Nephrology and Hypertension	5
Revista Española de Enfermedades Digestivas	1
Rheumatology International	1
RMD Open	2
Supportive Care in Cancer	1

The EMBO Journal	1
The International Journal of Biochemistry & Cell Biology	1
The Journal of Urology	1
Virchows Archiv: European Journal of Pathology	1
TOTAL	**65**

5. Journals of the EN-ES^{EU} Corpus

Journals	**Nº of OSRAs**
Acta Histochemica	2
Actas Urológicas Españolas	1
Annals of Intensive Care	3
Archives of osteoporosis	1
BMJ Open Gastroenterology	1
Bone	1
Brain, Behavior, and Immunity	1
Environmental Pollution	2
Epilepsy & Behavior	1
Frontiers in Molecular Neuroscience	1
Heart	1
International Journal of Cardiology	6
Journal of Affective Disorders	1
Journal of Cellular and Molecular Medicine	2
Journal of Clinical Anesthesia	1
Journal of Clinical Psychopharmacology	1
Journal of Medical Genetics	1
Journal of Translational Medicine	4
Molecular and Cellular Endocrinology	1
Molecular Neurobiology	2
Molecular Therapy - Nucleic Acids	2
Neurobiology of disease	1
Nutrition and Metabolism	4
Pediatric Research	3
Pharmacological Reports	1
PLOS Genetics	4
PLOS Medicine	1
PLOS One	1
Psychoneuroendocrinology	2
SpringerPlus	2
The Company of Biologists	1
The Journal of Allergy and Clinical Immunology	5
Toxicologia	1
Translational Research	3
TOTAL	**65**

Annex 3 – Representation of topics in CoRA according to weight of OSRAs keywords



Appendixes

Appendix 1 – OSRAs comprising the PT-EU corpus, in alphabetical order

1. Almeida, J., J. Monteiro, J. A. Silva, S. Bertoquini and J. Polónia (2016). "Os valores da pressão arterial aórtica e índice de aumentação central em indivíduos com hipertensão da bata branca são mais próximos dos indivíduos normotensos do que dos hipertensos tratados para idênticas idades, género e pressão noturna." Revista Portuguesa de Cardiologia 35(11): 559-567.
2. Bagueixa, M. A., M. H. Pimentel and M. J. Iglesias (2017). "Fragilidade no idoso internado num Serviço de Ortopedia." Revista Portuguesa de Ortopedia e Traumatologia 25(3): 173-184.
3. Barros, C., A. Gomes and E. Pinto (2013). "Estado de saúde e estilos de vida dos idosos portugueses: O que mudou em 7 anos?" Arquivos de Medicina 27(6): 242-247.
4. Basílio, N., A. S. Vitorino and J. M. Nunes (2017). "Caracterização da empatia em internos de medicina geral e familiar." Revista Portuguesa de Medicina Geral e Familiar 33(3): 171-175.
5. Braga, I., F. Branco, J. Cabral, N. Louro, V. Cavadas and A. Fraga (2014). "Litíase urinária no século XXI: análise bibliométrica de publicações na última década." Acta Urológica Portuguesa 32(1): 12-19.
6. BROEIRO-GONÇALVES, P. (2017). "Morbilidade em Idosos Dependentes ao Cuidado das Equipas Domiciliárias da Rede Nacional de Cuidados Continuados Integrados na Região de Lisboa e Vale do Tejo: Estudo Transversal Observacional." Acta Medica Portuguesa 30.
7. Cacela, D., A. Fiarresga, L. Branco, A. Galrinho, P. Rio, M. Selas and R. Ferreira (2015). "Terapêutica percutânea da insuficiência mitral: experiência inicial com o dispositivo MitraClip." Revista Portuguesa de Cardiologia 34(9): 515-524.
8. Carmo, A. d., A. C. Queiroz, F. E. Fontes, J. M. Pego, R. Tomé and F. Rodrigues (2017). "Avaliação do Efeito da Concentração da Hemoglobina e do Volume Globular Médio na Determinação da Hemoglobina A1c: Estudo Retrospetivo." Medicina Interna 24(2): 92-97.
9. Carolino, F., D. Silva, E. D. d. Castro and J. R. Cernadas (2016). "Desafios no diagnóstico de hipersensibilidade a inibidores da bomba de protões." Revista Portuguesa de Imunoalergologia 24(4): 219-225.
10. Correia, L. M., A. Barros and M. L. Brazão (2017). "Polifarmácia, Fármacos Inapropriados e Interacções Medicamentosas nas Prescrições de Doentes Nonagenários." Medicina Interna 24(1): 24-29.
11. Cruz, C., R. Reis, I. Didenko, E. Tomaz and F. Inácio (2017). "O questionário CARATkids e a espirometria na avaliação do controlo da asma." Revista Portuguesa de Imunoalergologia 25(2): 115-125.
12. Custódio, S., S. Lemos, M. Dias and C. Oliveira (2007). "Avaliação de uma série de 361 tumores benignos do ovário submetidos a tratamento cirúrgico."
13. Custódio, S., S. Saleiro, M. Dias and C. Oliveira (2007). "Sarcoma da mama: avaliação de uma série de 11 casos."
14. de Oliveira, S., A. Azenha, A. P. Sousa, J. P. Pinheiro and A. T. A. Santos (2016). "Eficácia da vibroestimulação peniana após lesão vertebromedular." Acta Urológica Portuguesa 33(1): 16-21.

15. de Oliveira, T. M. R., A. J. C. Romão, P. M. S. de Oliveira, S. R. S. Gaspar, F. M. G. Guerreiro and T. M. M. Lopes (2016). "Oxigenoterapia hiperbárica na cistite rádica hemorrágica." Acta Urológica Portuguesa 33(1): 1-5.
16. Dores, H., J. F. Santos, P. Dinis, F. M. Costa, L. Mendes, J. Monge, A. Freitas, P. de Araújo Gonçalves, N. Cardim and M. Mendes (2016). "Variabilidade na interpretação do eletrocardiograma do atleta: mais uma limitação na avaliação pré-competitiva." Revista Portuguesa de Cardiologia 36(6): 443-449.
17. Dores, J. A., P. Kronenberg, P. B. Santos, S. Ferreira and F. C. Gomes (2016). "Oncocitoma renal: tem a URO-TC utilidade no diagnóstico histológico?" Acta Urológica Portuguesa 33(3): 98-103.
18. Eliseu, L., R. Cardoso, N. Almeida, P. Amaro and C. Sofia (2014). "Sépsis em gastrenterologia: uma entidade subvalorizada?" GE Jornal Português de Gastrenterologia 21(4): 131-137.
19. Espinheira, M., M. Grilo, G. Rocha, B. Guedes and H. Guimarães (2011). "Síndrome de aspiração meconial-experiência de um centro terciário." Revista Portuguesa de Pneumología 17(2): 71-76.
20. Espírito Santo, R., C. Salgado, S. Valente and J. Saldanha (2017). "Qualidade dos registos no Boletim de Saúde da Grávida: a importância para o neonatologista." Nascer e Crescer 26(1): 11-20.
21. Eusébio, A., C. Araújo, M. Andrade and A. Duarte (2016). "Escherichia coli nas infeções urinárias da comunidade: comensal ou patogénica?" Acta Urológica Portuguesa 33(2): 37-42.
22. Fernandes, A., M. Cassandra, J. Trigo, J. Nascimento, M. C. Cachulo, R. Providência, M. Costa and L. Gonçalves (2016). "Endocardite de dispositivos, revisão com base na experiência de um centro." Revista Portuguesa de Cardiologia 35(6): 351-358.
23. Fernandes, M. A., S. Miranda, P. Marcelino, I. Mega, J. Machado, R. Perdigoto and E. Barroso (2017). "Mielinólise Centropôntica e Extrapôntica: Experiência de um Centro de Transplante Hepático." Medicina Interna 24(2): 119-123.
24. Fernandes, V., J. Ramalho, M. J. Santos, N. Oliveira and M. L. Pereira (2015). "Diabetes e hiperglicemia: fatores de prognóstico na pneumonia adquirida na comunidade." Revista Portuguesa de Endocrinologia, Diabetes e Metabolismo 10(2): 133-140.
25. Ferreira, C., H. Ferreira, M. Alves, C. Tavares, L. Macedo and Â. Dias (2017). "Estudo PaSeFi: o que ensinam os pais sobre sexualidade aos seus filhos." Nascer e Crescer 26(3): 164-170.
26. FERREIRA, C., H. FERREIRA, M. J. VIEIRA, M. COSTEIRA, L. BRANCO, Â. DIAS and L. MACEDO (2017). "Epidemiologia do Uso de Internet numa População Adolescente e Sua Relação com Hábitos de Sono." Acta médica portuguesa 30.
27. Firmino-Machado, J., J. Yaphe, M. J. Ribas and P. Costa (2017). "O local ideal para a prestação de serviços em alargamento de horário nos cuidados de saúde primários: análise dos custos e da perceção da qualidade dos serviços." Revista Portuguesa de Medicina Geral e Familiar 33(2): 92-104.
28. Fonseca, S., A. Barbosa, Z. Melnikova, C. Silva, S. Silva, V. Alves and J. V. Barreto (2017). "Pneumonias Pneumocócicas e Pneumonias por Influenza A: Estudo Comparativo." Medicina Interna 24(2): 106-111.
29. Freitas, M. M., M. d. L. F. Silva, J. Ribeiro, N. O. Teles, C. Candeias and E. Bronze-da-Rocha (2013). "Análise das regiões subteloméricas em 1180 doentes com atraso mental por FISH e MLPA." Arquivos de Medicina 27(1): 10-14.

30. Goes, A. R., G. Câmara, I. Loureiro, G. Bragança, L. S. Nunes and M. Bourbon (2015). "«Papa Bem»: investir na literacia em saúde para a prevenção da obesidade infantil." Revista portuguesa de saúde pública 33(1): 12-23.
31. Goulão, B., O. Santos, V. Alarcão, R. Portugal, M. Carreira and I. do Carmo (2015). "Prevalência de excesso de peso nos imigrantes brasileiros e africanos residentes em Portugal." Revista Portuguesa de Saúde Pública 33(1): 24-32.
32. Guedes, M. and C. Rego (2016). "Estudo HIPOGAIA: monitorização da hipocoagulação oral com dicumarínicos no concelho de Gaia." Revista Portuguesa de Cardiologia 35(9): 459-465.
33. Guimarães, C., M. Seidi, S. C. Alves, V. Fonseca, M. Irimia and A. Ramos (2016). "Síndrome de asma crítica numa unidade de cuidados intensivos em Portugal." Revista Portuguesa de Imunoalergologia 24(3): 155-162.
34. Lapa, P., R. Silva, T. Saraiva, A. Figueiredo, R. Ferreira, G. Costa and J. P. Lima (2016). "PET/CT com Fluorocolina-F18 no estadiamento inicial do carcinoma da próstata." Acta Urológica Portuguesa 33(3): 87-97.
35. Magalhães, J., B. Rosa, M. J. Moreira, M. Barbosa, A. Rebelo, S. Leite and J. Cotter (2014). "Hiperhomocisteínemia-Uma ameaça oculta da doença inflamatória intestinal?" GE Jornal Português de Gastrenterologia 21(4): 155-160.
36. Marques, F., C. Fonseca, A. R. Nunes, A. Belo, D. Brilhante and J. Cortez (2016). "Contextualizando a elevada prevalência de anemia na população portuguesa: perceção, caracterização e preditores: Um sub-estudo do EMPIRE." Medicina Interna 23(4): 26-38.
37. Melo, L., J. Duarte, D. Roque, I. F. d. Oliveira, A. Faustino, J. Caetano and S. Oliveira (2017). "Endocardite Infecciosa: Casuística do Departamento de Medicina Interna de um Hospital." Medicina interna 24(1): 19-23.
38. Mendes, P., V. P. Cardoso and J. Yaphe (2017). "Stress e burnout em internos de medicina geral e familiar da zona Norte de Portugal: estudo transversal." Revista Portuguesa de Medicina Geral e Familiar 33(1): 16-28.
39. Mota, A. F., B. K. Cardoso, M. F. Jordão, E. Tomazº, L. Caturra and F. Inácio (2017). "Reações anafiláticas em crianças admitidas numa Unidade de urgência." Revista Portuguesa de Imunoalergologia 25(1): 39-49.
40. Mota, R., F. Gomes and D. Ayres-de-Campo (2007). "Tratamento da gravidez ectópica não-rota com metotrexato intramuscular em “dose única”: uma experiência de 5 anos." Acta Obstet Ginecol Port 1(1): 5-9.
41. Neves, Â. (2017). "Atitudes e práticas dos médicos de família do ACeS de Matosinhos face à obesidade." Revista Portuguesa de Medicina Geral e Familiar 33(3): 188-198.
42. Nogueira, A., C. Teixeira, C. Ferreira, S. Ferreira, T. Pinto and T. Ribeiro (2015). "Tendências temporais da infeção pelo vírus da imunodeficiência humana em Portugal: 1984-2013." Arquivos de Medicina 29(6): 148-152.
43. Pereira, H., R. C. Teles, M. Costa, P. C. d. Silva, V. d. G. Ribeiro, V. Brandão, D. Martins, F. Matias, F. Pereira-Machado, J. Baptista, P. F. e. Abreu, R. Santos, A. Drummond, H. C. d. Carvalho, J. Calisto, J. C. Silva, J. L. Pipa, J. Marques, P. Sousa, R. Fernandes, R. C. Ferreira, S. Ramos, E. Oliveira and M. Almeida (2016). "Angioplastia primária em Portugal entre 2002-2013. Atividade segundo o Registo Nacional de Cardiologia de Intervenção." Revista portuguesa de cardiologia 35(7-8): 395-404.

44. Pereira, T. A., C. Neves, C. Esteves, D. Carvalho, L. Delgado and J. L. Medina (2015). "Hipotiroidismo subclínico, tiroidite autoimune e fatores de risco cardiovascular." Arquivos de Medicina 29(3): 69-74.
45. Pinho-Costa, L., S. Moreira, C. Azevedo, P. Azevedo, E. Castro, H. Sousa and M. Melo (2014). "APOLO I: controlo da hipocoagulação na fibrilhação auricular." Revista Portuguesa de Cardiologia 34(5): 337-345.
46. Pires, F. S., P. C. Mota, N. Melo, D. Costa, J. Jesus, R. Cunha, S. Guimarães, C. Souto-Moura and A. Morais (2013). "Fibrose pulmonar idiopática: apresentação clínica, evolução e fatores de prognóstico basais numa coorte portuguesa." Revista Portuguesa de Pneumología 19(1): 19-27.
47. Pires, S., M.-J. Festas, T. Soares, H. Amorim, J. Santoalha, A. Henriques and F. Parada (2014). "Pistas auditivas musicais na fisioterapia em grupo de doentes com Parkinson." Arquivos de Medicina 28(6): 162-166.
48. Portugal, G., L. M. Branco, A. Galrinho, M. M. Carmo, A. T. Timóteo, J. Feliciano, J. Abreu, S. D. Oliveira, L. Batarda and R. C. Ferreira (2016). "Importância da deformação longitudinal na deteção da cardiotoxicidade induzida por quimioterapia e na identificação de padrões específicos de afetação segmentar." Revista Portuguesa de Cardiologia 36(1): 9-15.
49. Primo, J., H. Gonçalves, A. Macedo, P. Russo, T. Monteiro, J. Guimarães and O. Costa (2017). "Prevalência da fibrilhação auricular paroxística numa população avaliada por monitorização contínua de 24 horas." Revista Portuguesa de Cardiologia 36(7-8): 535-546.
50. Raposo, J., R. Soares, A. Rebelo, R. Simões, A. Gonçalves and F. Carneiro (2016). "Devemos negar os benefícios do ácido tranexâmico na artroplastia total do joelho? Um novo protocolo." Revista Portuguesa de Ortopedia e Traumatologia 24(4): 237-246.
51. Ribeiro, C., A. Relvas, L. Carvalho, V. Costa, L. Gomes and M. Costa (2017). "Proteção solar: Conhecimentos e hábitos na população pediátrica." Nascer e Crescer 26(1): 31-35.
52. Ribeiro, S. P., R. B. Costa and C. P. Dias (2017). "Macrossomia Neonatal: Fatores de Risco e Complicações Pós-parto." Nascer e Crescer 26(1): 21-30.
53. ROCHA, J., P. BRANDÃO, A. MELO, S. TORRES, L. MOTA and F. COSTA (2017). "Avaliação da Incontinência Urinária na Gravidez e no Pós-Parto: Estudo Observacional." Acta Medica Portuguesa 30.
54. Rodrigues, E. B., A. P. Silva, E. Coelho and P. V. João (2017). "Tradução e propriedades psicométricas da versão portuguesa do non arthritic hip score (NAHS)." Revista Portuguesa de Ortopedia e Traumatologia 25(3): 160-172.
55. Rodrigues, G. M., J. Albuquerque, F. B. Gonçalves, A. Quintas, R. Abreu, R. Ferreira, N. Camacho, H. Valentim, A. Garcia, M. E. Ferreira and L. Mota Capitão (2015). "Correção endovascular de aneurismas da aorta abdominal em doentes com anatomia desfavorável: resultados institucionais a curto e médio prazo." Angiologia e Cirurgia Vascular 11(3): 158-165.
56. Rodrigues, V., E. Dias, P. Mota, A. Cordeiro and F. Botelho (2015). "Uso de ácido acetilsalicílico, metformina e estatinas e o cancro da próstata: impacto sobre as características patológicas e risco de recidiva bioquímica." Acta Urológica Portuguesa 32(2): 78-85.

57. Sangalho, I., M. P. Barbosa and M. B. Ferreira (2017). "Anafilaxia-8 anos de internamentos no Serviço de Imunoalergologia do Centro Hospitalar de Lisboa Norte." Revista Portuguesa de Imunoalergologia 25(1): 27-38.
58. Sá-Rodrigues, A., P. Cacho-Rodrigues, P. Negrão, M. Ribeiro-Silva, R. Pinto and N. Neves (2016). "Artrodese atlantoaxial com parafusos translaminares em C2: Uma nova opção no tratamento das instabilidades atlanto-axiais." Revista Portuguesa de Ortopedia e Traumatologia 24(4): 270-276.
59. Sepúlveda, L., A. Meireles, P. Moreira, H. Dinis, V. Marques, F. Rolo and A. Mota (2016). "Próteses penianas no tratamento da disfunção erétil: a casuística de 13 anos." Acta Urológica Portuguesa 33(3): 75-80.
60. Silva, D., L. Pacheco-Figueiredo, C. Silva, F. Cruz and J. Silva (2016). "Fatores preditivos da recorrência vesical do carcinoma urotelial do trato urinário superior após nefroureterectomia radical." Acta Urológica Portuguesa 33(2): 43-50.
61. Sousa, R., J. Esteves, A. Sousa, M. Silva, J. Ramos, S. Lopes and A. Oliveira (2017). "Tratamento de Infeções Protésicas com Cirurgia de Revisão a Dois Tempos: resultados de um estudo prospetivo com abordagem protocolada." Revista Portuguesa de Ortopedia e Traumatologia 25(2): 79-91.
62. Timóteo, A. T., S. A. Rosa, M. A. Nogueira, A. Belo and R. C. Ferreira (2016). "Validação externa do score de risco ProACS para estratificação de risco de doentes com síndrome coronária aguda." Revista Portuguesa de Cardiologia 35(6): 323-328.
63. Trindade, I., D. Almeida, M. Romão, S. Rocha, S. Fernandes, V. Varela and M. Braga (2017). "Caracterização do grau de sobrecarga dos cuidadores de utentes dependentes da Unidade de Saúde Familiar USF Descobertas." Revista Portuguesa de Medicina Geral e Familiar 33(3): 178-186.
64. Vinha, A. J. and S. Sampaio (2015). "Correção de aneurisma por via endovascular: fatores de risco para oclusão de ramo." Angiologia e Cirurgia Vascular 11(3): 140-152.
65. Viveiros, A. S., M. Borges, R. Martins, B. Anahory and M. S. Cordeiro (2015). "Estudo LIDIA: risco de diabetes mellitus tipo 2 numa população rural dos Açores." Revista Portuguesa de Endocrinologia, Diabetes e Metabolismo 10(2): 124-127.

Appendix 2 – OSRAs comprising the ES-EU corpus, in alphabetical order

1. Aboal, J., M. N. Torras, D. B. Portell, C. Tirón, R. Brugada and P. L.-O. Ricón (2017). "Angioplastia primaria frente a fibrinolisis en pacientes alejados de un centro con hemodinámica." Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias 29(2): 99-104.
2. Aguilar, A. G., C. G. Viejo and M. B. d. l. Heras (2017). La proteína tuberina: diana terapéutica para activar autofagia y mitofagia evitando la progresión a la diabetes tipo 2. Anales de la Real Academia Nacional de Farmacia.
3. Alamán Valtierra, M., C. Simón Valencia, H. Fuertes Negro, A. Unzuet Galarza, B. Flores Somarriba and N. Halaihel Kassab (2016). "Epidemiología molecular de Bartonella henselae en gatos callejeros y de albergue en Zaragoza, España." Revista Española de Salud Pública 90: e40010.
4. Alberola, S., A. Oliver and J. M. Tomás (2016). "Validación de un modelo intercultural de envejecimiento exitoso en población española." Gerokomos 28(2): 63-67.
5. Alhames, K. A., P. R. Artacho, V. T. S. de la Maza, M. M. O. de Zárate, C. N. Bustos, C. F. Pérez, A. B. Pinedo, J. G. del Castillo and F. J. M. Sánchez (2017). "Escala INFURG-SEMES para el diagnóstico de apendicitis aguda en los pacientes de 2 a 20 años atendidos en los servicios de urgencias hospitalarios." Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias 29(4): 231-236.
6. Alonso, F. J., M. D. Carranza, J. Rueda and J. Naranjo (2014). "Composición corporal en escolares de primaria y su relación con el hábito nutricional y la práctica reglada de actividad deportiva." Revista andaluza de medicina del deporte 7(4): 137-142.
7. Alquezar, A., M. S. Bel, M. A. Rizzi, I. G. Saladich, M. Grau, A. Sionis and J. O. Llanos (2017). "Evaluación de una estrategia diagnóstica combinada con copeptina y troponina T ultrasensibles en el infarto de miocardio sin elevación del segmento ST en los servicios de urgencias." Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias 29(4): 237-244.
8. Álvarez, M., V. Bertomeu-González, M. F. Arcocha, P. Moriña, L. Tercedor, Á. F. de Loma, M. Pachón, A. García, M. Pardo, T. Datino, C. Alonso and J. Osca (2016). "Ablación con catéter no guiada por fluoroscopia. Resultados de un registro prospectivo multicéntrico." Revista Española de Cardiología 70(9): 699-705.
9. Andreu, Ó. M., P. L. Soriano, X. E. Roig, P. H. Puente, J. J. Rodríguez, V. Gil, C. Xipell, C. Sánchez, S. Aguiló and F. J. M. Sánchez (2017). "Atención prehospitalaria a los pacientes con insuficiencia cardiaca aguda en España: estudio SEMICA." Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias 29(4): 223-230.
10. Avanzas, P., I. Pascual, A. J. Muñoz-García, J. Segura, J. H. Alonso-Briales, J. S. d. Lezo, M. Pan, M. F. Jiménez-Navarro, J. López-Aguilera, J. M. Hernández-García and C. Morís (2016). "Seguimiento a largo plazo de pacientes con estenosis aórtica grave tratados con prótesis autoexpandible." Revista Española de Cardiología 70(4): 247-253.
11. Ballesteros-Peña, S., I. Fernández-Aedo and G. Vallejo-De la Hoz (2017). "Eficacia del cloruro de etilo en aerosol como anestésico local previo a la punción arterial: ensayo

clínico aleatorizado controlado con placebo." Emergencias (St. Vicenç dels Horts): 161-166.

12. Barge-Caballero, E., C. Barbeito-Caamaño, G. Barge-Caballero, D. Couto-Mallón, M. J. Paniagua-Martín, R. Marzoa-Rivas, M. Solla-Bucet, F. Estévez-Cid, J. M. Herrera-Noreña, J. J. Cuenca-Castillo, J. M. Vázquez-Rodríguez and M. G. Crespo-Leiro (2016). "Estado serológico frente a Toxoplasma gondii en receptores de trasplante cardiaco:¿ un factor pronóstico independiente?" Revista Española de Cardiología 69(12): 1160-1166.
13. Bautista, F., S. Gallego, A. Cañete, J. Mora, C. D. d. Heredia, O. Cruz, J. M. Fernándezc, S. Rivesd, P. Berlangac, R. Hladunb, A. J. Ribellesc, L. Maderoa, M. Ramíreza, R. F. Delgadoe, A. Pérez-Martínezf, C. Matag, A. Llortb, J. M. Brotoh, M. E. Celag, G. Ramírezh, C. Sábadob, T. Achai, I. Astigarragaj, A. Sastref, A. Muñozk, M. Guibelaldel and L. Moreno (2016). "Ensayos clínicos precoces en oncología pediátrica en España: una perspectiva nacional." Anales de Pediatría (English Edition) 87(3): 155-163.
14. Cabeza-Ruiz, R., N. Castro-Lemus, R. Centeno-Prada and J. Beas-Jiménez (2016). "Desplazamiento del centro de presiones en personas con síndrome de Down en bipedestación." Revista Andaluza de Medicina del Deporte 9(2): 62-66.
15. Calvo, C., I. Aguado, M. L. García-García, E. Ruiz-Chercoles, E. Díaz-Martinez, R. M. Albanil, O. Campelo, A. Olivas, L. Munóz-Gonzalez, F. Pozo, R. Fernandez-Arroyo, A. Fernandez-Rincón, A. Calderon, I. Casas and G. d. E. d. S. Recurrentes (2017). Infecciones virales respiratorias en una cohorte de niños durante el primer año de vida y su papel en el desarrollo de sibilancias. Anales de Pediatría, Elsevier.
16. Cambra, A. D., X. Rosselló, J. S. Roselló, M. Vila, A. Hidalgo, I. D. Rodríguez, R. L. Petracca, G. Pons-Lladó, J. O. Llanos and A. Sionis (2016). "Troponina T de alta sensibilidad y angiotomografía computarizada coronaria para el diagnóstico rápido del dolor torácico en el servicio de urgencias." Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias 28(1): 9-15.
17. Carbayo García, J. J., R. Tuesta Reina, J. F. Sastre García, J. J. Criado Álvarez, C. Gómez González and J. Rodríguez Losáñez (2014). "Valoración de la función renal en diabéticos tipo 2 y su adecuación al tratamiento antidiabético oral." Revista Clínica de Medicina de Familia 7(1): 8-13.
18. Castel, A., J. Miró and M. Rull (2007). "La escala de dolor BS-21: datos preliminares sobre su fiabilidad y validez para evaluar la intensidad del dolor en geriatría." Revista de la Sociedad Española del Dolor 14(4): 274-283.
19. Cequier, Á., A. Ariza-Sole, F. J. Elola, C. Fernandez-Perez, J. L. Bernal, J. V. Segura, A. Iñiguez and V. Bertomeu (2017). "Impacto en la mortalidad de diferentes sistemas de asistencia en red para el tratamiento del infarto agudo de miocardio con elevación del segmento ST. La experiencia de España." Revista española de cardiología 70(3): 155-161.
20. Concha Sanz, S., M. Torre Ruiz and A. Hurtado Aguilar (2014). "Consumo de tabaco y alcohol en 1º y 2º de ESO en una población rural." Revista Clínica de Medicina de Familia 7(3): 169-176.
21. Córdoba-Soriano, J. G., J. Jiménez-Mazuecos, A. R. Juárez, A. Gutiérrez-Díez, E. G. Ibañes, B. Samaniego-Lampón, I. Lozano, A. Gallardo-López, L. Díaz, R. Sanz-Ruiz, D.

Melehi, M. I. Barrionuevo-Sánchez, J. Rondán-Murillo, J. M. Vegas-Valle and J. Elízaga (2016). "Seguridad y factibilidad de la intervención coronaria percutánea ambulatoria en pacientes seleccionados: datos de un registro multicéntrico español." Revista Española de Cardiología 70(7): 535-542.

22. de Villar, L. O. P., S. A. García, M. J. L. Pérez, J. J. A. Cuenca, V. B. Caballer and E. S. Ortí (2016). "Comparación de un programa de ejercicio intradiálisis frente a ejercicio domiciliario sobre capacidad física funcional y nivel de actividad física." Enfermería Nefrológica 18(1): 42.
23. Degayón, V. T., F. J. M. Pérez, J. M. T. Murillo, M. J. F. Dáder, M. I. B. Parejo and M. A. C. Hernández (2016). "Resultados negativos asociados a la medicación en los pacientes con fibrilación auricular permanente atendidos en un servicio de urgencias hospitalario." Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias 28(2): 75-82.
24. del Valle Garrido López, M., C. Sesmero Ramos, A. Ortigosa Barriola and E. Gruss Vergara (2015). "Valoración de la implantación del seguimiento ecográfico del acceso vascular autólogo." Enfermería Nefrológica 18(4): 260-264.
25. Delgado-Calle, J., M. Alonso, J. Ortiz, A. Montero, C. Garcés, C. Sañudo, M. Pérez-Aguilar, M. Pérez Núñez and J. Riancho (2014). "Análisis comparativo del epigenoma del tejido óseo y de osteoblastos primarios." Revista de Osteoporosis y Metabolismo Mineral 6(2): 35-39.
26. Diago, E. B., L. L. del Val, S. S. Lasaosa, E. L. Garcia and A. V. Alebesque (2016). "Relación entre el trastorno de conducta del sueño REM y el trastorno de control de impulsos en pacientes con enfermedad de Parkinson." Neurología 32(8): 494-499.
27. Domínguez Sánchez-Migallón, P. (2014). "Cambios en el control metabólico de los pacientes diabéticos tipo 2 de un centro de salud." Revista Clínica de Medicina de Familia 8(1): 11-18.
28. Fernández-Friera, L., J. M. García-Ruiz, A. García-Álvarez, R. Fernández-Jiménez, J. Sánchez-González, X. Rossello, S. Gómez-Talavera, G. J. López-Martín, G. Pizarro, V. Fuster and B. Ibáñez (2016). "Impacto del territorio miocárdico infartado en la cuantificación del área en riesgo mediante cardiorresonancia magnética." Revista Española de Cardiología 70(5): 323-330.
29. Gallego-Delgado, M., E. González-López, F. Muñoz-Beamud, J. Buades, L. Galán, J. L. Muñoz-Blanco, J. Sánchez-González, B. Ibáñez, J. G. Mirelis and P. García-Pavía (2016). "El volumen extracelular detecta la amiloidosis cardiaca y está correlacionado con el deterioro neurológico en la amiloidosis familiar relacionada con la transtiretina." Revista Española de Cardiología 69(10): 923-930.
30. García-Giralt, N., L. De-Ugarte, G. Yoskovitz, R. Güerri, D. Grinberg, X. Nogués, L. Mellibovsky, S. Balcells and S. Díez-Pérez (2015). "Estudio del patrón de expresión de microRNAs en el hueso osteoporótico." Revista de Osteoporosis y Metabolismo Mineral 8(1): 5-14.
31. Garzón-Maldonado, F., M. Gutiérrez-Bedmar, N. García-Casares, F. Pérez-Errázquin, A. Gallardo-Tur and M. M.-V. Torres (2017). "Calidad de vida relacionada con la salud en cuidadores de pacientes con enfermedad de Alzheimer." Neurología 32(8): 508-515.

32. Gómez, P. A., A. M. Castaño-León, D. Lora, S. Cepeda and A. Lagares (2017). "Evolución temporal en las características de la tomografía computarizada, presión intracraneal y tratamiento quirúrgico en el traumatismo craneal grave: análisis de la base de datos de los últimos 25 años en un servicio de neurocirugía." Neurocirugía 28(1): 1-14.
33. Gómez-Cantorna, C., M. Clemente, C. Bugallo-Carrera and M. Gandoy-Crego (2016). "Cuidados paliativos gerontológicos: influencia de las condiciones laborales y burnout en el personal de enfermería." Gerokomos 27(3): 91-96.
34. Gómez-Hernández, A., N. B. Redondo, L. P. Loaiza, Ó. E. Illanes, S. Díaz-Castroverde and M. B. de las Heras (2016). Papel de la isoforma A del receptor de la insulina y del IGF-1R en el crecimiento de la placa aterosclerótica. Anales de la Real Academia Nacional de Farmacia.
35. Isabel Portillo Villares, Eunate Arana-Arri, Isabel Idigoras Rubio, Josep Alfons Espinás Piñol, Francisco Pérez Riquelme, Mariola de la Vega Prieto, Alvaro González Aledo, Elena Oceja Setien, Mercedes Vanaclocha Espi, Josefa Ibáñez Cabanell, Dolores Salas Trejo and g. CRIBEA (2017). "Lesiones detectadas en seis programas poblacionales de cribado de cáncer colorrectal en España. Proyecto CRIBEA." Revista Española de Salud Pública 91: 201702021.
36. José, M., S. Brugaletta, J. A. G. Hospital, J. A. Baz, A. P. de Prado, R. L. Palop, B. Cid, T. G. Camarero, A. Diego and F. G. de Carlos (2017). "Angioplastia primaria en mayores de 75 años. Perfil de pacientes y procedimientos, resultados y predictores pronósticos en el registro ESTROFA IM+ 75." Revista Española de Cardiología 70(2): 81-87.
37. López González, A., L. Díaz Rodríguez, Á. Novo Casas, S. Cid Armada and M. Mojón Barcia (2016). "Evaluación de la efectividad y satisfacción del apósito con Gluconato de Clorhexidina 3M Tegaderm en el cuidado del catéter central tunelizado para hemodiálisis." Enfermería Nefrológica 19(1): 56-62.
38. López, S., C. Faro, L. Lopetegui, E. Pujol-Ribera, M. Monteagudo, J. Cobo and M. I. Fernández (2017). "Impacto del abuso sexual durante la infancia-adolescencia en las relaciones sexuales y afectivas de mujeres adultas." Gaceta Sanitaria 31: 210-219.
39. Lorente Antoñanzas, R., J. L. Varona Malumbres, F. Antoñanzas Villar and J. Rejas Gutiérrez (2016). "La vacunación anti-neumocócica con la vacuna conjugada 13-valente en población inmunocompetente de 65 años: análisis del impacto presupuestario en España aplicando un modelo de transmisión dinámica." Revista Española de Salud Pública 90: e40001.
40. Marín-Méndez, J., M. Borra-Ruiz, M. Álvarez-Gómez and C. S. Esperón (2016). "Desarrollo psicomotor y dificultades del aprendizaje en preescolares con probable trastorno por déficit de atención e hiperactividad. Estudio epidemiológico en Navarra y La Rioja." Neurología 32(8): 487-493.
41. Martín, J. d. T., E. F. Millán, E. L. Mollinedo, F. E. Pons and C. Á. Escolá (2016). Papel del péptido insulinotrópico dependiente de glucosa en la programación nutricional del síndrome metabólico. Anales de la Real Academia Nacional de Farmacia.
42. Mateo, A. J., R. S. García, A. M. Sarmiento, L. G. Guidet, J. Gálvez and R. García-Domenech (2016). Aplicación de la Topología Molecular para la predicción de la actividad frente a leishmania de un grupo de compuestos derivados del pirrol [1, 2-α] quinoxalina. Anales de la Real Academia Nacional de Farmacia.

43. Mateos Rodríguez, A. A., A. Andrés Belmonte, F. D. Río Gallegos and E. Coll (2017). "Factores que influyen en la evolución de los injertos de donantes tras muerte cardiaca extrahospitalaria." Emergencias (Sant Vicenç dels Horts): 167-172.
44. Moyano Santiago, M. A. and J. M. Rivera Lirio (2017). "Aspectos relevantes para el diseño de planes de salud sostenibles orientados a los grupos de interés. Una propuesta basada en la guía ISO 26000: 2010." Revista Española de Salud Pública 91: 201701005.
45. Navas, P., J. Tenorio, C. A. Quezada, E. Barrios, G. Gordo, P. Arias, M. L. Meseguer, A. Santos-Lozano, J. P. Doza, P. Lapunzina and P. Escribano Subías (2016). "Análisis de los genes BMPR2, TBX4 y KCNK3 y correlación genotipo-fenotipo en pacientes y familias españolas con hipertensión arterial pulmonar." Revista Española de Cardiología 69(11): 1011-1019.
46. Pacho, C., M. Domingo, R. Núñez, J. Lupón, P. Moliner, M. d. Antonio, B. González, J. Santesmases, E. Vela, J. Tor and A. Bayes-Genis (2017). "Una consulta específica al alta (STOP-HF-Clinic) reduce los reingresos a 30 días de los pacientes ancianos y frágiles con insuficiencia cardiaca." Revista Española de Cardiología 70(8): 631-638.
47. Parro Moreno, A., M. Santiago Pérez, V. Abraira Santos, J. L. Aréjula Torres, A. Díaz Holgado, A. Gandarillas Grande, J. M. Morales Asencio and P. Serrano Gallardo (2016). "Control de la diabetes mellitus en población adulta según las características del personal de enfermería de atención primaria de la Comunidad de Madrid: análisis multinivel." Revista Española de Salud Pública 90: e40005.
48. Pérez Vega, F. J., M. T. Gutiérrez Vázquez, J. R. Lorenzo Peñuelas, J. F. Domínguez Bermúdez, J. C. Armario Hita and G. d. Castro Maqueda (2015). "Alternativa al cierre de heridas crónicas mediante injertos de Reverdin y factores de crecimiento en cirugía menor ambulatoria." Gerokomos 26(1): 34-39.
49. Petronila Gómez, L., S. Aragón Chicharro and B. Calvo Morcuende (2017). "Caídas en ancianos institucionalizados: valoración del riesgo, factores relacionados y descripción." Gerokomos 28(1): 2-8.
50. Porras-Povedano, M., A. Roldán-Garrido and V. Santacruz-Hamer (2017). "Brote epidémico por Tos ferina en Écija (Sevilla), 2016." Revista Española de Salud Pública 91: 201701008.
51. Ramos-Fernández, J. M., D. Moreno-Pérez, M. Gutiérrez-Bedmar, A. Hernández-Yuste, A. M. Cordón-Martínez, G. Milano-Manso and A. Urda-Cardona (2017). "Predicción de la evolución de la bronquiolitis por virus respiratorio sincitial en lactantes menores de 6 meses." Revista Española de Salud Pública 91: 201701006.
52. Reyes-García, R., P. Rozas-Moreno, A. García-Martín, B. García-Fontana, S. Morales-Santana and M. Muñoz-Torres (2015). "Dickkopf1 (DKK1), metabolismo óseo y enfermedad ateroesclerótica en pacientes con diabetes mellitus tipo 2." Revista de Osteoporosis y Metabolismo Mineral 8(1): 24-29.
53. Rivero, N. L., I. G. A. Santos, J. L. L. Soto and C. V. Gómez (2016). Aportaciones de la microcalorimetría al diagnóstico precoz de infecciones bacteriana. Anales de la Real Academia Nacional de Farmacia.
54. Rodríguez-Mena, R., J. Piquer-Belloch, J. L. Llácer-Ortega, P. Riesgo-Suárez and V. Rovira-Lillo (2017). "Anatomía de los pedúnculos cerebelosos en 3D basada en

microdisección de fibras y demostración a través de tractografía." Neurocirugía 28(3): 111-123.

55. Sáez-Jiménez, R. and J. Bonis (2015). "Estudio descriptivo sobre el uso de antiinflamatorios no esteroideos por vía intramuscular para el tratamiento de la lumbalgia aguda en las consultas de Atención Primaria en España durante 2002-2011." Revista Clínica de Medicina de Familia 8(2): 103-109.
56. Saltijeral, A., L. P. d. Isla, R. Alonso, O. Muñiz, J. L. Díaz-Díaz, F. Fuentes, N. Mata, R. d. Andrés, G. Díaz-Soto, J. Pastor, J. M. Pinilla, D. Zambón, X. Pinto, L. Badimón and P. Mata (2017). "Consecución de objetivos terapéuticos de colesterol LDL en niños y adolescentes con hipercolesterolemia familiar. Registro longitudinal SAFEHEART." Revista Española de Cardiología 70(6): 444-450.
57. Santos-Moriano, P., L. Fernández-Arrojo, B. Rodríguez-Colinas, A. Ballesteros and F. J. Plou (2015). Síntesis enzimática de fructooligosacáridos estimuladores de la microbiota colónica. Anales de la Real Academia Nacional de Farmacia.
58. Serna-Cuéllar, E. and L. Santamaría-Solís (2013). "Protocolo de extracción y procesamiento de células madre adultas del tejido adiposo abdominal: coordenadas del cirujano plástico en la investigación traslacional." Cirugía Plástica Ibero-Latinoamericana 39: s44-s50.
59. Vasco-Aguas, K., C. Ferrando-Hernández, E. Barrio-Miguel, M. Álvarez-Izquierdo, J. Gálvez and R. García-Domenech (2017). Predicción de la actividad anti-Trypanosoma brucei rhodesiense de un grupo de 3, 5-Difenilisoxazoles dicationicos por medio de la Topología Molecular. Anales de la Real Academia Nacional de Farmacia.
60. Vecina, S. T., J. M. Duarte, M. O. Marcos, M. G. R. Navarro, V. Borillo, L. S. J. Gago, F. R. Egea and M. C. Borrás (2016). "Estudio sobre la reducción de eventos adversos en pacientes y problemas de bioseguridad de los profesionales derivados de la aplicación de catéteres vasculares en urgencias." Emergencias 28: 89-96.
61. Velarde-García, J. F., R. Luengo-González, R. González-Hervías, S. González-Cervantes, B. Álvarez-Embarba and D. Palacios-Ceña (2017). "Dificultades para ofrecer cuidados al final de la vida en las unidades de cuidados intensivos. La perspectiva de enfermería." Gaceta Sanitaria 31: 299-304.
62. Vicente, M. D., M. C. L. Agara, J. L. Barrios and A. C. Blasco (2017). "Pielonefritis aguda complicada y no complicada en urgencias: indicadores de proceso y resultado." Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias 29(1): 27-32.
63. Villanueva Álvarez, E., M. Fernández Rodríguez, E. Viano Pérez and M. Amorín Bayón (2017). "Fiabilidad en la medición de la temperatura corporal con un termómetro timpánico en pacientes geriátricos." Gerokomos 28(2): 68-72.
64. Viñas Casasola, M. J., P. Fernández Navarro, M. L. Fajardo Rivas, J. L. Gurucelain Raposo and J. Alguacil Ojeda (2017). "Distribución municipal de la incidencia de los tumores más frecuentes en un área de elevada mortalidad por cáncer." Gaceta sanitaria 31: 100-107.
65. Wijers, I. G., A. Sánchez Gómez and J. A. Taveira Jiménez (2017). "Estudio espacial de la sífilis infecciosa y la infección gonocócica en un servicio de salud pública de área de Madrid." Revista Española de Salud Pública 91: e201706033.

Appendix 3 – OSRAs comprising the EN-GB corpus, in alphabetical order

1. Alicia C. Thornton, Sophie Jose, Sanjay Bhagani, David Chadwick, David Dunn, Richard Gilson, Janice Main, Mark Nelson, Alison Rodger, Chris Taylor, Elaney Youssef, Clifford Leen, Mark Gompels, Stephen Kegg, Achim Schwenk and Caroline Sabin (2017). "Hepatitis B, hepatitis C, and mortality among HIV-positive individuals." AIDS (London, England) 31(18): 2525.
2. Bevis, M., M. Marshall, T. Rathod and E. Roddy (2016). "The association between gout and radiographic hand, knee and foot osteoarthritis: a cross-sectional study." BMC musculoskeletal disorders 17(1): 1-7.
3. Black, J. A., R. K. Simmons, C. E. Boothby, M. J. Davies, D. Webb, K. Khunti, G. H. Long and S. J. Griffin (2015). "Medication burden in the first 5 years following diagnosis of type 2 diabetes: findings from the ADDITION-UK trial cohort." BMJ Open Diabetes Research and Care 3(1): e000075.
4. Body, R., C. Boachie, A. McConnachie, S. Carley, P. Van Den Berg and F. E. Lecky (2017). "Feasibility of the Manchester Acute Coronary Syndromes (MACS) decision rule to safely reduce unnecessary hospital admissions: a pilot randomised controlled trial." Emergency Medicine Journal 34(9): 586-592.
5. Body, R., E. Carlton, M. Sperrin, P. S. Lewis, G. Burrows, S. Carley, G. McDowell, I. Buchan, K. Greaves and K. Mackway-Jones (2016). "Troponin-only Manchester Acute Coronary Syndromes (T-MACS) decision aid: single biomarker re-derivation and external validation in three cohorts." Emergency Medicine Journal 34(6): 349-356.
6. Bowen, L., A. Shaw, M. D. Lyttle and S. Purdy (2017). "The transition to clinical expert: enhanced decision making for children aged less than 5 years attending the paediatric ED with acute respiratory conditions." Emergency Medicine Journal 34(2): 76-81.
7. Brain, K., B. Carter, K. J. Lifford, O. Burke, A. Devaraj, D. R. Baldwin, S. Duffy and J. K. Field (2017). "Impact of low-dose CT screening on smoking cessation among high-risk participants in the UK Lung Cancer Screening Trial." Thorax 72(10): 912-918.
8. Brown, J., A. Roy, R. Harris, S. Filson, M. Johnson, I. Abubakar and M. Lipman (2016). "Respiratory symptoms in people living with HIV and the effect of antiretroviral therapy: a systematic review and meta-analysis." Thorax 72(4): 355-366.
9. Cavany, S. M., T. Sumner, E. Vynnycky, C. Flach, R. G. White, H. L. Thomas, H. Maguire and C. Anderson (2017). "An evaluation of tuberculosis contact investigations against national standards." Thorax 72(8): 736-745.
10. Collier, S., F. Matjiu, G. Jones, M. Harber and S. Hopkins (2013). "A prospective study comparing contamination rates between a novel mid-stream urine collection device (Peezy) and a standard method in renal patients." Journal of clinical pathology 67(2): 139-142.
11. Connolly, S., K. Kotseva, C. Jennings, A. Atrey, J. Jones, A. Brown, P. Bassett and D. Wood (2017). "Outcomes of an integrated community-based nurse-led cardiovascular disease prevention programme." Heart 103(11): 840-847.
12. Coulthard, M. G., H. J. Lambert, S. J. Vernon, E. W. Hunter, M. J. Keir and J. N. Matthews (2014). "Does prompt treatment of urinary tract infection in preschool children prevent renal scarring: mixed retrospective and prospective audits." Archives of disease in childhood 99(4): 342-347.

13. Curtis, E. M., R. van der Velde, R. J. Moon, J. P. van den Bergh, P. Geusens, F. de Vries, T. P. van Staa, C. Cooper and N. C. Harvey (2016). "Epidemiology of fractures in the United Kingdom 1988–2012: variation with age, sex, geography, ethnicity and socioeconomic status." Bone 87: 19-26.
14. Davies, F. C., F. E. Lecky, R. Fisher, M. Fragoso-Iiguez and T. J. Coats (2017). "Major trauma from suspected child abuse: a profile of the patient pathway." Emergency medicine journal 34(9): 562-567.
15. Dunwell, T. L. and P. W. Holland (2017). "A sister of NANOG regulates genes expressed in pre-implantation human development." Open biology 7(4): 170027.
16. Eisen, S., S. Sukhani, A. Brightwell, S. Stoneham and A. Long (2013). "Peer mentoring: evaluation of a novel programme in paediatrics." Archives of disease in childhood 99(2): 142-146.
17. Ellingford, J. M., S. Barton, S. Bhaskar, J. O'Sullivan, S. G. Williams, J. A. Lamb, B. Panda, P. I. Sergouniotis, R. L. Gillespie, S. P. Daiger, G. Hall, T. Gale, I. C. Lloyd, P. N. Bishop, S. C. Ramsden and G. C. M. Black (2016). "Molecular findings from 537 individuals with inherited retinal disease." Journal of medical genetics 53(11): 761-767.
18. Evans, C. A., R. Rosser, J. S. Waby, J. Noirel, D. Lai, P. C. Wright, E. A. Williams, S. A. Riley, J. P. Bury and B. M. Corfe (2015). "Reduced keratin expression in colorectal neoplasia and associated fields is reversible by diet and resection." BMJ open gastroenterology 2(1): e000022.
19. Fisk, M., J. Cheriyan, D. Mohan, C. M. McEniery, J. Forman, J. R. Cockcroft, J. H. F. Rudd, R. Tal-Singer, N. S. Hopkinson, M. I. Polkey and I. B. Wilkinson (2018). "Vascular inflammation and aortic stiffness: potential mechanisms of increased vascular risk in chronic obstructive pulmonary disease." Respiratory research 19(1): 1-10.
20. Fulford, A. J., K. K. Ong, C. E. Elks, A. M. Prentice and B. J. Hennig (2014). "Progressive influence of body mass index-associated genetic markers in rural Gambians." Journal of medical genetics 52(6): 375-380.
21. Hamer, M., R. A. Hackett, S. Bostock, A. I. Lazzarino, L. A. Carvalho and A. Steptoe (2014). "Objectively assessed physical activity, adiposity, and inflammatory markers in people with type 2 diabetes." BMJ Open Diabetes Research and Care 2(1).
22. Harrison, O. B., K. Cole, J. Peters, F. Cresswell, G. Dean, D. W. Eyre, J. Paul and M. C. Maiden (2017). "Genomic analysis of urogenital and rectal Neisseria meningitidis isolates reveals encapsulated hyperinvasive meningococci and coincident multidrug-resistant gonococci." Sexually transmitted infections 93(6): 445-451.
23. Hawley, C., M. Sakr, S. Scapinello, J. Salvo and P. Wrenn (2017). "Traumatic brain injuries in older adults—6 years of data for one UK trauma centre: retrospective analysis of prospectively collected data." Emergency medicine journal 34(8): 509-516.
24. Holmes, G. K. and A. Muirhead (2017). "Epidemiology of coeliac disease in a single centre in Southern Derbyshire 1958–2014." BMJ open gastroenterology 4(1): e000137.
25. Hughes, K. R., L. C. Harnisch, C. Alcon-Giner, S. Mitra, C. J. Wright, J. Ketskemety, D. van Sinderen, A. J. Watson and L. Hall (2017). "Bifidobacterium breve reduces apoptotic epithelial cell shedding in an exopolysaccharide and MyD88-dependent manner." Open biology 7(1): 160155.
26. Kelly, S., S. Kramer, A. Schwede, P. Maini, K. Gull and M. Carrington (2012). "Genome organization is a major component of gene expression control in response to stress and during the cell division cycle in trypanosomes." Open Biology 2(4): 120033.

27. Kemp, A. M., S. A. Maguire, D. Nuttall, P. Collins, and F. Dunstan (2013). "Bruising in children who are assessed for suspected physical abuse." Archives of disease in childhood 99(2): 108-113.
28. Livermore, P., D. Eleftheriou and L. Wedderburn (2016). "The lived experience of juvenile idiopathic arthritis in young people receiving etanercept." Pediatric Rheumatology 14(1): 1-6.
29. McCahon, D., A. Roalfe and D. A. Fitzmaurice (2017). "An evaluation of a coagulation system (Xprecia Stride) for utilisation in anticoagulation management." Journal of clinical pathology 71(1): 20-26.
30. McDonagh, B., S. M. Scullion, A. Vasilaki, N. Pollock, A. McArdle and M. J. Jackson (2016). "Ageing-induced changes in the redox status of peripheral motor nerves imply an effect on redox signalling rather than oxidative damage." Free Radical Biology and Medicine 94: 27-35.
31. McEwan, P., B. L. Thorsted, M. Wolden, J. Jacobsen and M. Evans (2015). "Healthcare resource implications of hypoglycemia-related hospital admissions and inpatient hypoglycemia: retrospective record-linked cohort studies in England." BMJ Open Diabetes Research and Care 3(1): e000057.
32. McGrath, E., A. Jones and M. Field (2016). "Acute stress increases ad-libitum alcohol consumption in heavy drinkers, but not through impaired inhibitory control." Psychopharmacology 233(7): 1227-1234.
33. McMillan, T., P. McSkimming, J. Wainman-Lefley, L. Maclean, J. Hay, A. McConnachie and W. Stewart (2016). "Long-term health outcomes after exposure to repeated concussion in elite level: rugby union players." Journal of Neurology, Neurosurgery & Psychiatry 88(6): 505-511.
34. McWilliams, D. F., E. Ferguson, A. Young, P. D. Kiely and D. A. Walsh (2016). "Discordant inflammation and pain in early and established rheumatoid arthritis: Latent Class Analysis of Early Rheumatoid Arthritis Network and British Society for Rheumatology Biologics Register data." Arthritis research & therapy 18(1): 1-12.
35. Mercer, C. H., K. G. Jones, A. M. Johnson, R. Lewis, K. R. Mitchell, K. Gravningen, S. Clifton, C. Tanton, P. Sonnenberg, K. Wellings, J. A. Cassell and C. S. Estcourt (2016). "How can we objectively categorise partnership type? A novel classification of population survey data to inform epidemiological research and clinical practice." Sexually transmitted infections 93(2): 129-136.
36. Miller, K. A., S. R. F. Twigg, S. J. McGowan, J. M. Phipps, A. L. Fenwick, D. Johnson, S. A. Wall, P. Noons5, K. E. M. Rees, E. A. Tidey, J. Craft, J. Taylor, J. C. Taylor, J. A. C. Goos, S. M. A. Swagemakers, I. M. J. Mathijssen, P. J. v. d. Spek, H. Lord, T. Lester, N. Abid, D. Cilliers, J. A. Hurst, J. E. V. Morton, E. Sweeney, A. Weber, L. C. Wilson and A. O. M. Wilkie (2016). "Diagnostic value of exome and whole genome sequencing in craniosynostosis." Journal of medical genetics 54(4): 260-268.
37. Miners, A. H., C. D. Llewellyn, V. L. Cooper, E. Youssef, A. J. Pollard, M. Lagarde, C. Sabin, E. Nixon, M. Sachikonye, N. Perry and M. Fisher (2016). "A discrete choice experiment to assess people living with HIV's (PLWHIV's) preferences for GP or HIV clinic appointments." Sexually transmitted infections 93(2): 105-111.
38. Norris, T., K. Hawton, J. Hamilton-Shield and E. Crawley (2016). "Obesity in adolescents with chronic fatigue syndrome: an observational study." Archives of disease in childhood 102(1): 35-39.

39. O'Neill, F., M. Charakida, E. Topham, E. McLoughlin, N. Patel, E. Sutill, C. W. M. Kay, F. D'Aiuto, U. Landmesser, P. C. Taylor and J. Deanfield (2017). "Anti-inflammatory treatment improves high-density lipoprotein function in rheumatoid arthritis." Heart 103(10): 766-773.
40. Owen-Casey, M. P., R. Sim, H. T. Cook, C. A. Roufosse, J. D. Gillmore, J. A. Gilbertson, C. A. Hutchison, and A. J. Howie (2014). "Value of antibodies to free light chains in immunoperoxidase studies of renal biopsies." Journal of clinical pathology 67(8): 661-666.
41. Palfreyman, J., J. Graham-Brown, C. Caminade, P. Gilmore, D. Otranto and D. J. Williams (2018). "Predicting the distribution of Phortica variegata and potential for Thelazia callipaeda transmission in Europe and the United Kingdom." Parasites & vectors 11(1): 1-8.
42. Pearson, S., A. J. Williamson, R. Blance, T. C. Somervaille, S. Taylor, N. Azadbakht, A. D. Whetton and A. Pierce (2017). "Proteomic analysis of JAK2V617F-induced changes identifies potential new combinatorial therapeutic approaches." Leukemia 31(12): 2717-2725.
43. Raemdonck, K., K. Baker, N. Dale, E. Dubuis, F. Shala, M. G. Belvisi and M. A. Birrell (2016). "CD4+ and CD8+ T cells play a central role in a HDM driven model of allergic asthma." Respiratory research 17(1): 1-17.
44. Richman, S. D., J. Fairley, R. Butler and Z. C. Deans (2016). "How close are we to standardised extended RAS gene mutation testing? The UK NEQAS evaluation." Journal of clinical pathology 70(1): 58-62.
45. Salt, E., D. Van der Windt, L. Chesterton, F. Mainwaring, N. Ashwood and N. Foster (2017). "Physiotherapist-led suprascapular nerve blocks for persistent shoulder pain: Evaluation of a new service in the UK." Musculoskeletal care 16(1): 214-221.
46. Saxon, J. A., J. C. Thompson, M. Jones, J. M. Harris, A. M. Richardson, T. Langheinrich, D. Neary, D. M. Mann and J. S. Snowden (2017). "Examining the language and behavioural profile in FTD and ALS-FTD." Journal of Neurology, Neurosurgery & Psychiatry 88(8): 675-680.
47. Sinnett, C. G., D. P. Letley, G. L. Narayanan, S. R. Patel, N. R. Hussein, A. M. Zaitoun, K. Robinson and J. C. Atherton (2016). "Helicobacter pylori vacA transcription is genetically-determined and stratifies the level of human gastric inflammation and atrophy." Journal of Clinical Pathology 69(11): 968-973.
48. Slatter, D. A. and R. W. Farndale (2015). "Structural constraints on the evolution of the collagen fibril: convergence on a 1014-residue COL domain." Open biology 5(5): 140220.
49. Smith, A. D., K. Tilling, S. M. Nelson and D. A. Lawlor (2015). "Live-birth rate associated with repeat in vitro fertilization treatment cycles." Jama 314(24): 2654-2662.
50. Stiles, V. H., B. S. Metcalf, K. M. Knapp and A. V. Rowlands (2017). "A small amount of precisely measured high-intensity habitual physical activity predicts bone health in pre-and post-menopausal women in UK Biobank." International Journal of Epidemiology 46(6): 1847-1856.
51. Summers, J. A., J. Peacock, B. Coker, V. McMillan, M. Ofuya, C. Lewis, S. Keevil, R. Logan, J. McLaughlin and F. Reid (2016). "Multicentre prospective survey of SeHCAT provision and practice in the UK." BMJ open gastroenterology 3(1).
52. North, T. L., Ben-Shlomo, Y., Cooper, C., Deary, I. J., Gallacher, J., Kivimaki, M., Kumari, M., Martin, R. M., Pattie, A., Sayer, A. A., Starr, J. M., Wong, A., Kuh, D., Rodriguez, S.

& Day, I. N. (2015). "A study of common Mendelian disease carriers across ageing British cohorts: meta-analyses reveal heterozygosity for alpha 1-antitrypsin deficiency increases respiratory capacity and height." Journal of medical genetics 53(4): 280-288.

53. Toleman, M. S., E. R. Watkins, T. Williams, B. Blane, B. Sadler, E. M. Harrison, F. Coll, J. Parkhill, B. Nazareth, N. M. Brown, and S. J. Peacock (2017). "Investigation of a cluster of sequence type 22 methicillin-resistant Staphylococcus aureus transmission in a community setting." Clinical Infectious Diseases 65(12): 2069-2077.
54. Trump, N., A. McTague, H. Brittain, A. Papandreou, E. Meyer, A. Ngoh, R. Palmer, D. Morrogh, C. Boustred, J. A. Hurst, L. Jenkins, M. A. Kurian and R. H. Scott (2015). "Improving diagnosis and broadening the phenotypes in early-onset seizure and severe developmental delay disorders through gene panel analysis." Journal of medical genetics 53(5): 310-317.
55. Turnbull, A., M. Osborn and N. Nicholas (2015). "Hospital autopsy: endangered or extinct?" Journal of clinical pathology 68(8): 601-604.
56. Turner, P. J., M. H. Gowland, V. Sharma, D. Ierodiakonou, N. Harper, T. Garcez, R. Pumphrey and R. J. Boyle (2015). "Increase in anaphylaxis-related hospitalizations but no increase in fatalities: an analysis of United Kingdom national anaphylaxis data, 1992-2012." Journal of Allergy and Clinical Immunology 135(4): 956-963. e951.
57. Turney, B. W., J. M. Reynard, J. G. Noble and S. R. Keoghane (2011). "Trends in urological stone disease." BJU international 109(7): 1082-1087.
58. W.J. Rodgersa, J. Chatawayb, K. Schmiererc, D. Rogd, I. Galeae, A. Akbaria, K. Tuite-Daltona, H. Lockhart-Jonesa, D. Griffithsa, D.G. Noblea, K.H. Jonesa, A. Al-Dinf, M. Cranerg, N. Evangelouh, P. Harmani, T. Harrowerj, J. Hobartk, H. Husseyinl, M. Kastim, C. Kippsn, G. McDonnello, C. Owenp, O. Pearsonq, W. Rashidr, H. Wilsons and D. V. Forda (2018). "Validating the portal population of the United Kingdom multiple sclerosis register." Multiple sclerosis and related disorders 24: 3-10.
59. Watson CM, Crinnion LA, Murphy H, Newbould M, Harrison SM, Lascelles C, Antanaviciute A, Carr IM, Sheridan E, Bonthron DT and S. A (2015). "Deficiency of the myogenic factor MyoD causes a perinatally lethal fetal akinesia." Journal of medical genetics 53(4): 264-269.
60. Watson E, Shinkins B, Frith E, Neal D, Hamdy F, Walter F, Weller D, Wilkinson C, Faithfull S, Wolstenholme J, Sooriakumaran P, Kastner C, Campbell C, Neal R, Butcher H, Matthews M, Perera R and R. P. (2016). "Symptoms, unmet needs, psychological well-being and health status in survivors of prostate cancer: implications for redesigning follow-up." BJU international 117(6B): E10-E19.
61. Williams, M. C., A. Hunter, A. Shah, V. Assi, S. Lewis, K. Mangion, C. Berry, N. A. Boon, E. Clark, M. Flather, J. Forbes, S. McLean, G. Roditi, E. J. v. Beek, A. D. Timmis and D. E. Newby (2017). "Symptoms and quality of life in patients with suspected angina undergoing CT coronary angiography: a randomised controlled trial." Heart 103(13): 995-1001.
62. Wilson, E. C., J. A. Usher-Smith, J. Emery, P. Corrie and F. M. Walter (2018). "A modeling study of the cost-effectiveness of a risk-stratified surveillance program for melanoma in the United Kingdom." Value in Health 21(6): 658-668.
63. Woodhead, Z. V., J. Crinion, S. Teki, W. Penny, C. J. Price and A. P. Leff (2017). "Auditory training changes temporal lobe connectivity in ‘Wernicke’s aphasia’: a randomised trial." Journal of Neurology, Neurosurgery & Psychiatry 88(7): 586-594.

64. Woodhouse, J. M., N. Davies, A. McAvinchey and B. Ryan (2013). "Ocular and visual status among children in special schools in Wales: the burden of unrecognised visual impairment." Archives of disease in childhood 99(6): 500-504.
65. Woods, L. M., B. Rachet, D. O'Connell, G. Lawrence and M. P. Coleman (2016). "Impact of deprivation on breast cancer survival among women eligible for mammographic screening in the West Midlands (UK) and New South Wales (Australia): Women diagnosed 1997–2006." International journal of cancer 138(10): 2396-2403.

Appendix 4 – OSRAs comprising the EN-PT^{EU} corpus, in alphabetical order

1. Adragão, P., P. Carmo, D. Cavaco, J. Carmo, A. Ferreira, F. Moscoso Costa, M. Carvalho, J. Mesquita, R. Quaresma, F. Belo Morgado and M. Mendes (2017). "Relationship between rotors and complex fractionated electrograms in atrial fibrillation using a novel computational analysis." Revista Portuguesa de Cardiologia (English Edition) 36(4): 233-238.
2. Afonso, C., S. Costa, C. Cardoso, R. Oliveira, H. Lourenço, A. Viula, I. Batista, I. Coelho and M. Nunes (2015). "Benefits and risks associated with consumption of raw, cooked, and canned tuna (Thunnus spp.) based on the bioaccessibility of selenium and methylmercury." Environmental research 143: 130-137.
3. Alves, M. G., A. D. Martins, P. I. Moreira, R. A. Carvalho, M. Sousa, A. Barros, J. Silva, S. Pinto, T. Simões and P. F. Oliveira (2015). "Metabolic fingerprints in testicular biopsies from type 1 diabetic patients." Cell and tissue research 362(2): 431-440.
4. Areia, M., M. Dinis-Ribeiro and F. Rocha Gonçalves (2014). "Cost-utility analysis of endoscopic surveillance of patients with gastric premalignant conditions." Helicobacter 19(6): 425-436.
5. Azevedo, A., P. Braga, A. Rodrigues, L. Santos, B. Melica, J. Ribeiro, F. Sampaio, R. Fontes-Carvalho, M. Fonseca, A. Dias and V. Gama Ribeiro (2017). "Percutaneous closure of periprosthetic paravalvular leaks: A viable alternative to surgery?" Revista portuguesa de cardiologia 36(7-8): 489-494.
6. Barros-Barbosa, A. R., A. L. Fonseca, S. Guerra-Gomes, F. Ferreirinha, A. Santos, R. Rangel, M. G. Lobo, P. Correia-de-Sá and J. M. Cordeiro (2016). "Up-regulation of P2X7 receptor–mediated inhibition of GABA uptake by nerve terminals of the human epileptic neocortex." Epilepsia 57(1): 99-110.
7. Belino, C., A. Coelho, S. Pereira, D. Lopes, A. M. Gomes and A. Ventura (2017). "Survival of hemodialysis patients: A new reality?" Portuguese Journal of Nephrology & Hypertension 31(1): 37-41.
8. Belo, L., H. Nascimento, M. Kohlova, E. Bronze-da-Rocha, J. Fernandes, E. Costa, C. Catarino, L. Aires, H. F. Mansilha, P. Rocha-Pereira, A. Quintanilha, C. Rêgo and A. Santos-Silva (2014). "Body fat percentage is a major determinant of total bilirubin independently of UGT1A1* 28 polymorphism in young obese." PloS one 9(6): e98467.
9. Bettencourt, A., A. M. Silva, C. Carvalho, B. Leal, E. Santos, P. P. Costa and B. M. Silva (2014). "The role of KIR2DS1 in multiple sclerosis-KIR in Portuguese MS patients." Journal of neuroimmunology 269(1): 52-55.
10. Branco, J., A. Rodrigues, N. Gouveia, M. Eusébio, S. Ramiro, P. Machado, L. da Costa, A. Mourão, I. Silva, P. Laires, A. Sepriano, F. Araújo, S. Gonçalves, P. Coelho, V. Tavares, J. Cerol, J. Mendes, L. Carmona and H. Canhão (2016). "Prevalence of rheumatic and musculoskeletal diseases and their impact on health-related quality of life, physical function and mental health in Portugal: results from EpiReumaPt–a national health survey." RMD open 2(1): e000166.
11. Canena, J., M. Liberato, L. Meireles, I. Marques, C. Romão, A. P. Coutinho, B. C. Neves and P. M. Veiga (2015). "A non-randomized study in consecutive patients with postcholecystectomy refractory biliary leaks who were managed endoscopically with the use of multiple plastic stents or fully covered self-expandable metal stents (with videos)." Gastrointestinal endoscopy 82(1): 70-78.

12. Carvalho, C., S. Calvisi, B. Leal, A. Bettencourt, A. Marinho, I. Almeida, F. Farinha, P. Costa, B. Silva and C. Vasconcelos (2013). "CCR 5-D elta32: implications in SLE development." International journal of immunogenetics 41(3): 236-241.
13. Castro, A., C. Santos, H. Meireles, J. Silva and P. Teixeira (2015). "Food handlers as potential sources of dissemination of virulent strains of Staphylococcus aureus in the community." Journal of infection and public health 9(2): 153-160.
14. Correia-Costa, L., A. C. Afonso, F. Schaefer, J. T. Guimarães, M. Bustorff, A. Guerra, H. Barros and A. Azevedo (2015). "Decreased renal function in overweight and obese prepubertal children." Pediatric research 78(4): 436.
15. Correia-Melo, C., F. D. Marques, R. Anderson, G. Hewitt, R. Hewitt, J. Cole, B. M. Carroll, S. Miwa, J. Birch, A. Merz, M. D. Rushton, M. Charles, D. Jurk, S. W. Tait, R. Czapiewski, L. Greaves4, G. Nelson, M. Bohlooly-Y, S. Rodriguez-Cuenca, A. Vidal-Puig, D. Mann, G. Saretzki, G. Quarato, D. R. Green, P. D. Adams, T. v. Zglinicki, V. I. Korolchuk and J. F. Passos (2016). "Mitochondria are required for pro-ageing features of the senescent phenotype." The EMBO journal 35(7): 724-742.
16. Cunha, C., S. Pereira, J. C. Fernandes and V. P. Dias (2017). "24-hour ambulatory blood pressure monitoring in chronic kidney disease and its influence on treatment." Portuguese Journal of Nephrology & Hypertension 31(1): 31-36.
17. Cunha-Miranda, L., H. Santos, C. Miguel, C. Silva, F. Barcelos, J. Borges, R. Trinca, V. Vicente and T. Silva (2015). "Validation of Portuguese-translated computer touch-screen questionnaires in patients with rheumatoid arthritis and spondyloarthritis, compared with paper formats." Rheumatology international 35(12): 2029-2035.
18. David, S., A. Mateus, E. L. Duarte, J. Albuquerque, C. Portugal, L. Sancho, J. Lavinha and G. Gonçalves (2015). "Determinants of the sympatric host-pathogen relationship in tuberculosis." PloS one 10(11): e0140625.
19. De Jesus, B. B., S. P. Marinho, S. Barros, A. Sousa-Franco, C. Alves-Vale, T. Carvalho and M. Carmo-Fonseca (2017). "Silencing of the lncRNA Zeb2-NAT facilitates reprogramming of aged fibroblasts and safeguards stem cell pluripotency." Nature communications 9(1): 1-11.
20. Diana Teixeira, Diogo Pestana, Cristina Santos, Luísa Correia-Sá, Cláudia Marques, Sónia Norberto, Manuela Meireles, Ana Faria, Ricardo Silva, Gil Faria, Carla Sá, Paula Freitas, António Taveira-Gomes, Valentina Domingues, Cristina Delerue-Matos, Conceição Calhau and R. Monteiro (2015). "Inflammatory and cardiometabolic risk on obesity: role of environmental xenoestrogens." The Journal of Clinical Endocrinology & Metabolism 100(5): 1792-1801.
21. Dias, J. L., J. M. Pina, N. V. Costa, S. Carmo, C. Leal, T. Bilhim, R. M. Marques and L. C. Pinheiro (2016). "The utility of apparent diffusion coefficient values in the risk stratification of prostate cancer using a 1.5 T magnetic resonance imaging without endorectal coil." Acta Urológica Portuguesa 33(3): 81-86.
22. Duarte-Pereira, S., I. Pereira-Castro, S. S. Silva, M. G. Correia, C. Neto, L. T. da Costa, A. Amorim and R. M. Silva (2016). "Extensive regulation of nicotinate phosphoribosyltransferase (NAPRT) expression in human tissues and tumors." Oncotarget 7(2): 1973.
23. Espinar, M. J., I. M. Miranda, S. Costa-de-Oliveira, R. Rocha, A. G. Rodrigues and C. Pina-Vaz (2015). "Urinary tract infections in kidney transplant patients due to Escherichia Coli and Klebsiella Pneumoniae-producing extended-spectrum β-lactamases: risk factors and molecular epidemiology." PloS one 10(8): e0134737.

24. Faria, G., A. Gonçalves, R. Cunha, J. Guimarães, C. Calhau, J. Preto and A. Taveira-Gomes (2015). "Beyond central adiposity: liver fat and visceral fat area are associated with metabolic syndrome in morbidly obese patients." International Journal of Surgery 14: 75-79.
25. Fernandes, A., N. Almeida, A. M. Ferreira, A. Casela, D. Gomes, F. Portela, E. Camacho and C. Sofia (2015). "Left-Sided Portal Hypertension: A Sinister Entity." GE Portuguese journal of gastroenterology 22(6): 234-239.
26. Ferreira, P. C., J. Barbosa, J. M. Amarante, J. Carvalho, A. G. Rodrigues and A. C. Silva (2015). "Associated injuries in pediatric patients with facial fractures in Portugal: Analysis of 1416 patients." Journal of Cranio-Maxillofacial Surgery 43(4): 437-443.
27. Fontes, F., M. Gonçalves, S. Maia, S. Pereira, M. Severo and N. Lunet (2017). "Reliability and validity of the Pittsburgh Sleep Quality Index in breast cancer patients." Supportive Care in Cancer 25(10): 3059-3066.
28. Garrido, P., L. M. Pedro, R. F. e Fernandes, L. Silvestre, G. Sousa, C. Martins and J. F. e Fernandes (2016). "Endovascular treatment of synchronous and metachronous aneurysms of the thoracic aorta. Is there an increase in the procedural risk?" Angiologia e Cirurgia Vascular 12(4): 226-233.
29. Gonçalves, G., J. Frade, M. Nascimento, J. R. Mesquita and C. Nunes (2016). "Persistence of rubella and mumps antibodies, following changes in the recommended age for the second dose of MMR vaccine in Portugal." Epidemiology & Infection 144(15): 3139-3147.
30. Jesus, T. T., P. F. Oliveira, J. Silva, A. Barros, R. Ferreira, M. Sousa, C. Y. Cheng, B. M. Silva and M. G. Alves (2016). "Mammalian target of rapamycin controls glucose consumption and redox balance in human Sertoli cells." Fertility and sterility 105(3): 825-833. e823.
31. Ladeiras-Lopes, R., N. Bettencourt, N. Ferreira, F. Sampaio, G. Pires-Morais, L. Santos, B. Melica, A. Rodrigues, P. Braga, A. Leite-Moreira, J. Silva-Cardoso and V. Gama (2016). "CT myocardial perfusion and coronary CT angiography: Influence of coronary calcium on a stress–rest protocol." Journal of cardiovascular computed tomography 10(3): 215-220.
32. Lopes, F., M. Barbosa, A. Ameur, G. Soares, J. de Sá, A. Dias, G. Oliveira, P. Cabral, T. Temudo, E. Calado, I. Cruz, J. Vieira, R. Oliveira, S. Esteves, S. Sauer, I. Jonasson, A. Syvänen, U. Gyllensten, D. Pinto and P. Maciel (2016). "Identification of novel genetic causes of Rett syndrome-like phenotypes." Journal of medical genetics 53(3): 190-199.
33. Magriço, R., J. P. Santos, S. Colaço, S. Dias and A. Ramos (2017). "Recombinant tissue plasminogen activator plus citrate versus citrate alone as catheter lock for tunnelled catheters of haemodialysis." Portuguese Journal of Nephrology & Hypertension 31(2): 100-107.
34. Marques, I., M. Moura, R. Cabrera, A. Pinto, J. Simões-Pereira, C. Santos, F. Menezes, D. Montezuma, R. Henrique, M. Rodrigues Teixeira, V. Leite and B. Cavaco (2017). "Identification of somatic TERT promoter mutations in familial nonmedullary thyroid carcinomas." Clinical endocrinology 87(4): 394-399.
35. Martins, M., J. Boavida, J. Raposo, F. Froes, B. Nunes, R. Ribeiro, M. Macedo and C. Penha-Gonçalves (2016). "Diabetes hinders community-acquired pneumonia outcomes in hospitalized patients." BMJ Open Diabetes Research and Care 4(1).

36. Mesquita, I., H. Sousa, F. Carvalho and F. Nolasco (2017). "Renal pathology in HCV infected patients-Report of 148 patients and review of the literature." Portuguese Journal of Nephrology & Hypertension 31(2): 91-99.
37. Miranda, A. M., C. Garcia, V. A. Bento and S. Pinto (2017). "Urinary tract infections under 24 months old: Is it possible to predict the risk of renal scarring?" Portuguese Journal of Nephrology & Hypertension 31(2): 224-230.
38. Monteiro, M., N. Moreira, J. Pinto, A. S. Pires-Luís, R. Henrique, C. Jerónimo, M. d. L. Bastos, A. M. Gil, M. Carvalho and P. Guedes de Pinho (2017). "GC-MS metabolomics-based approach for the identification of a potential VOC-biomarker panel in the urine of renal cell carcinoma patients." Journal of cellular and molecular medicine 21(9): 2092-2105.
39. Moura, J., J. Rodrigues, M. Gonçalves, C. Amaral, M. Lima and E. Carvalho (2017). "Impaired T-cell differentiation in diabetic foot ulceration." Cellular & molecular immunology 14(9): 758-769.
40. Nascimento, H., E. Costa, S. Rocha, C. Lucena, P. Rocha-Pereira, C. Rêgo, H. F. Mansilha, A. Quintanilha, L. Aires, J. Mota, A. Santos-Silva and L. Belo (2014). "Adiponectin and markers of metabolic syndrome in obese children and adolescents: impact of 8-mo regular physical exercise program." Pediatric research 76(2): 159.
41. Nora, M., M. Guimarães, R. Almeida, P. Martins, G. Gonçalves, M. Santos, T. Morais, C. Freitas and M. P. Monteiro (2014). "Excess body mass index loss predicts metabolic syndrome remission after gastric bypass." Diabetology & metabolic syndrome 6(1): 1-7.
42. Oliveira-Ramos, F., M. Eusébio, F. M Martins, A. Mourão, C. Furtado, R. Campanilho-Marques, I. Cordeiro, J. Ferreira, M. Cerqueira, R. Figueira, I. Brito, H. Canhão, M. Santos, J. Melo-Gomes and J. Fonseca (2016). "Juvenile idiopathic arthritis in adulthood: fulfilment of classification criteria for adult rheumatic diseases, long-term outcomes and predictors of inactive disease, functional status and damage." RMD open 2(2): e000304.
43. Patricio, A., D. F. Cruz, J. V. Silva, A. Padrão, B. R. Correia, L. Korrodi-Gregório, R. Ferreira, N. Maia, S. Almeida, J. Lourenço, V. Silva and M. Fardilha (2016). "Relation between seminal quality and oxidative balance in sperm cells." Acta Urológica Portuguesa 33(1): 6-15.
44. Pena, M., M. de Almeida, E. van Dam, K. Ahring, A. Bélanger-Quintana, K. Dokoupil, H. Gokmen-Ozel, A. Lammardo, A. MacDonald, M. Robert and J. Rocha (2016). "Protein substitutes for phenylketonuria in Europe: access and nutritional composition." European journal of clinical nutrition 70(7): 785-789.
45. Pereira, C., S. Queirós, A. Galaghar, H. Sousa, P. Pimentel-Nunes, C. Brandão, L. Moreira-Dias, R. Medeiros and M. Dinis-Ribeiro (2014). "Genetic variability in key genes in prostaglandin E2 pathway (COX-2, HPGD, ABCC4 and SLCO2A1) and their involvement in colorectal cancer development." PLoS One 9(4): e92000.
46. Pereira, C. D., M. Severo, J. R. Araújo, J. T. Guimarães, D. Pestana, A. Santos, R. Ferreira, A. Ascensão, J. Magalhães, I. Azevedo, R. Monteiro and M. J. Martins (2014). "Relevance of a hypersaline sodium-rich naturally sparkling mineral water to the protection against metabolic syndrome induction in fructose-fed sprague-dawley rats: A biochemical, metabolic, and redox approach." International journal of endocrinology 2014.

47. Pestana, D., G. Faria, C. Sá, V. Fernandes, D. Teixeira, S. Norberto, A. Faria, M. Meireles, C. Marques, L. Correia-Sá, A. Cunha, J. Guimarães, A. Taveira-Gomes, A. Santos, V. Domingues, C. Delerue-Matos, R. Monteiro and C. Calhau (2014). "Persistent organic pollutant levels in human visceral and subcutaneous adipose tissue in obese individuals—Depot differences and dysmetabolism implications." Environmental research 133: 170-177.
48. Póvoa, P., I. Martin-Loeches, P. Ramirez, L. Bos, M. Esperatti, J. Silvestre, G. Gili, G. Goma, E. Berlanga, M. Espasa, E. Gonçalves, A. Torres and A. Artigas (2016). "Biomarker kinetics in the prediction of VAP diagnosis: results from the BioVAP study." Annals of intensive care 6(1): 1-11.
49. Queirós, P., H. Pinheiro, J. Carvalho, P. Oliveira, I. Gullo, F. Carneiro, G. M. Almeida and C. Oliveira (2015). "KRAS mutations in microsatellite instable gastric tumours: impact of targeted treatment and intratumoural heterogeneity." Virchows Archiv 467(4): 383-392.
50. Rangel, I., A. Gonçalves, C. De Sousa, S. Leite, M. Campelo, E. Martins, S. Amorim, B. Moura, J. S. Cardoso and M. J. Maciel (2014). "Iron deficiency status irrespective of anemia: a predictor of unfavorable outcome in chronic heart failure patients." Cardiology 128(4): 320-326.
51. Rato, L., M. G. Alves, A. I. Duarte, M. S. Santos, P. I. Moreira, J. E. Cavaco and P. F. Oliveira (2015). "Testosterone deficiency induced by progressive stages of diabetes mellitus impairs glucose metabolism and favors glycogenesis in mature rat Sertoli cells." The international journal of biochemistry & cell biology 66: 1-10.
52. Ribeiro, I., R. Pinho, A. Rodrigues, J. Silva, A. Ponte, J. Rodrigues and J. Carvalho (2015). "What is the long-term outcome of a negative capsule endoscopy in patients with obscure gastrointestinal bleeding?" Revista Española de Enfermedades Digestivas 107(12): 753-758.
53. Ribeiro, J., M. Malta, A. Galaghar, F. Silva, L. P. Afonso, R. Medeiros and H. Sousa (2017). "P53 deregulation in Epstein-Barr virus-associated gastric cancer." Cancer letters 404: 37-43.
54. Ribeiro, R., A. P. Araújo, A. Coelho, R. Catarino, D. Pinto, A. Araújo, C. Calçada, C. Lopes and R. Medeiros (2006). "A functional polymorphism in the promoter region of leptin gene increases susceptibility for non-small cell lung cancer." European Journal of Cancer 42(8): 1188-1193.
55. Ribeiro-Rodrigues, T., T. Laundos, R. Pereira-Carvalho, D. Batista-Almeida, R. Pereira, V. Coelho-Santos, A. Silva, R. Fernandes, M. Zuzarte, F. Enguita, M. Costa, P. Pinto-do-Ó, M. Pinto, P. Gouveia, L. Ferreira, J. Mason, P. Pereira, B. Kwak, D. Nascimento and H. Girão (2017). "Exosomes secreted by cardiomyocytes subjected to ischaemia promote cardiac angiogenesis." Cardiovascular Research 113(11): 1338-1350.
56. Rodrigues, J., J. Carmo, L. Carvalho, P. Barreiro and C. Chagas (2015). "Endoscopic submucosal dissection for gastrointestinal superficial lesions: initial experience in a single Portuguese center." GE Portuguese journal of gastroenterology 22(5): 190-197.
57. Santos, M. P. C., C. Palmela, R. Ferreira, E. Barjas, A. A. Santos, R. Maio and M. Cravo (2016). "Self-expandable metal stents for colorectal cancer: from guidelines to clinical practice." GE Portuguese journal of gastroenterology 23(6): 293-299.
58. Silva, A. P., F. Mendes, A. Fragoso, T. Jeronimo, A. Pimentel, K. Gundlach, J. Büchel, N. Santos and P. L. Neves (2015). "Altered serum levels of FGF-23 and magnesium are independent risk factors for an increased albumin-to-creatinine ratio in type 2

diabetics with chronic kidney disease." Journal of diabetes and its complications 30(2): 275-280.

59. Silva, I., F. Ferreirinha, M. T. Magalhães-Cardoso, M. Silva-Ramos and P. Correia-de-Sá (2015). "Activation of P2Y6 receptors facilitates nonneuronal adenosine triphosphate and acetylcholine release from urothelium with the lamina propria of men with bladder outlet obstruction." The Journal of urology 194(4): 1146-1154.
60. Silva, J., F. Cerqueira and R. Medeiros (2015). "Y chromosome DNA in cervicovaginal self-collected samples of childbearing age women: Implications for epitheliotropic sexually transmitted infections?" Life sciences 139: 62-68.
61. Sílvia, A., Pires-Luísa, F. Lobo, M. Vieira-Coimbra, P. Costa-Pinheiro, L. Antunes, J. Oliveira, R. Henrique and C. Jerónimo (2015). "MST1R methylation as a diagnostic biomarker in renal cell tumors." Acta Urológica Portuguesa 32(2): 64-70.
62. Simões, T., A. Queirós, A. T. Marujo, S. Valdoleiros, P. Silva and I. Blickstein (2015). "Outcome of monochorionic twins conceived by assisted reproduction." Fertility and sterility 104(3): 629-632.
63. Tavares-Silva, M., M. Alaa, S. Leite, J. Oliveira-Pinto, L. Lopes, A. F. Leite-Moreira and A. P. Lourenço (2017). "Dose–Response Head-to-Head Comparison of Inodilators Dobutamine, Milrinone, and Levosimendan in Chronic Experimental Pulmonary Hypertension." Journal of cardiovascular pharmacology and therapeutics 22(5): 485-495.
64. Teixeira, D., C. Marques, D. Pestana, A. Faria, S. Norberto, C. Calhau and R. Monteiro (2016). "Effects of xenoestrogens in human M1 and M2 macrophage migration, cytokine release, and estrogen-related signaling pathways." Environmental toxicology 31(11): 1496-1509.
65. Timóteo, M. A., I. Carneiro, I. Silva, J. B. Noronha-Matos, F. Ferreirinha, M. Silva-Ramos and P. Correia-de-Sá (2014). "ATP released via pannexin-1 hemichannels mediates bladder overactivity triggered by urothelial P2Y6 receptors." Biochemical pharmacology 87(2): 371-379.

Appendix 5 – OSRAs comprising the EN-ES^{EU} corpus, in alphabetical order

1. Aguirre-Bermeo, H., I. Morán, M. Bottiroli, S. Italiano, F. J. Parrilla, E. Plazolles, F. Roche-Campo and J. Mancebo (2016). "End-inspiratory pause prolongation in acute respiratory distress syndrome patients: effects on gas exchange and mechanics." Annals of intensive care 6(1): 1-8.
2. Alcubierre, N., E. M. Navarrete-Muñoz, E. Rubinat, M. Falguera, J. Valls, A. Traveset, M.-B. Vilanova, J. R. Marsal, M. Hernandez and M. Granado-Casas (2016). "Association of low oleic acid intake with diabetic retinopathy in type 2 diabetic patients: a case–control study." Nutrition & metabolism 13(1): 1-7.
3. Alegret, J. M., N. Martínez-Micaelo, G. Aragonès and R. Beltrán-Debón (2016). "Circulating endothelial microparticles are elevated in bicuspid aortic valve disease and related to aortic dilation." International journal of cardiology 217: 35-41.
4. Alfonso, M., L. R. F. Faro, I. Oliveira and R. Duran (2015). "Mediation of a glutamate antagonist, a NOS inhibitor and antioxidants with–SH groups on striatal dopamine release induced by clothianidin." Revista de Toxicología 32(2): 135-139.
5. Almazán, M. V., E. Ortega, R. M. Torres, M. Tovar, J. Romero, M. Á. López-Casado, L. Jáimez, J. Jiménez-Jáimez, A. Ballesteros and J. Caballero-Villarraso (2015). "Diagnostic screening for subclinical celiac disease using a rapid test in children aged 2–4." Pediatric research 78(3): 280-285.
6. Aras, L. M., J. Isla and A. Mingorance-Le Meur (2015). "The European patient with Dravet syndrome: results from a parent-reported survey on antiepileptic drug use in the European population with Dravet syndrome." Epilepsy & Behavior 44: 104-109.
7. Arrebola, J. P., R. Ocaña-Riola, A. L. Arrebola-Moreno, M. Fernández-Rodríguez, P. Martin-Olmedo, M. F. Fernández and N. Olea (2014). "Associations of accumulated exposure to persistent organic pollutants with serum lipids and obesity in an adult cohort from Southern Spain." Environmental pollution 195: 9-15.
8. Artacho-Cordon, F., M. Fernández-Rodríguez, C. Garde, E. Salamanca, L. Iribarne-Durán, P. Torné, J. Expósito, L. Papay-Ramírez, M. Fernández and N. Olea (2015). "Serum and adipose tissue as matrices for assessment of exposure to persistent organic pollutants in breast cancer patients." Environmental research 142: 633-643.
9. Bárez-López, S., A. Montero-Pedrazuela, D. Bosch-García, C. Venero and A. Guadaño-Ferraz (2017). "Increased anxiety and fear memory in adult mice lacking type 2 deiodinase." Psychoneuroendocrinology 84: 51-60.
10. Barrio, J., C. L. Errando, G. San Miguel, B. I. Salas, J. Raga, J. L. Carrión, J. García-Ramón and J. Gallego (2016). "Effect of depth of neuromuscular blockade on the abdominal space during pneumoperitoneum establishment in laparoscopic surgery." Journal of clinical anesthesia 34: 197-203.
11. Calvo, D., J. P. Flórez, I. Valverde, J. Rubín, D. Pérez, M. G. Vasserot, J. Rodríguez-Reguero, P. Avanzas, J. M. de la Hera and J. Gómez (2016). "Surveillance after cardiac arrest in patients with Brugada syndrome without an implantable defibrillator: an alarm effect of the previous syncope." International journal of cardiology 218: 69-74.
12. Casado, J., M. Sánchez, V. Garcés, L. Manzano, J. M. Cerqueiro, F. Epelde, D. García-Escrivá, J. Pérez-Silvestre, J. L. Morales and M. Montero-Pérez-Barquero (2017). "Influence of renal dysfunction phenotype on mortality in decompensated heart

failure with preserved and mid-range ejection fraction." International journal of cardiology 243: 332-339.

13. Cremades, A., J. Rio-Garcia, A. Lambertos, C. López-Garcia and R. Peñafiel (2016). "Tissue-specific regulation of potassium homeostasis by high doses of cationic amino acids." SpringerPlus 5(1): 616.
14. Crespo, I., B. San-Miguel, A. Fernández, J. O. de Urbina, J. González-Gallego and M. J. Tuñón (2015). "Melatonin limits the expression of profibrogenic genes and ameliorates the progression of hepatic fibrosis in mice." Translational research 165(2): 346-357.
15. Currás-Freixes, M., L. Inglada-Pérez, V. Mancikova, C. Montero-Conde, R. Letón, I. Comino-Méndez, M. Apellániz-Ruiz, L. Sánchez-Barroso, M. A. Sánchez-Covisa and V. Alcázar (2015). "Recommendations for somatic and germline genetic testing of single pheochromocytoma and paraganglioma based on findings from a series of 329 patients." Journal of medical genetics: jmedgenet-2015-103218.
16. Domingo-Calap, P., J. M. Cuevas and R. Sanjuán (2009). "The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages." PLoS Genet 5(11): e1000742.
17. Escribà-Garcia, L., C. Alvarez-Fernández, M. Tellez-Gabriel, J. Sierra and J. Briones (2017). "Dendritic cells combined with tumor cells and α-galactosylceramide induce a potent, therapeutic and NK-cell dependent antitumor immunity in B cell lymphoma." Journal of translational medicine 15(1): 1-11.
18. Fernandez-Lopez, R., I. del Campo, C. Revilla, A. Cuevas and F. de la Cruz (2014). "Negative feedback and transcriptional overshooting in a regulatory network for horizontal gene transfer." PLoS genetics 10(2): e1004171.
19. Forés, M., L. Simón-Carrasco, L. Ajuria, N. Samper, S. González-Crespo, M. Drosten, M. Barbacid and G. Jiménez (2017). "A new mode of DNA binding distinguishes Capicua from other HMG-box factors and explains its mutation patterns in cancer." PLoS genetics 13(3): e1006622.
20. Fuster-García, C., G. García-García, E. González-Romero, T. Jaijo, M. D. Sequedo, C. Ayuso, R. P. Vázquez-Manrique, J. M. Millán and E. Aller (2017). "USH2A gene editing using the CRISPR system." Molecular Therapy-Nucleic Acids 8: 529-541.
21. García-Cerro, S., N. Rueda, V. Vidal, S. Lantigua and C. Martínez-Cué (2017). "Normalizing the gene dosage of Dyrk1A in a mouse model of Down syndrome rescues several Alzheimer's disease phenotypes." Neurobiology of disease 106: 76-88.
22. Gascon, M., M. Casas, E. Morales, D. Valvi, A. Ballesteros-Gómez, N. Luque, S. Rubio, N. Monfort, R. Ventura and D. Martínez (2015). "Prenatal exposure to bisphenol A and phthalates and childhood respiratory tract infections and allergy." Journal of Allergy and Clinical Immunology 135(2): 370-378. e377.
23. Gil-Cayuela, C., E. Roselló-LLetí, E. Tarazón, A. Ortega, J. Sandoval, L. Martínez-Dolz, J. Cinca, E. Jorge, J. R. González-Juanatey and F. Lago (2017). "Thyroid hormone biosynthesis machinery is altered in the ischemic myocardium: An epigenomic study." International journal of cardiology 243: 27-33.
24. Gonzalez-Fernandez, C., A. Arevalo-Martin, B. Paniagua-Torija, I. Ferrer, F. J. Rodriguez and D. Garcia-Ovejero (2017). "Wnts are expressed in the ependymal region of the adult spinal cord." Molecular neurobiology 54(8): 6342-6355.
25. Heredia-Rodríguez, M., M. T. Peláez, I. Fierro, E. Gómez-Sánchez, E. Gómez-Pesquera, M. Lorenzo, F. J. Álvarez-González, J. Bustamante-Munguira, J. M. Eiros and J. F.

Bermejo-Martin (2016). "Impact of ventilator-associated pneumonia on mortality and epidemiological features of patients with secondary peritonitis." Annals of intensive care 6(1): 1-9.

26. Hernández-Alvarez, N., J. M. P. Acevedo, E. Quintero, I. F. Vázquez, M. García-Eliz, J. de la Revilla Negro, J. C. García and M. Hernández-Guerra (2017). "Effect of season and sunlight on viral kinetics during hepatitis C virus therapy." BMJ open gastroenterology 4(1): e000115.
27. Jiménez-López, E., A. I. Aparicio, E. M. Sánchez-Morla, R. Rodríguez-Jiménez, E. Vieta and J. L. Santos (2017). "Neurocognition in patients with psychotic and non-psychotic bipolar I disorder. A comparative study with individuals with schizophrenia." Journal of affective disorders 222: 169-176.
28. Lopez-Legarrea, P., R. de la Iglesia, I. Abete, I. Bondia-Pons, S. Navas-Carretero, L. Forga, J. A. Martinez and M. A. Zulet (2013). "Short-term role of the dietary total antioxidant capacity in two hypocaloric regimes on obese with metabolic syndrome symptoms: the RESMENA randomized controlled trial." Nutrition & metabolism 10(1): 1-11.
29. López-Valcárcel, B. G., J. Librero, A. García-Sempere, L. M. Peña, S. Bauer, J. Puig-Junoy, J. Oliva, S. Peiró and G. Sanfélix-Gimeno (2017). "Effect of cost sharing on adherence to evidence-based medications in patients with acute coronary syndrome." Heart 103(14): 1082-1088.
30. Lorente, L., M. M. Martín, A. Pérez-Cejas, P. Abreu-González, L. Ramos, M. Argueso, J. J. Cáceres, J. Solé-Violán and A. Jiménez (2016). "Association between total antioxidant capacity and mortality in ischemic stroke patients." Annals of intensive care 6(1): 1-6.
31. Martinez, F., P. Marín-Reina, A. Sanchis-Calvo, A. Perez-Aytés, S. Oltra, M. Roselló, S. Mayo, S. Monfort, J. Pantoja and C. Orellana (2015). "Novel mutations of NFIX gene causing Marshall-Smith syndrome or Sotos-like syndrome: one gene, two phenotypes." Pediatric research 78(5): 533-539.
32. Martinez-Hervás, S., M. M. Mendez, J. Folgado, C. Tormos, P. Ascaso, M. Peiró, J. T. Real and J. F. Ascaso (2017). "Altered Semmes–Weinstein monofilament test results are associated with oxidative stress markers in type 2 diabetic subjects." Journal of translational medicine 15(1): 1-8.
33. Martinez-Pinilla, E., A. Navarro, C. Ordonez, E. del Valle and J. Tolivia (2015). "Apolipoprotein D subcellular distribution pattern in neuronal cells during oxidative stress." Acta histochemica 117(6): 536-544.
34. Martín-Merino, E., C. Huerta-Álvarez, D. Prieto-Alhambra and D. Montero-Corominas (2017). "Cessation rate of anti-osteoporosis treatments and risk factors in Spanish primary care settings: a population-based cohort analysis." Archives of osteoporosis 12(1): 39.
35. Mata, A., L. Urrea, S. Vilches, F. Llorens, K. Thüne, J.-C. Espinosa, O. Andréoletti, A. M. Sevillano, J. M. Torres and J. R. Requena (2017). "Reelin expression in Creutzfeldt-Jakob disease and experimental models of transmissible spongiform encephalopathies." Molecular neurobiology 54(8): 6412-6425.
36. Mazzeo, C., J. A. Cañas, M. P. Zafra, A. R. Marco, M. Fernández-Nieto, V. Sanz, M. Mittelbrunn, M. Izquierdo, F. Baixaulli and J. Sastre (2015). "Exosome secretion by eosinophils: a possible role in asthma pathogenesis." Journal of Allergy and Clinical Immunology 135(6): 1603-1613.

37. Meca-Cortés, O., M. Guerra-Rebollo, C. Garrido, S. Borrós, N. Rubio and J. Blanco (2017). "CRISPR/Cas9-mediated knockin application in cell therapy: a non-viral procedure for bystander treatment of glioma in mice." Molecular Therapy-Nucleic Acids 8: 395-403.
38. Mela, V., O. Hernandez, C. Hunsche, F. Diaz, J. A. Chowen and M. De la Fuente (2017). "Administration of a leptin antagonist during the neonatal leptin surge induces alterations in the redox and inflammatory state in peripubertal/adolescent rats." Molecular and cellular endocrinology 454: 125-134.
39. Menendez, C., P. Castillo, M. J. Martínez, D. Jordao, L. Lovane, M. R. Ismail, C. Carrilho, C. Lorenzoni, F. Fernandes and T. Nhampossa (2017). "Validity of a minimally invasive autopsy for cause of death determination in stillborn babies and neonates in Mozambique: an observational study." PLoS medicine 14(6): e1002318.
40. Milara, J., J. Lluch, P. Almudever, J. Freire, Q. Xiaozhong and J. Cortijo (2014). "Roflumilast N-oxide reverses corticosteroid resistance in neutrophils from patients with chronic obstructive pulmonary disease." Journal of Allergy and Clinical Immunology 134(2): 314-322. e319.
41. Molina-Infante, J., A. Arias, J. Barrio, J. Rodríguez-Sánchez, M. Sanchez-Cazalilla and A. J. Lucendo (2014). "Four-food group elimination diet for adult eosinophilic esophagitis: a prospective multicenter study." Journal of Allergy and Clinical Immunology 134(5): 1093-1099. e1091.
42. Moreno-Indias, I., W. Oliva-Olivera, A. Omiste, D. Castellano-Castillo, S. Lhamyani, A. Camargo and F. J. Tinahones (2016). "Adipose tissue infiltration in normal-weight subjects and its impact on metabolic function." Translational Research 172: 6-17. e13.
43. Moya-Perez, A., A. Perez-Villalba, A. Benitez-Paez, I. Campillo and Y. Sanz (2017). "Bifidobacterium CECT 7765 modulates early stress-induced immune, neuroendocrine and behavioral alterations in mice." Brain, behavior, and immunity 65: 43-56.
44. Muñoz, M., M. Rosso, A. Carranza and R. Coveñas (2017). "Increased nuclear localization of substance P in human gastric tumor cells." Acta histochemica 119(3): 337-342.
45. Muñoz, M., M. Rosso and R. Coveñas (2017). "The NK-1 receptor antagonist L-732,138 induces apoptosis in human gastrointestinal cancer cell lines." Pharmacological Reports 69(4): 696-701.
46. Nogués, X., D. Prieto-Alhambra, R. Güerri-Fernández, N. Garcia-Giralt, J. Rodriguez-Morera, L. Cos, L. Mellibovsky and A. D. Pérez (2017). "Fracture during oral bisphosphonate therapy is associated with deteriorated bone material strength index." Bone 103: 64-69.
47. Orriach, J. G., M. G. Ortega, A. R. Fernandez, M. R. Aliaga, M. M. Cortes, D. A. Villanueva, A. F. Vela, J. A. Torres, C. S. Fernandez and E. M. Gonzalez (2017). "Cardioprotective efficacy of sevoflurane vs. propofol during induction and/or maintenance in patients undergoing coronary artery revascularization surgery without pump: A randomized trial." International journal of cardiology 243: 73-80.
48. Padilla-Fernández, B., M. García-Cenador, P. Rodríguez-Marcos, J. López-Marcos, P. Antúnez-Plaza, J. Silva-Abuín, D. López-Montañés, F. García-Criado and M. Lorenzo-Gómez (2017). "Experimental murine model of renal cancer." Actas Urológicas Españolas (English Edition) 41(7): 445-450.
49. Peñalver, A., J. A. Campos-Sandoval, E. Blanco, C. Cardona, L. Castilla, M. Martín-Rufián, G. Estivill-Torrús, R. Sánchez-Varo, F. J. Alonso and M. Pérez-Hernández (2017).

"Glutaminase and MMP-9 downregulation in cortex and hippocampus of LPA1 receptor null mice correlate with altered dendritic spine plasticity." Frontiers in molecular neuroscience 10: 278.

50. Redondo, N., E. Nova, A. Gheorghe, L. E. Díaz, A. Hernández and A. Marcos (2017). "Evaluation of Lactobacillus coryniformis CECT5711 strain as a coadjuvant in a vaccination process: a randomised clinical trial in healthy adults." Nutrition & metabolism 14(1): 1-9.
51. Revuelta-López, E., C. Soler-Botija, L. Nasarre, A. Benitez-Amaro, D. de Gonzalo-Calvo, A. Bayes-Genis and V. Llorente-Cortés (2016). "Relationship among LRP1 expression, Pyk2 phosphorylation and MMP-9 activation in left ventricular remodelling after myocardial infarction." Journal of cellular and molecular medicine 21(9): 1915-1928.
52. Ronzoni, G., A. Del Arco, F. Mora and G. Segovia (2016). "Enhanced noradrenergic activity in the amygdala contributes to hyperarousal in an animal model of PTSD." Psychoneuroendocrinology 70: 1-9.
53. Ruiz-Andres, O., M. D. Sanchez-Niño, P. Cannata-Ortiz, M. Ruiz-Ortega, J. Egido, A. Ortiz and A. B. Sanz (2016). "Histone lysine crotonylation during acute kidney injury in mice." Disease models & mechanisms 9(6): 633-645.
54. Ruiz-García, R., S. Mora, G. Lozano-Sánchez, L. Martínez-Lostao, E. Paz-Artal, J. Ruiz-Contreras, A. Anel, L. I. González-Granado, D. Moreno-Pérez and L. M. Allende (2015). "Decreased activation-induced cell death by EBV-transformed B-cells from a patient with autoimmune lymphoproliferative syndrome caused by a novel FASLG mutation." Pediatric research 78(6): 603-608.
55. Inogés, S., Tejada, S., de Cerio, A. L. D., Pérez-Larraya, J. G., Espinós, J., Idoate, M. A., . Domínguez, P.D., García de Eulate, R., Aristu, J., Bendandi, M., Pastor, F., Alonso, M., Andreu, E., Prósper Cardoso, F., and & Valle, R. D. (2017). A phase II trial of autologous dendritic cell vaccination and radiochemotherapy following fluorescence-guided surgery in newly diagnosed glioblastoma patients. Journal of translational medicine, 15(1), 1-12.
56. Sánchez, G. F. L., S. G. Víllora and A. D. Suárez (2016). "Level of habitual physical activity in children and adolescents from the Region of Murcia (Spain)." SpringerPlus 5(1): 386.
57. Sánchez-López, V., V. Vila-Liante, E. Arellano-Orden, T. Elías-Hernández, L. A. Ramón-Nuñez, L. Jara-Palomares, V. Martínez-Sales, L. Gao and R. Otero-Candelera (2015). "High correlation between 2 flow cytometry platforms in the microparticles analysis using a new calibrated beads strategy." Translational Research 166(6): 733-739.
58. Sánchez-Sarasúa, S., S. Moustafa, Á. García-Avilés, M. F. López-Climent, A. Gómez-Cadenas, F. E. Olucha-Bordonau and A. M. Sánchez-Pérez (2016). "The effect of abscisic acid chronic treatment on neuroinflammatory markers and memory in a rat model of high-fat diet induced neuroinflammation." Nutrition & metabolism 13(1): 1-11.
59. Sarmento-Cabral, A., V. Herrero-Aguayo, M. D. Gahete, J. P. Castaño and R. M. Luque (2017). "Obesity and metabolic dysfunction severely influence prostate cell function: role of insulin and IGF1." Journal of cellular and molecular medicine 21(9): 1893-1904.
60. Sirvent, S., I. Soria, C. Cirauqui, B. Cases, A. I. Manzano, C. M. Diez-Rivero, P. A. Reche, J. López-Relaño, E. Martínez-Naves and J. Jiménez-Barbero (2016). "Novel vaccines targeting dendritic cells by coupling allergoids to nonoxidized mannan enhance

allergen uptake and induce functional regulatory T cells through programmed death ligand 1." Journal of Allergy and Clinical Immunology 138(2): 558-567. e511.

61. Talamillo, A., L. Herboso, L. Pirone, C. Pérez, M. Gonzalez, J. Sánchez, U. Mayor, F. Lopitz-Otsoa, M. S. Rodriguez and J. D. Sutherland (2013). "Scavenger receptors mediate the role of SUMO and Ftz-f1 in Drosophila steroidogenesis." PLoS Genet 9(4): e1003473.
62. Torner, N., D. Carnicer-Pont, J. Castilla, J. Cayla, P. Godoy and A. Dominguez (2011). "Auditing the management of vaccine-preventable disease outbreaks: the need for a tool." PloS one 6(1): e15699.
63. Urdinguio, R. G., M. I. Torró, G. F. Bayón, J. Alvarez-Pitti, A. F. Fernández, P. Redon, M. F. Fraga and E. Lurbe (2016). "Longitudinal study of DNA methylation during the first 5 years of life." Journal of translational medicine 14(1): 1-12.
64. Vicent, L., J. Velásquez-Rodríguez, M. J. Valero-Masa, F. Díez-Delhoyo, H. González-Saldívar, V. Bruña, C. Devesa, M. Juárez, I. Sousa-Casasnovas and F. Fernández-Avilés (2017). "Predictors of high Killip class after ST segment elevation myocardial infarction in the era of primary reperfusion." International journal of cardiology 248: 46-50.
65. Zabala, A., M. Bustillo, I. Querejeta, M. Alonso, O. Mentxaka, A. González-Pinto, A. Ugarte, J. J. Meana, M. Gutiérrez and R. Segarra (2017). "A Pilot Study of the Usefulness of a Single Olanzapine Plasma Concentration as an Indicator of Early Drug Effect in a Small Sample of First-Episode Psychosis Patients." Journal of clinical psychopharmacology 37(5): 569.