http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

New Insights in the Design and Compilation of Digital Bilingual Lexicographical Products: The Case of the Diccionarios Valladolid-UVa

Pedro A. Fuertes-Olivera, *International Centre for Lexicography, University of Valladolid, Spain; and Department of Afrikaans and Dutch, University of Stellenbosch, South Africa (pedro@emp.uva.es)*

Sven Tarp, *International Centre for Lexicography, University of Valladolid, Spain; Department of Afrikaans and Dutch, University of Stellenbosch, South Africa; and Centre for Lexicography, University of Aarhus, Denmark (st@cc.au.dk)*
and
Peter Sepstrup, *Ordbogen A/S, Odense, Denmark (pse@ordbogen.com)*

Abstract: This contribution deals with a new digital English–Spanish–English lexicographical project that started as an assignment from the Danish high-tech company Ordbogen A/S which signed a contract with the University of Valladolid (Spain) for designing and compiling a digital lexicographical product that is economically and commercially feasible and can be used for various purposes in connection with its expansion into new markets and the launching of new tools and services which make use of lexicographical data. The article presents the philosophy underpinning the project, highlights some of the innovations introduced, e.g. the use of logfiles for compiling the initial lemma list and the order of compilation, and illustrates a compilation methodology which starts by assuming the relevance of new concepts, i.e. object and auxiliary languages instead of target and source languages. The contribution also defends the premise that the future of e-lexicography basically rests on a close cooperation between research centers and high-tech companies which assures the adequate use of disruptive technologies and innovations.

Keywords: DICTIONARY CONCEPT, EMPIRICAL RESOURCES, LOGFILES, NGRAM VIEWER, INTERNET AS A CORPUS, COMPILATION METHODS, LEXICOGRAPHICAL DATA, ONLINE DICTIONARIES, INTEGRATED DICTIONARIES, WRITING ASSISTANTS, L2-RECEPTION DICTIONARIES, L2-PRODUCTION DICTIONARIES, TRANSLATION DICTIONARIES

Opsomming: Nuwe insig in die ontwerp en samestelling van digitale tweetalige leksikografiese produkte: Die geval van die Diccionarios Valladolid-UVa. In hierdie bydrae word aandag geskenk aan 'n nuwe digitale Engels–Spaans–Engelse leksikogra-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

fiese projek wat begin is in opdrag van die Deense hoëtegnologiemaatskappy Ordbogen A/S. 'n Ooreenkoms is gesluit met die Universiteit van Vallodolid (Spanje) vir die ontwerp en vervaardiging van 'n digitale leksikografiese produk wat ekonomies en kommersieel uitvoerbaar is en wat gebruik kan word vir verskillende doeleindes wat verband hou met die uitbreiding daarvan na nuwe markte en die bekendstelling van nuwe hulpmiddels en dienste wat leksikografiese data benut. Die artikel bespreek die filosofie onderliggend aan die projek, belig sommige van die vernuwende elemente wat bekendgestel is, soos die gebruik van log-lêers vir die samestelling van die aanvanklike lemmalys en die volgorde van die samestelling. Die samestellingsmetodologie wat begin by die aanname dat vernuwende konsepte toepaslik is, word ook geïllustreer, d.w.s. primêre en sekondêre tale in plaas van doel- en brontale. In hierdie bydrae word die aaname dat die toekoms van e-leksikografie fundamenteel berus op die noue samewerking tussen navorsingssentrums en hoëtegnologiemaatskappye wat die voldoende gebruik van ontwrigtende tegnologieë en vernuwende elemente verseker, verdedig.

Sleutelwoorde: WOORDEBOEKKONSEP, EMPIRIESE HULPBRONNE, LOG-LÊERS, NGRAM VIEWER, DIE INTERNET AS 'N KORPUS, SAMESTELLINGSMETODES, LEKSIKOGRAFIESE DATA, AANLYN WOORDEBOEKE, GEÏNTEGREERDE WOORDEBOEKE, SKRYFHULPMIDDELS, L2-RESEPSIE-WOORDEBOEKE, L2-PRODUKSIE-WOORDEBOEKE, VERTALENDE WOORDEBOEKE

> "Q: Did you do consumer research on the iMac when you were developing it?
>
> A**:** No. We have a lot of customers, and we have a lot of research into our installed base. We also watch industry trends pretty carefully. But in the end, for something this complicated, it's really hard to design products by focus groups. A lot of times, people don't know what they want until you show it to them. That's why a lot of people at Apple get paid a lot of money, because they're supposed to be on top of these things."
>
> (Interview with Steve Jobs in *Business W*eek, Reinhardt (1998))

1. Introduction

In this contribution, we will first briefly discuss the history and philosophy behind a new digital English–Spanish–English lexicographical project which started in 2017 and is expected to see its first practical results launched from 2020 onwards. We will then go into details about the experience to date in the design and compilation of the product which, in various aspects, is innovative and based on cutting-edge technology with the use of completely new lexicographical methods guided by the Function Theory of Lexicography; cf. Fuertes-Olivera and Tarp (2014).

The project started as an assignment from the Danish high-tech company Ordbogen A/S, an international provider of online dictionary portals (ordbogen.com, lemma.com) as well as language services 24/7. Due to its technological innovation and unique business model, both of which have received

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

several national and international prizes, this company has since 2000 completely surpassed the traditional publishing houses and is now the dominant provider of online dictionaries in Denmark with a clear intention to increase its market share also in the neighboring countries. It is therefore an interesting partner for any lexicographer with a novel idea to be implemented or a dictionary to be distributed on a commercial basis.

As to the project discussed in this contribution, Ordbogen A/S wanted a digital lexicographical product that is economically and commercially viable and can be used for various purposes which are in line with its expansion into new markets and the launching of new tools and services which make use of lexicographical data. The Danish company therefore made contact with the International Centre for Lexicography at the University of Valladolid (Spain), with which it was already collaborating. The collaboration was on two other major online projects, the English–Spanish Accounting Dictionaries (available since 2012) and a set of monolingual Spanish dictionaries (under construction), both of which are to be commercialized under the brand name *Diccionarios Valladolid-UVa*; see Fuertes-Olivera (forthcoming).

The contract signed between the two partners stipulates, among other things, that the Danish company provides the technological support for the project, including the Dictionary Writing System (DWS) with lexicographical database, interfaces, search engines and grammar, as well as part of the empirical basis. The Spanish counterpart is in charge of the practical production of the required lexicographical data by means of highly specialized human resources, and project management. In addition, Ordbogen A/S finances part of the production costs at the International Centre for Lexicography with the Centre and the University of Valladolid providing the remaining funds. The project develops on a contractual basis as an international cooperation between two independent partners, each with their specific know-how and experience. To our knowledge, this represents a rather atypical lexicographical project model as projects of a similar scope in most cases are carried out either directly in the publishing houses or by independent entities and lexicographers who subsequently offer their products to the former. So far, the experience has been highly positive. For instance, Ordbogen A/S has agreed to transfer 50,000 euros a year to the International Centre for Lexicography for paying contracted and freelance lexicographers working on this project. These lexicographers are expected to compile around 25,000 senses a year, including definitions, collocations, synonyms, antonyms, examples and other data. This means that this bilingual project is expected to use approximately 2 euros per sense.

2. Lexicography, technology and current trends

During its more than four thousand years of existence as a cultural practice, lexicography has always depended strongly on the available technology in order to compile and present its products which, until now, have mainly taken

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

the form of dictionaries, although history has also known other forms of lexicographic endeavour. Hanks (2013: 507), for instance, reports how, at the dawn of European lexicography (500 B.C.), "it was customary for Greek scribes to insert glosses into manuscript copies of the works of Homer and other earlier writers" in order to explain "obsolete and unusual words".

These early *context-adapted* lexicographical glosses, which later developed into separate glossaries, allow for two important conclusions which we think are undervalued in the scholarly literature: 1) that lexicographers, as a matter of fact, do not compile dictionaries but lexicographical data which subsequently can be used for different purposes, among them, and notably, to edit dictionaries; and 2) that the standardized dictionary which was totally dominant in the printed environment is not the only type of lexicographical product known to history. Both conclusions are highly relevant for the correct interpretation of the current tendencies in lexicography where new disruptive technologies are turning the discipline upside down.

Prior to the advent of computer and information technologies, the introduction of the printing press more than 500 years ago had, in many respects, a similar impact on the discipline. A lot has been written about this phenomenon and some of its consequences (see, for instance, Hanks 2010), whereas other consequences have been less adequately dealt with, although they may not be less important in the long run.

In conclusion, it can be established that the introduction of printing technology implied big changes in:

— the production and presentation of the lexicographical product;
— the empirical basis with the increased use of index cards based on written texts;
— the design of dictionary articles with the incorporation of new data categories;
— the distribution and use of dictionaries;
— the number of users;
— the topics treated in dictionaries; and
— the research areas of scholarly interest.

To this can be added the growing social prestige of lexicographers, some of whom became nationally and internationally famous personalities, as well as the fact that lexicography turned into an increasingly successful business. Over a few hundred years, printing technology led to an almost total revolution of the discipline.

A similar thing is happening today where the technological innovations affect lexicography in its four main dimensions, i.e. the production, presentation, usage and financing of the lexicographical product. Fuertes-Olivera (2016) refers to the current situation as a "Cambrian explosion" where new forms constantly appear and disappear. This indicates that the adaptation to the new technological environment is a complex process that is far from one-dimen-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

sional. Of special concern is the fact that the new technologies, especially the use of the Internet to make dictionaries available to their users, has undermined the existing business model and thrown lexicography into a sort of identity crisis where many publishing houses have reduced or even closed down their lexicographical sections due to dramatically reduced sales. Consequently, the continuous production of high-quality products is under attack. A new business model is therefore necessary but this is, obviously, the publishers' task — although nothing prevents lexicographers from contributing new ideas.

It is important to understand that the roots of the current crisis for lexicography are not only objective (disruptive technologies and an obsolete business model), but also subjective (ingrained habits and a frequently conservative approach to the new challenges). In this regard, lexicographers also have a big responsibility to the future of the discipline. They are above all challenged with the task of engaging in interdisciplinary collaboration with programmers and designers in order to guarantee still higher productivity without compromising quality and exploring new ways of presentation of the lexicographical product as the old static dictionary article is becoming increasingly obsolete. This presupposes a good dose of technological sensibility and understanding of the lexicographical potential of the continuous innovations, development of new compilation methods, and visionary thinking that offers new solutions to both new and old problems. In this perspective, Rundell and Kilgarriff (2011) have treated some of the technological and methodological advances in terms of the automation of the compilation process but the very title of their contribution leaves, understandably, an important question to be answered: "Where will it all end?"

In the following, we will take Rundell and Kilgarriff's (2011) reflections a little further and look into new methods of lexicographical data compilation. However, we are firmly convinced that this is not possible without knowing, or at least having a qualified idea of, the destination of these data and how they will eventually be presented to the users. A careful observation of the current trends related to this aspect of lexicography unveils four big transformations that are taking place simultaneously:

1. The first big transformation is from the *printed dictionary* to the *digital dictionary*. This process is still ongoing and characterized by a large number of dictionaries, especially online ones, that are either digitalized editions of already printed dictionaries or designed from scratch without taking into proper consideration the new options provided by the digital media.

2. The second transformation is from the traditional *stand-alone dictionary*, either printed or online, to the *integrated dictionary*, i.e. a dictionary integrated into other information tools such as e-readers, writing assistants or learning tools.

3. The third transformation is from the *standardized dictionary*, which is a typical result of the printed book format, to a more *personalized dictionary*

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

that adapts to the user's specific needs in each situation.

4. Finally, it is also possible to observe an inevitable move away from the *dictionary as such* to *lexicographical data for different uses*. Today, many publishing houses are increasingly receiving their revenue from selling lexicographical data. Many integrated information tools do not present dictionaries as such to their users but only a selection of data types taken from a lexicographical database.

The growing tendency to work with separate lexicographical data is also the reason why this contribution mostly refers to *lexicographical products* instead of dictionaries. In some cases, the data may even be taken from different sources. An example of this can be seen in Figure 1, which is a screenshot from a Danish–English writing assistant where the English equivalent *donation* and its Danish definition have been taken from two different digital dictionaries in order to get the best result for the user. This has become necessary because neither of the two dictionaries was originally planned and designed to be used as a support for a writing assistant.

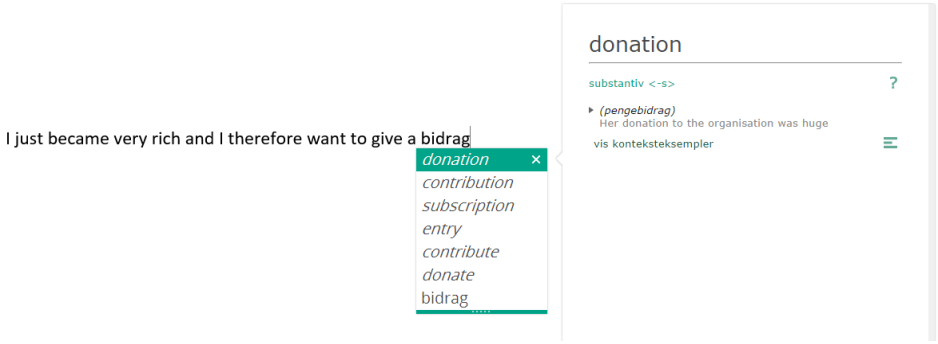

Figure 1: Screenshot from a Danish–English Write Assistant (Fisker 2018)

These four transformations are interwoven and herald a near future of integrated information tools that are based on digital platforms and provide *personalized service* by making use of lexicographical data. Personalized service is a general consumer demand in modern society and has therefore been a dream of many lexicographers in recent years; see, for instance, Rundell (2010).

In a lexicographical perspective, personalized service can be defined as the provision of the exact amount and types of data required to meet the user's needs in each concrete consultation, neither more nor less (Fuertes-Olivera and Tarp 2014: 64). This requires that the lexicographical data are of a high quality, that there are enough data, and that data and information overload is avoided as requested by Gouws and Tarp (2017).

Until a few years ago, it was conceived by many lexicographers, among

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

them Tarp (2011), that a personalized solution meeting these requirements would be something like "a set of components which customers can mix and match according to their needs" (Rundell 2007: 50). However, subsequent technological development has shown that such a solution, although innovative and useful in many aspects, is not the final word in this respect as it entails clear structural limitations, especially in terms of the resulting information costs, i.e. the time required to find and retrieve the needed information; see Nielsen (2008). It is now clear that a completely personalized service is only possible in an integrated information tool which, like a GPS, is designed to "observe" its users' behavior and prescribes the exact amount of data that is likely to meet their needs in each concrete case.

3. Presentation of the project

The combination of data from two different sources, which was shown in Figure 1, is a typical example of lexicographical databases that were prepared without knowing the exact use of the included data. In this case, the databases were designed several years before the work with the writing assistant started, and the problem was inevitable, at least to a certain extent. It nevertheless shows that meticulous work is required when starting a lexicographical project as small inadvertent "mistakes" could have big consequences at a later stage.

In the project we are discussing, the assignment from Ordbogen A/S was clear. The company wanted a bilingual lexicographical database that could feed two new products, namely a "traditional" online English–Spanish–English dictionary (to be included into the portal *Diccionarios Valladolid-Uva*) as well as a Spanish–English Write Assistant like the one described by Tarp et al. (2017). Both products are intended for native Spanish-speaking users.

After signing the contract, the lexicographers at the International Centre for Lexicography in Valladolid were tasked with 1) establishing the respective lexicographical functions, 2) framing a project concept including the required data categories, 3) preparing a compilation methodology that guarantees productivity and quality, and 4) engaging with programmers and designers from Ordbogen A/S in order to design a Dictionary Writing System suited for this particular project; cf. Fuertes-Olivera (forthcoming).

As to the functions, it was evident that the Write Assistant in the first instance had to assist Spanish users when writing texts in English and, secondarily, when translating from Spanish into English. These two functions are also applicable to the dictionary section of the project to which should be added two other functions that were exclusively relevant to the dictionary, i.e. assistance with reception of English texts and English–Spanish translation.

Based on these prioritized functions, a list of data categories to be included in the lexicographical database was drawn up. These categories included formal grammar, definitions, synonyms, antonyms, collocations, example sentences, etc. Apart from that, and due to the design and functionality of both the

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

Write Assistant and the dictionary, the data categories were divided into their smallest relevant parts so that they could be presented separately to the users who are expected to work on devices with varying screen sizes (laptops, tablets and smartphones).

As to the presentation of the product, it is predicted that the dictionary will be made available on the Internet with different function buttons which allow the users to get more specialized and individualized service such as:

— English definitions
— English grammar
— English synonyms and antonyms
— English collocations and examples
— Spanish–English translation
— Spanish–English translation of collocations and examples
— Spanish–English translation of fixed expressions
— English–Spanish translation
— English–Spanish translation of collocations and examples
— English–Spanish translation of fixed expressions
— Etc.

As to the Write Assistant, its design and functionality is still being improved. However, at this point it is clear that the tool demands English equivalents to Spanish words, Spanish definitions of English words (to be used as meaning discrimination), English inflectional forms, English synonyms and antonyms, English collocations and example sentences, etc. This suggests that the data categories already envisaged for the dictionary project are sufficient in order to meet the requirement of the Write Assistant.

As mentioned, both the dictionary and the writing assistant are intended for native Spanish-speaking users. This means that English in both cases is the language where the users need to have special assistance, whereas Spanish is used both to access and explain English, and as a lexicographical metalanguage. We therefore call these two languages *object language* (English) and *auxiliary language* (Spanish), respectively. These two terms are used with a different meaning than the one defined by Hartmann and James 1998 in their *Dictionary of Lexicography*. In this respect, they break with the terminology which is traditionally used to describe bilingual dictionaries (source language, target language, both of them treated as object languages) and which was coined in a period when practical solutions to users' needs were influenced and also limited by the existing technology, especially due to the restraints of the printed book format. We do not find this old terminology to be the most adequate and helpful if lexicography is expected to make the most out of modern computer and information technology as it may constitute a mental barrier that stands in the way of developing new solutions.

The new terminology makes us focus on the object language, i.e. English. It is English that has to be described and explained to the Spanish users. It is in

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

English where they need instructions on how to write and produce texts. Spanish is "only" used to access the English words and expressions as well as to explain these and give indications on how to use them in context. This means that Spanish is not going to be treated at the same level as English.

This approach has direct consequences for the methodology used in the project. Whereas traditional mono-directional, biscopal dictionary projects usually take their point of departure in the users' native language, the Valladolid project does the opposite. It starts with a selection and description of English lemmata including separation in senses, definitions, Spanish equivalents, grammar, etc. An automatic and simultaneous inversion is then made where the Spanish equivalents to one or more English lemmata become new lemmata whereas the English lemmata become equivalents with the brief Spanish definitions used as meaning discrimination. This inversion is, of course, revised by the lexicographers who also rely on an independent list of Spanish lemma candidates as will be explained in paragraphs 4 and 5.

The described compilation methodology is then in close collaboration with programmers and designers from Ordbogen A/S, incorporated into the Dictionary Writing System which, so far, has proved to be very user-friendly, efficient and economical in terms of productivity and quality.

4. Empirical basis

In the scholarly literature, there is a long and rich reflection on the most adequate empirical basis of the different types of dictionary. As it was briefly mentioned in the historical overview in paragraph 2, as is the case with the compilation and presentation of the lexicographical product, its empirical sources also change over time as the result of the continuous technological development. Since the 1960s, and especially since the disruptive publication of the *Collins Cobuild English Language Dictionary* in 1987, there has been a strong reliance on still bigger corpora as the main empirical source of dictionaries (see Sinclair 1987, Bergenholtz 1996, Kilgarriff 1997, Atkins and Rundell 2008, and Hanks 2012, among many others). The positive results of this development are indisputable and excellent dictionaries have been produced with this point of departure. However, the generalized use of corpora also gives rise to new questions and challenges, especially in the light of new technological innovations such as digital dictionaries, the Internet and logfiles. This is the case with the selection of lemmata to dictionaries with a limited lemma stock. Some lexicographers, like Kilgarriff (2013), advocate, at least until recently, the use of corpora in these cases:

> Building a headword list is the most obvious way to use a corpus for making a dictionary. *Ceteris paribus,* if a dictionary is to have N words in it, they should be the N words from the top of the corpus frequency list. (Kilgarriff 2013: 79)

However, there are two questions which in our opinion have not been paid suf-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

ficient attention, namely 1) whether users actually consult the lemmata included in dictionaries, and 2) the relationship between corpus frequency and lexicographical frequency, i.e. the frequency with which the users consult the words in a dictionary. As to the first question, Bergenholtz and Norddahl (2012) have reported that the study of logfiles shows that a considerable number of words have never been consulted in an online Danish dictionary after more than 20 million lookups. The dictionary in question is a big general one with more than hundred thousand lemmata and the conclusion may therefore not be representative for dictionaries with a more reduced lemma stock as the ones to which Kilgarriff (2013) refers. However, research into logfiles by other scholars confirms another of Bergenholtz and Norddahl's (2012) conclusions, namely that there is a certain, and therefore lexicographical relevant, discrepancy between the most frequent words in a corpus and the words most frequently looked up in dictionaries; see De Schryver et al. (2006) and Trap-Jensen et al. (2014).

This last conclusion implies that it would be better to start a lexicographical project with a reduced lemma stock with lemmata selected from logfiles instead of a corpus, and then use the method recommended by Bergenholtz and Johnsen (2005) and De Schryver (2013), among others, to supplement the lemma list with additional lemmata that appear in the logfiles once the dictionary has been published online. This is, at least, the method used in the project discussed in this contribution, which uses four main empirical bases: logfiles; Ngram Viewer; the Internet; and existing dictionaries. These are used in the above order and nobody working in the project can change the order, as this could clearly jeopardize the whole project, as we will show in the following paragraphs. This is critical for the project as we believe that someone initially consulting a dictionary will be clearly influenced in their lexicographical work by the data found in the consulted dictionary.

4.1 Logfiles

As already mentioned, our bilingual project started in 2017 with an initial lemma list of around 20,000 English words and 16,000 Spanish words. These were compiled at Ordbogen A/S headquarters in Odense (Denmark) by using big data analytics for around two months.

The process comprises several stages and is based on an analysis of around one million daily searches in several of the company's dictionaries, e.g. an English–Spanish/Spanish–English dictionary, an English–German/German–English dictionary, an English monolingual dictionary, a Spanish monolingual dictionary, and so on. Around 80% of the searches can be matched, i.e. the same search is identified in the logfiles of different dictionaries and can therefore be interpreted with the aim of identifying the most popular dictionary articles in the dictionaries under scrutiny. After two months of work with the logfiles of the searches, which amount to more than 60 million, IT people at Ordbogen A/S were able to produce the above-mentioned lists of 20,000 Eng-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

lish words and 16,000 Spanish words. They comprise the words most searched for in the period under analysis and were used by the editor of the project for compiling the initial lemma lists of the bilingual project. Below, we copy some of the searched words, most of which are currently lemmata in this project:

— English words starting with "ang-": *angel, angelic, angelica, anger, angered, angina, angiogenesis, angiogram, angioplasty, angle, angled, angler, Anglican, anglicize, angling, Anglo, Anglo-American, Anglo-Danish, Anglo-Saxon, Anglophile, Anglophobe, Anglophone, Angola, angrily, angry, angst, anguish, anguished, angular, angular momentum, angularity,* and *angulation.*

— English words starting with "bed-": *bed, bed linen, bed-sitting room, bedazzle, bedbug, bedchamber, bedclothes, bedcover, bedding, bedevil, bedew, bedfellow, bedlam, Bedouin, bedpan, bedplate, bedpost, bedraggled, bedridden, bedrock, bedroom, bedside, bedsit, bedsore, bedspread, bedspring, bedstead, bedstead canopy, bedtime* and *bedtime story.*

— English words starting with "defe-": *defeasible, defeat, defeated, defeatist, defecate, defecation, defect, defection, defective, defector, defence, defenceless, defend, defendant, defender, defenestrate, defenestration, defensible, defensive, defensively, defensiveness, defer, deference, deferent, deferential, deferment, deferral,* and *deferred income.*

— English words starting with "equ-": *equable, equal, equalization, equalize, equalizer, equality, equally, equanimity, equanimous, equate, equation, equator, equatorial, equestrian, equidistant, equdistribution, equifinality, equilateral, equilibrate, equilibration, equilibrist, equilibrium, equine, equinox, equip, equipment, equipoise, equipotential, equipped, equitable, equitably, equity, equity capital, equity ratio, equity warrant, equivalence, equivalency, equivalent, equivalently, equivocal, equivocality, equivocally, equivocate,* and *equivocation.*

A comparison of the above words with the lemma list of the *Oxford English–Spanish Dictionary* indicates three main findings. Firstly, the degree of matching between them is high: 74% of the most searched words are also found in the lemma list of the *Oxford English–Spanish Dictionary*. Secondly, there are 36 searched words (26%) that are not included in the Oxford list and these are basically either formal technical words, e.g. medicine words, or multiword lemmata, i.e. extended units of meaning (Rundell 2018). Thirdly, this second list, composed of the searched words that are not lemmatized in the Oxford dictionary, offers some clues about users' interest for some semantic fields, — these are: medicine, sports, law, sex, and economics — and for multiword lemmata. In addition, they also indicate the adequacy of logfiles for offering more than possible lemmata: they can also help lexicographers to disambiguate meanings and offer interesting data for crafting additional data types, typically sentence examples and collocations, i.e. chunks of words that offer clues on the meaning and use of the lemma and/or equivalent.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

4.2 Ngram Viewer

As already mentioned, in this bilingual project we have used three other empirical bases apart from the logfiles: Ngram Viewer; the Internet; and existing dictionaries. Ngram Viewer is being used for four main lexicographical purposes. Firstly, it is used for augmenting the initial lemma list with "extended units of meaning", i.e. strings of recurrent words that adhere to Sinclair's idiom principle which assumes that language users regularly resort to "an inventory of semi-preconstructed phrases that constitute single choices" (Sinclair 1991: 110). In our bilingual project these are lemmatized when they refer to bearers of meaning, e.g. they refer to material things, feelings and emotions, human beings, etc. both in their literal and figurative meanings. For instance, we have searched in Ngram Viewer (English) the following strings: *air* *_ADJ, *air *_NOUN, *_NOUN air, *_VERB air, *_ADJ air, air *_PREP, air *_CONJ* and *air *_VERB.*

An analysis of the hits as well as the results obtained during the process of compilation, e.g. by means of Google searches and look ups in existing dictionaries (see below), have resulted in around 100 multiword lemmata with "air": *air current*, *air dry*, *air offensive*, *air pollution*, *air force*, *air conditioning*, *air temperature*, *air pressure*, *air flow*, *air transport*, *air space*, *air pump*, *compressed air rifle*, *air rifle scope*, *confined air*, *air letter*, *air ambulance*, *air assault*, *air attack*, *air ball*, *air brake*, *air bubble*, *air cargo*, *air cleaner*, *air conditioner*, *air conditioning unit*, *air dam*, *air filter*, *air freight*, *air freshener*, *air gap*, *air gauge*, *air guitar*, *air hammer*, *air hockey*, *air hockey table*, *air horn*, *air hostess*, *air hunger*, *air intake*, *air kiss*, *air leak*, *air lock*, *air mail*, *air map*, *air marshal*, *air mass*, *air mattress*, *air mile*, *air out*, *air piracy*, *air pocket*, *air power*, *air pressure*, *air rage*, *air raid*, *air scoop*, *air spring*, *air strike*, *air taxi*, *air traffic*, *air traffic control*, *air travel*, *air valve*, *air vent*, *air waybill*, *air-condition*, *air-conditioned*, *air-cooled*, *air-dried*, *air-filled*, *air-raid shelter*, *air-sea rescue*, *air-to-air*, *air-to-ground*, *air-to-surface*, *air traffic controller*, *airbag*, *airbase*, *airbed*, *airtight*, *airway*, *air lane*, *on-air*, *airing*, *airman*, *airwoman*, *air gun*, *hot air*, *airfare and airdrop*.

It is interesting to highlight that around 40% of these multiword lemmata are not lemmatized in the *Oxford English–Spanish Dictionary*. This finding indicates that the use of several empirical bases may be needed for compiling online dictionaries, especially because of the disappearance of the space constraints associated with a printed dictionary. For instance, neither *air ball* nor *airball* is lemmatized in the *Oxford English–Spanish Dictionary*, although it is frequently used in Spanish television during the broadcasting of a basketball match where an *airball* is "an unblocked shot which misses the basket, the rim, and the backboard entirely" (*Wikipedia*).

Secondly, Ngram Viewer is used for deciding which word variety is used as lemma and which other varieties are included but not lemmatized. For example, the English varieties "color" and "colour" refer to the same reality and are totally synonymous for Spanish native speakers, who are the main users of these

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

products. We normally lemmatize the most frequent variety and include the other varieties in several data fields, e.g. as synonyms with their corresponding tag (e.g. UK English or US English) or as not recommended (Figure 2). This decision does not hinder users' searches in a reception situation: the dictionary entry for "color" and "colour" is the same and will be recovered searching "color" or "colour".

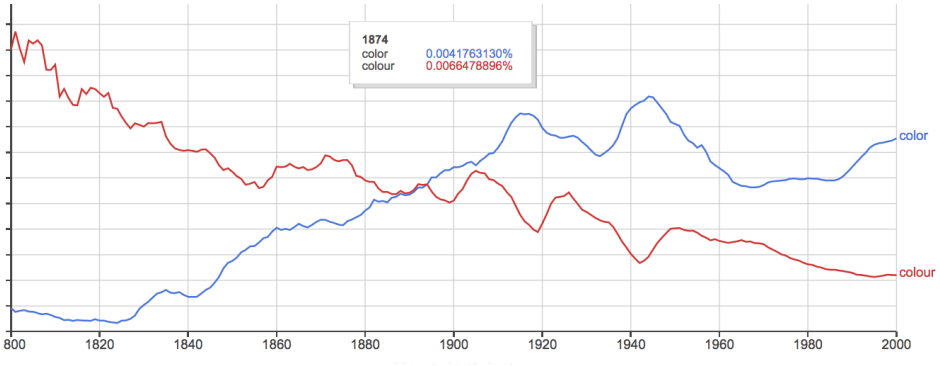

Figure 2: Comparison of **color** and **colour** with Ngram Viewer

Thirdly, we also use Ngram Viewer for checking the inflections and grammar forms of all lemmata, especially those lemmata that can be problematic for users, e.g. countable and uncountable English nouns, masculine and/or feminine Spanish nouns, and so on. For instance **air power** is described as "mass noun" in the *Oxford English Dictionary*. However, Ngram Viewer shows that "air powers" is used in English, especially during the Second World war (Figure 3):

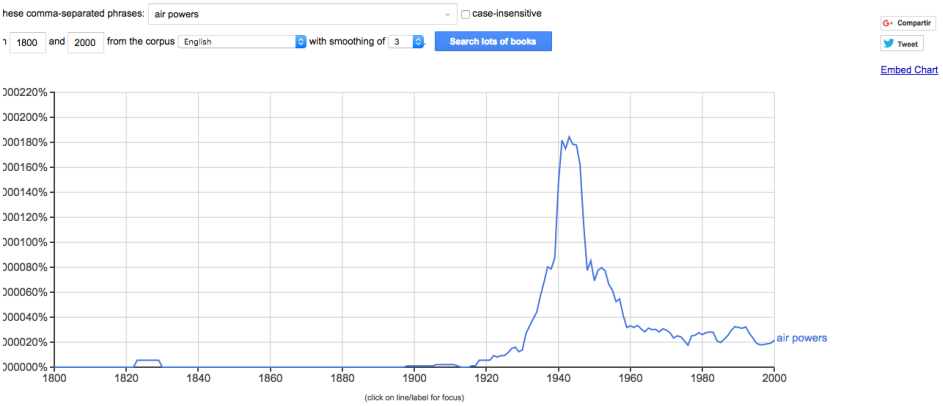

Figure 3: Uses of **air powers** with Ngram Viewer

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

Hence, in our project, **air power** is lemmatized as uncountable and countable, each with its own grammar, definition, examples, synonyms, and so on (examples 1 and 2):

air power
flexions
air power, air powers
Definition
unidad del ejército de un país encargada de todo lo relacionado con el ejército del aire
Equivalent
fuerza aérea
Example
China had "become one of the major air powers of the world"
China se ha convertido en una de las principales fuerzas aéreas del mundo

Example (1): Extract for **air power** (countable)

air power
flexions
Sin flexion (uncountable)
Definition
1. fortaleza o capacidad del ejército del aire de un nación para defender sus territorios o atacar otros
Equivalent
poder aéreo
Example
Military air power was used to protect relief efforts.
El poder aéreo militar se usó para proteger labores humanitarias.

Definition
2. energía producida por la acción del viento sobre un molino o aerogenerador
Equivalent
energía eólica
Example
In 2008, the U.S. became the world's leading provider of air power.
En 2008, EE.UU. se convirtió en el proveedor líder de energía eólica en el mundo.

Example (2): Extract for **air power** (uncountable)

Finally, we are also using Ngram Viewer for identifying frequent combinations of particular words that have not been lemmatized as multiword lemmata but are included in example sentences or other parts of the dictionary articles, especially when they offer something relevant such as a translation pattern as the different types of "air" recorded in Figure 4. All of them have similar Spanish translations and are therefore adequate for machine learning and neural network software: "aire de la noche", "aire de la mañana", "aire de la habi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

tación", "aire de la montaña", "aire de la tarde", "aire del mar", "aire salado", aire del verano" or "aire veraniego", "aire del desierto", and "aire del país":

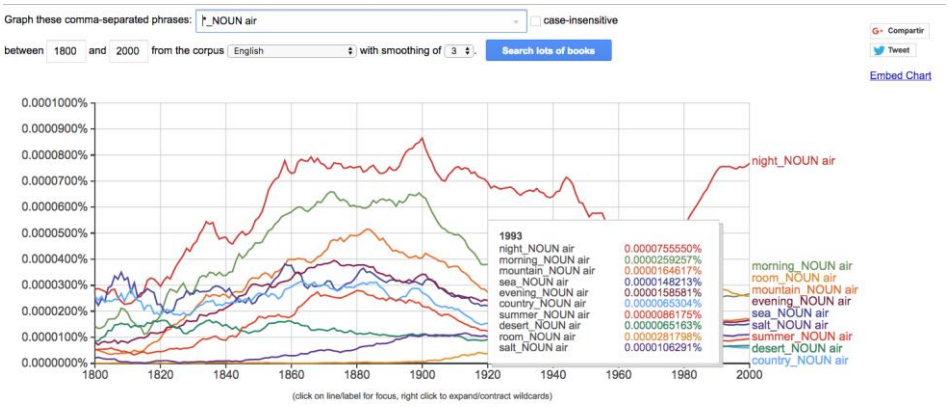

Figure 4: Searching *_NOUN air with Ngram Viewer

4.3 Internet

The Internet is also one of our main empirical bases. As shown in Tarp and Fuertes-Olivera (2016), we use it for crafting definitions, selecting different types of sentence examples, synonyms and antonyms, and so on. In Tarp and Fuertes-Olivera, we have shown that the analysis of Google minitexts, i.e. three-line texts that appear as a result of a Google search, has resulted in dictionary articles that, on average, describe 30% more different senses than existing dictionaries. Examples (1) and (2) above corroborate this reflection. **Air power** in this bilingual project has three senses (In all the dictionaries consulted, it has only one sense. For instance, the *Oxford English Dictionary* describes **air power** as "The ability to defend or attack by means of aircraft."). One of the senses recorded in this bilingual project refers to "wind energy", and this sense is obtained by analyzing texts such as the following, all of them extracted from the Internet, and recorded in the dictionary entry either as collocations or as sentence examples (example 3):

Collocations

— that air power is considered one of the purest energy sources
— the use of air power and solar installations
— the air power, biomass and waste treatment sectors
— air power, solar energy and other renewable energies

Examples

— In 2008, the U.S. became the world's leading provider of air power.

Example (3): Collocations and examples extracted from the Internet for crafting one of the senses of **air power**

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

The Internet is also used for four more key lexicographical tasks. First, we *always* consult *Wikipedia* for explaining *terms*, i.e. words describing concepts in specialized fields (Humbley 2018). For instance, **angioplasty** (one of the words from the logfiles), is explained in terms of the data reported in *Wikipedia*, especially with regard to following: (a) a definition for our Spanish users; (b) several synonyms, e.g. **balloon angioplasty**; (c) several types of **angioplasty,** all of which are also lemmatized, e.g. **coronary angioplasty; peripheral angioplasty; carotid angioplasty;** (d) its Spanish equivalent; (e) a usage note indicating that this medical procedure did not exist until 1964 and is therefore not found in texts before such year; (g) collocations and sentence example(s), e.g. "Angioplasty is typically used to treat atherosclerosis." (h) link(s) to the *Wikipedia* entry, images, and so on.

Secondly, we also use *Wikipedia* and other available texts, e.g. maps, lists of cities, rivers, oceans, seas and mountains for completing and describing our lemmata. For instance, the *Wikipedia* entry for **Amazon River** offers reliable data on its left and right tributaries (they are lemmatized when their length is more than 1,000 km) as well as other data for crafting its definition. In addition, several blogs offer interesting data about its flora and fauna, ecosystem, characteristics of the rainforest, etc. All these texts are analyzed and some of their data are included.

Thirdly, we also use the Internet for searching for texts that can be of use for our users such as free online pronouncing dictionaries (e.g. https://www.howjsay.com/), images, and so on. We include a link to unprotected texts, especially to texts produced by individuals or companies whose business model is based on the number of clicks, i.e. their revenues come from clicks, no matter where these are done.

Finally, we also use the Internet for finding equivalents. For such a task, we query Google with the lemma, some words related with its meaning(s), and the expression "in Spanish" or "in English" if we are searching for the Spanish or English equivalent. For instance, for finding the English equivalent of Spanish "cobro revertido", we googled "cobro revertido in English", and obtained the English equivalent "reverse charge". Analyzing it was very fruitful: we discovered that **reverse charge** is a synonym of US English **collect call** as well as a term related with the accrual of VAT (lemma **VAT reverse charge**), the charging of batteries, and a trick in pen spinning. In existing dictionaries **reverse charge** is only explained as UK English for "a telephone call paid for by the recipient" (*Oxford English Dictionary*).

4.4 Existing dictionaries

Existing dictionaries are also used as empirical bases. They are consulted once the rest of the empirical bases have been used for compiling the dictionary entries. This consultation has three main purposes.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

Firstly, we check for any possible missing sense. In such a situation the lexicographer in charge of a particular dictionary entry must analyze whether the missing sense is still in use, e.g. by googling the lemma with some key parts of the definition or equivalent found.

Secondly, we check for possible grammar discrepancies (e.g. countable or uncountable nouns), lexicographical notes (e.g. a usage note about a particular lemma), formal and informal tags, and so on. If something new is found we double-check it before incorporating it in the dictionary entry.

Finally, we consult existing dictionaries for finding information about geographical varieties, something that is difficult to obtain from the rest of the empirical bases.

5. Phases and steps in the compilation process

As indicated in the previous section, there are three main phases, each with their own sub-phases or steps in the compilation process of this bilingual project.

5.1 First phase

The first phase comprises the work in the English–Spanish section of the project. It starts with the editor of the project analyzing the logfiles submitted by Ordbogen A/S, i.e. the list of around 20,000 English words, with two main aims, deciding which of the most searched words will be lemmatized in the project and establishing the order of compilation of each lemma, as it is expected that the project will go public before it has been completed. This order of compilation is important as we have found out that users do not search randomly but tend to search for specific words. For instance, an analysis of the around 2,300 English words starting with "a-" submitted by Ordbogen A/S to the editor shows that around 25% of them refer to five topics: medicine; sports; law; sex and economics.

The second step comprises the working in the 'Lemma section' of the Dictionary Writing System (Figure 5). This is the editing tool with up to 20 slots, i.e. fields for including lexicographical data and administrative data. The editor of the project enters the basic grammar data of the lemma and decides who will finish the rest of the dictionary entry ('Assigned user' in Figure 5), its status (Review, in progress, finished, start, etc.) and its log history, e.g. who has completed the entry, who has reviewed it, etc. This information is important for the editor, who can, for example, decide to assign the Spanish–English entry to the same lexicographer who has completed the English–Spanish one.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

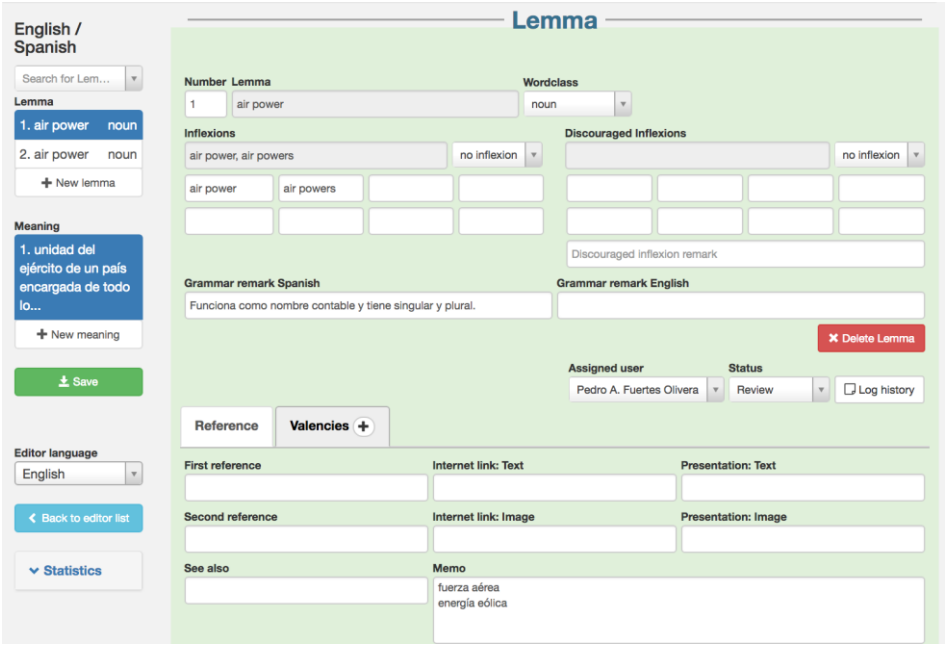

Figure 5: The 'Lemma section' of the Dictionary Writing System

By basic grammar data we mean the information that applies to the lemma in all situations, varieties, registers, etc. For most lemmas, this information comprises the following:

— Number: it differentiates between homonyms, for instance **air power** (countable) and **air power** (uncountable).

— Lemma: it records the dictionary form or canonical form of the lemma. For instance, in the logfiles we have found that users typically search for "clothes" instead of "cloth". In the project, however, we have lemmatized three examples of "cloth": (a) cloth as countable noun; (b) cloth as uncountable noun; and (c) cloth as singular noun in the collective noun "the cloth".

— Word class: it indicates the part of speech of the lemma.

— Inflexions: depending on the word class of the lemma, it stores singular and plural noun forms, comparative and superlative adjectival forms, some regular and irregular verbal forms, and so on.

— Discouraged inflexions: it also stores the above types of inflexions but with an indication that they are not recommended for some reason, e.g. **airball** is less used than **air ball**.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

— Grammar remark in Spanish: a grammar comment in Spanish, e.g. **air power** is a countable noun and has singular and plural forms (Lemma 1) and **air power** has no inflexions as it is an uncountable noun (Lemma 2).
— Reference: for internal reference, i.e. cross-references, or for external reference, e.g. a link to a free pronouncing dictionary.
— Valency: it includes syntactic information of the lemma, e.g. "someone plucks something from the air" in the lemma **pluck from the air**.

The third step in phase one comprises work in the 'Meaning section' of the Dictionary Writing System (Figure 6). The assigned lexicographer works in this section, which consists of up to 27 slots with the aim of offering five types of information: (a) meanings; (b) tags, for indicating the style and type of English, if necessary; (c) remarks, e.g. with this meaning it is only used in negative; (d) references, both internal cross-references and external references with links to homepages, e.g. a FLICKR image; and (e) synonyms and antonyms. It is important to highlight that all the information contained in this part is linked to the Arabic number on the left side. This serves for "bundling" all the data of the "meaning part" to each meaning, i.e. associating each meaning to its synonyms, antonyms, production notes, external and internal references, tags and so on. For instance, the meaning "wind energy" of **air power** (example 2 above) is always associated with the data types describing this meaning and use.

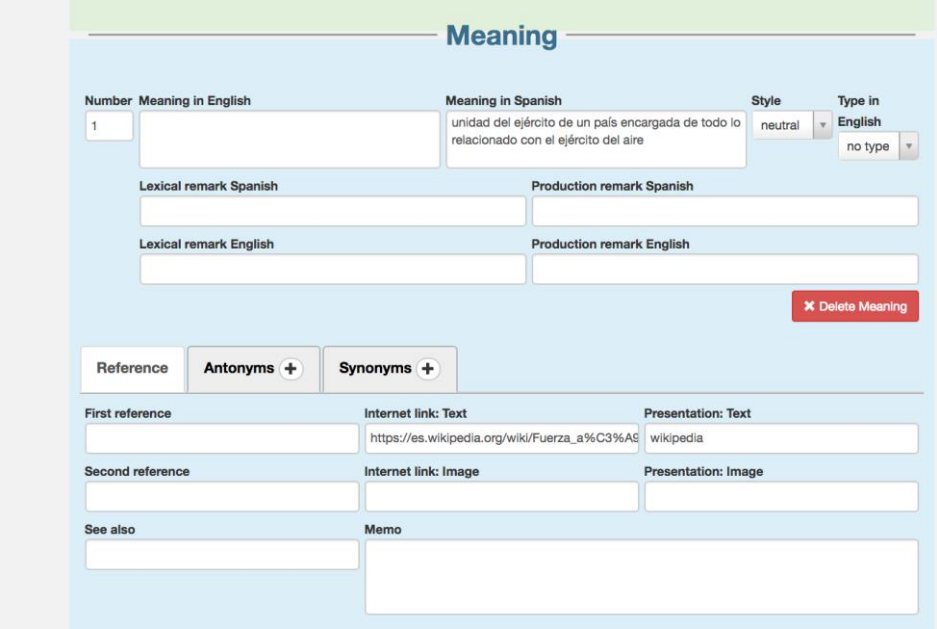

Figure 6: 'The Meaning section' of the Dictionary Writing System

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

The fourth step in phase one is working in the 'Translation section' of the Dictionary Writing System (Figure 7). It includes up to 20 slots, all of them concerned with the Spanish equivalent and the meaning and function of lemma and equivalent in context. In this section we also have an administrative button "Create lemma", which will be used in Phase 3 of the compilation process. Regarding the equivalent, lexicographers include the Spanish equivalent, its word class, grammar and contrastive remarks as well as syntactic information, e.g. that an English verb is only used with "something" and not with "someone". On average, we only include one equivalent per meaning, although there are exceptions. For instance, the English lemma **teacher** can refer to a male or female teacher. As the Spanish gender system is different we include the Spanish equivalents **profesor** and **profesora** (male and female teacher). This distinction is important for several reasons and has important consequences in our project. We will not comment on it further for reasons of space. Finally, the buttons "collocations", "examples" and "formation" in Figure 6 record data for contextualizing the meaning of the lemma in different translation situations, in which we can or cannot have the Spanish equivalent. For instance, one of the meanings of **leave up in the air** refers to an unsettled issue or plan. Its Spanish equivalent is "dejar en el aire". One of the collocations in this meaning is "that the whole matter was left up in the air for the whole weekend", which is translated into Spanish as "que todo quedó en el aire durante todo el fin de semana", i.e. the Spanish translation does not use the equivalent but an adaptation, i.e. "quedar en el aire" instead of "dejar en el aire".

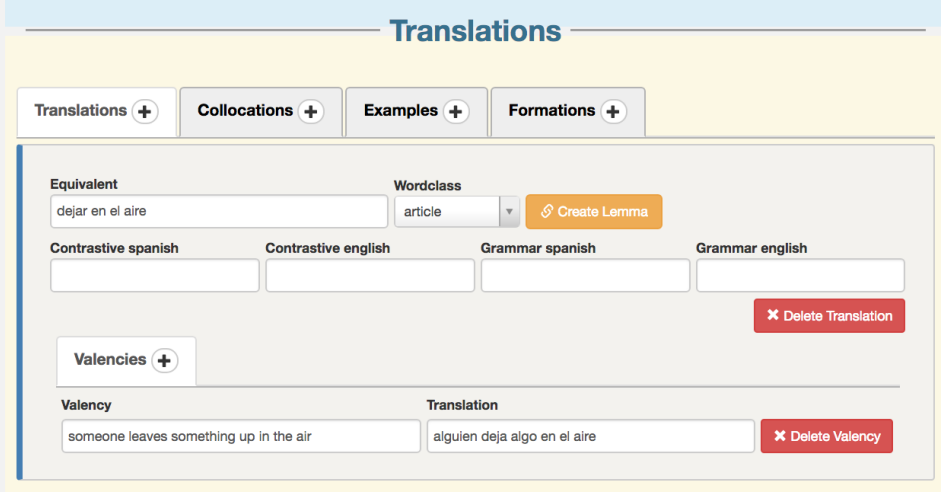

Figure 7: The 'Translation section' in the Dictionary Writing System

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

5.2 Second phase

Phase 2 consists of a single step, i.e. reviewing the dictionary entry completed in Phase 1. This phase is assigned to a member of the International Centre for Lexicography who must check the work done before assigning the status "finished" to the dictionary entry and sending it to the editor of the project for initiating Phase 3. The reviewing phase consists of correcting possible errors, deciding whether the collocations and examples support the meaning and checking possible omissions, e.g. by comparing the dictionary entry with those in existing dictionaries. Should the reviewer find omissions, he or she must analyze them before sending the entry back to the lexicographer with indications about the omissions found. So far, we have found a small number of omissions, less than 2% of the completed entries have been sent back to lexicographers due to omissions.

5.3 Third phase

Phase 3 starts with the editor of the project checking the Spanish equivalent compiled in Phase 1. The editor decides either to convert the equivalent into a Spanish lemma or to leave it only as equivalent. Accepting the equivalent as lemma means reversing the English–Spanish word list. The reversion occurs when the editor clicks "create lemma" (Figure 7) and adds the Wordclass of the equivalent. The "Create lemma" button changes to "Open lemma" (now in green) and clicking on it opens the 'Lemma section' of the Dictionary Writing System corresponding to the Spanish–English side. Figure 8 shows the results of the reversion. An interesting feature is the opening of a drop-down menu on the right side of the 'Lemma section'. This menu refers to the English–Spanish section in which the present lemma was an equivalent and is identified as 'Select to open article'.

Working in this section of the Dictionary Writing System includes all the steps commented on in the above paragraphs, and two more new steps, one for the editor and another one for the lexicographers of the project. The editor has to analyze the presence of the equivalent in the list of 16,000 Spanish words extracted from logfiles and submitted by Ordbogen A/S. If the equivalent is in this list, the equivalent is included in the Spanish lemma list and work continues as shown in the above paragraphs. In most cases, however, the equivalent is not in the list of most searched words. In such a situation, the editor also lemmatizes the equivalent but may postpone the order of compiling, a decision depending on two variables. First, the editor checks whether the equivalent which has now become a lemma is not treated in the monolingual Spanish dictionary, which is also part of the *Diccionarios Valladolid-UVa* and which has more than 50,000 completed lemmata at the time of writing this article (July 2018). In such a situation, the editor usually assigns the equivalent-turned-lemma over to a lexicographer and the work continues as explained in the previous

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

paragraphs. Second, the equivalent-turned-lemma is not completed in the monolingual Spanish dictionary and is not connected with one of the five topics mostly searched for by users and discovered by lexicographers by analyzing the log files of English words starting with "a-". In such a situation, working on the equivalent-turned-lemma is usually postponed.

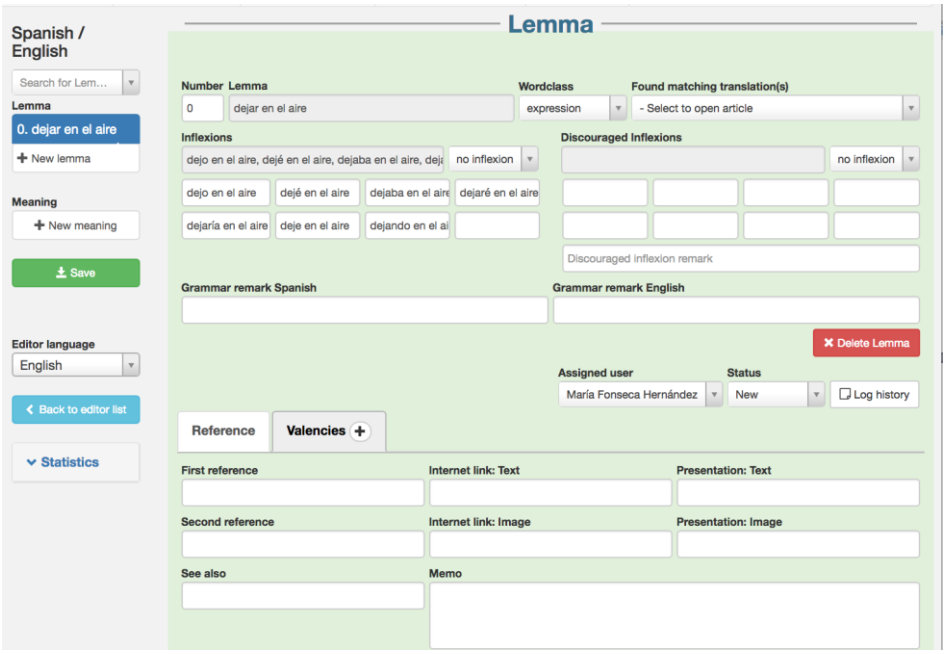

Figure 8: The 'Lemma section' in the Dictionary Writing System after reversion

Lexicographers have an additional step on this side of the bilingual project. They must evaluate the information found when clicking on "Select to open article" (upper right part of Figure 8), which corresponds to English lemmata already described and finished in the English–Spanish section of the bilingual project. The purpose of such an evaluation is to find the best possible English equivalent for the Spanish lemma. For instance, one of the meanings of **abate** is "to put an end to a law, decree, etc." The best Spanish equivalent for such a meaning is *revocar* and therefore *revocar* is lemmatized in the Spanish–English section of the dictionary. When lexicographers study the Spanish word *revocar* and start to explain its different meanings, they must decide that the best English equivalent for the above meaning is **revoke.** There are several reasons for using **revoke** instead of **abate** in such a situation. Three of them are important and illustrate our method of working. The first one is that **abate** is restricted to formal written legal texts, whereas **revoke** is used in a greater variety of texts.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

The second one is that **abate** has a lower frequency of use than **revoke** today (Ngram Viewer). And the third one is that **abate** shows a steep downward trend in use from 1800 to 2000 (Ngram Viewer). Hence, for the Spanish lemma **revocar,** its English equivalent is *revoke* whereas *abate* is a synonym that is assigned the tag "formal" and a synonym remark that explains that its use is restricted to formal English written texts.

Once lexicographers finish the steps already commented on they send their entries for reviewing and reviewers check their work before sending them back to the editor of the project who starts the process again. To sum up, the different phases and steps start with the English–Spanish section and continue with the Spanish–English section, which currently has half of the lemmata of the other section. The data will initially, and mainly, be used to feed the bilingual Spanish–English–Spanish dictionary as well as the Spanish–English Write Assistant, both for native Spanish-speaking users. In the long run, Ordbogen A/S is also planning an English–Spanish Write Assistant for English-speaking users. When the latter has to be prepared lexicographically, we will have to change our compilation order and have more or less the same amount of data in both sections of the bilingual project.

6. Conclusions

The bilingual project described in this paper offers four interesting conclusions for the future of e-lexicography. First, lexicographical data have an intrinsic economic value. This value can be realized provided it is prepared in such a way that it can be used for as many projects as possible, e.g. for developing writing assistants and online dictionaries, and for well-defined users and uses. Secondly, as lexicography is in the middle of a Cambrian explosion, the use of disruptive innovations is necessary to be competitive. For instance, this project shows that collaboration between research institutions and technological companies is fruitful as it guarantees funds, cutting-edge technology and knowledge. In addition, the project uses on a regular basis, systems, methods and resources that have not been used before on a large scale, e.g. logfiles for selecting the initial lemma list and the order of compilation, the Internet for searching for senses and the Ngram Viewer for searching for extended units of meaning. Thirdly, the project's point of departure is the idea that lexicography at its most abstract level is no more and no less than the science concerned with the theory and practice of *dictionaries*, i.e. dictionaries, encyclopedias, lexica, glossaries, vocabularies, terminological knowledge bases, and other information tools covering areas of knowledge and its corresponding language. And finally, the project is based on new ideas and concepts that have not been used so far in the scholarly literature related to bilingual projects, e.g. the existence of an object language and an auxiliary one and the interrelationship of the big transformations affecting today's lexicography.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

Acknowledgments

Special thanks go to the Ministerio de Economía y Competividad for financial support (grant FFI2014-52462-P). Our special thanks to two anonymous reviewers and to Elsabé Taljard for their comments on a previous draft of this paper.

References

Dictionaries

Oxford English Dictionary: https://en.oxforddictionaries.com/.
Oxford Inglés–Español/Español–Inglés: https://es.oxforddictionaries.com/.
Wikipedia: https://www.wikipedia.org/.

Other literature

Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Bergenholtz, H. 1996. Korpusbaseret Leksikografi. *LexicoNordica* 3: 1-15.

Bergenholtz, H. and M. Johnsen. 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes* 34: 117-141.

Bergenholtz, H. and B. Norddahl. 2012. Ordbogsartikler som ingen læser. *LexicoNordica* 19: 207-223.

De Schryver, G.-M. 2013. The Concept of Simultaneous Feedback. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 548-556. Berlin: Walter de Gruyter.

De Schryver, G.-M., D. Joffe, P. Joffe and S. Hillewaert. 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.

Fisker, K. (Ed.). 2018. *Write Assistant. Danish–English*. Odense: Ordbogen A/S.

Fuertes-Olivera, P.A. 2016. A Cambrian Explosion in Lexicography: Some Reflections for Designing and Constructing Specialised Online Dictionaries. *International Journal of Lexicography* 29(2): 226-247.

Fuertes-Olivera, P.A. Forthcoming. Designing and Making Commercially-driven Dictionary Portals: The *Diccionarios Valladolid-UVa. Lexicography.*

Fuertes-Olivera, P.A. and S. Tarp. 2014. Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography. Berlin/Boston: De Gruyter.

Gouws, R.H. and S. Tarp. 2017. Information Overload and Data Overload in Lexicography. *International Journal of Lexicography* 30(4): 389-415.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1460

Hanks, P. 2010. Lexicography, Printing Technology, and the Spread of Renaissance Culture. Dykstra, A. and T. Schoonheim (Eds.). 2010. *Proceedings of the XIV Euralex International Congress,* Leeuwarden, *6–10 July* 2010*:* 988-1016. Leeuwarden: Fryske Akademy.

Hanks, P. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.

Hanks, P. 2013. Lexicography from Earliest Times to the Present. Keith, A. (Ed.). 2013. *The Oxford Handbook of the History of Linguistics*: 503-536. Oxford: Oxford University Press.

Hartmann, R.R.K. and G. James. 1998. *Dictionary of Lexicography.* London/New York: Routledge.

Humbley, J. 2018. Specialised Dictionaries. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 317-334. London/New York: Routledge.

Kilgarriff, A. 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10(2): 135-155.

Kilgarriff, A. 2013. Using Corpora as Data Sources for Dictionaries. Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*: 77-96. London: Bloomsbury.

Ngram Viewer: https://books.google.com/ngrams.

Nielsen, S. 2008. The Effect of Lexicographical Information Costs on Dictionary Making and Use. *Lexikos* 18: 170-189.

Reinhardt, A. 1998. Steve Jobs: 'There's Sanity Returning'. *Business Week*, 25 May 1998. https://www.bloomberg.com/news/articles/1998-05-25/steve-jobs-theres-sanity-returning. Accessed on 29 June 2018.

Rundell, M. 2007. The Dictionary of the Future. Granger, S. (Ed.). 2007. *Optimizing the Role of Language in Technology-enhanced Learning*: 49-51. https://hal.archives-ouvertes.fr/hal-00197203/document/. Accessed on 30 June 2018.

Rundell, M. 2010. What Future for the Learner's Dictionary? Kernerman, I. and P. Bogaards (Eds.). 2010. *English Learners' Dictionaries at the DSNA 2009*: 169-175. Jerusalem: Kdictionaries.

Rundell, M. 2018. Searching for Extended Units of Meaning — and What To Do When You Find Them. *Lexicography.* https://doi.org/10.1007/s40607-018-0042-1. Accessed on 4 July 2018.

Rundell, M. and A. Kilgarriff. 2011. Automating the Creation of Dictionaries: Where Will It All End? Meunier, F., S. de Cock, G. Gilquin and M. Paquot (Eds.). 2011. *A Taste for Corpora. In Honour of Sylviane Granger*: 257-282. Amsterdam/Philadelphia: John Benjamins.

Sinclair, J.M. 1987. Introduction. *Collins Cobuild English Language Dictionary*: xv-xxi. London: HarperCollins.

Tarp, S. 2011. Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. Fuertes-Olivera, P.A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 54–70. London/New York: Continuum.

Tarp, S. and P.A. Fuertes-Olivera. 2016. Advantages and Disadvantages in the Use of Internet as a Corpus: The Case of the Online Dictionaries of Spanish Valladolid-UVa. *Lexikos* 26: 273-295.

Tarp, S., K. Fisker and P. Sepstrup. 2017. L2 Writing Assistants and Context-aware Dictionaries: New Challenges to Lexicography. *Lexikos* 27: 494-521.

Trap-Jensen, L., H. Lorentzen and N.H. Sørensen. 2014. An Odd Couple — Corpus Frequency and Look-up Frequency: What Relationship? *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research* 2(2): 94-113. https://doi.org/10.4312/slo2.0.2014.2.94-113. Accessed on 4 July 2018.