ELIA
Estudios de Lingüística Inglesa Aplicada

THE DYNAMICS OF TYPOGRAPHICAL ERROR REPRODUCTION: OPTIMISING FORMAL CORRECTNESS IN THREE SPECIALISED BILINGUAL DICTIONARIES

DINÁMICAS DE REPRODUCCIÓN DE ERRATAS: OPTIMIZACIÓN DE LA CORRECCIÓN FORMAL EN TRES DICCIONARIOS ESPECIALIZADOS BILINGÜES

Santiago Rodríguez-Rubio
Universidad Pablo de Olavide, Spain
santirrm@hotmail.com

Nuria Fernández-Quesada
Universidad Pablo de Olavide, Spain
nferque@upo.es

Abstract

It is only through an extreme concern for accuracy and the understanding of typographical errors that authors can turn specialised dictionaries into high quality reference works. This paper describes patterns of typographical error reproduction in three specialised English-Spanish dictionaries. We approach intratextual error reproduction (within a particular dictionary), either through related subentries or through non-related subentries. In addition, we compare the frequency of errors between dictionaries written by institutional lexicographers and works written by freelance professionals. The purpose is to provide a model for typographical error detection and analysis that may contribute to formal correctness in reference works. The reason is twofold: a) dictionaries are expected

to be high-standard primary tools for language professionals; b) data quality is essential for a wide variety of utilities, ranging from dictionary writing systems and writing assistants to corpus tools.

Keywords: data quality, data reusability, specialised bilingual lexicography, typographical error reproduction.

Resumen

Los diccionarios especializados no pueden ser considerados obras de referencia de calidad si sus autores no prestan una especial atención a la corrección y si no entienden el fenómeno de la reproducción de las erratas. Este artículo describe patrones de reproducción de erratas en tres diccionarios especializados inglés-español. Abordamos la reproducción intratextual de erratas (en un diccionario en particular), tanto en subentradas relacionadas como no relacionadas. Además, comparamos la frecuencia de erratas en diccionarios elaborados por lexicógrafos institucionales con la de obras realizadas por profesionales independientes. El objetivo es ofrecer un modelo de detección y análisis de erratas que contribuya a la corrección formal en obras de referencia, por dos motivos: a) se supone que los diccionarios deben ser herramientas esenciales de alto nivel para los profesionales del lenguaje; b) la calidad de los datos es fundamental para una amplia gama de herramientas, desde programas de elaboración de diccionarios (*dictionary writing systems*) hasta asistentes de escritura y herramientas relacionadas con córpora.

Palabras clave: calidad de los datos, reciclaje de los datos, lexicografía especializada bilingüe, reproducción de erratas.

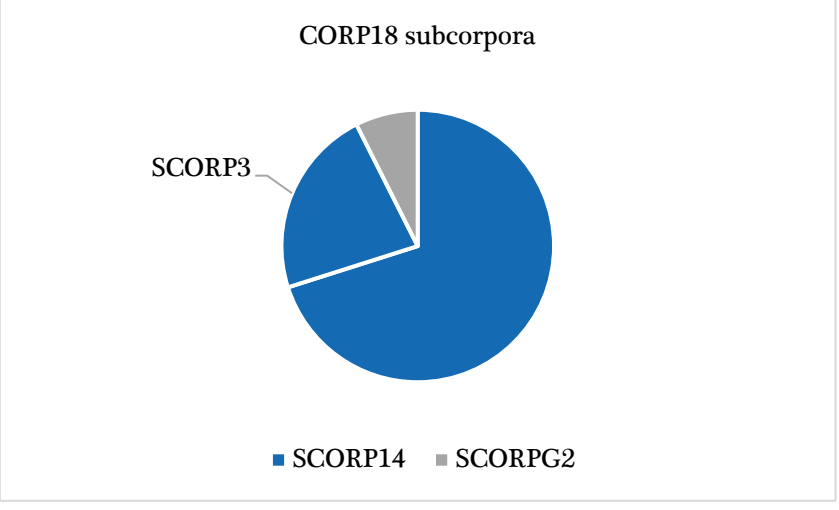
1. Introduction

This article presents the manual analysis of three specialised bilingual paper dictionaries: *Diccionario de Fiscalidad Internacional y Aduanas* (*Ariel*, 2009), *Diccionario de Comercio Internacional*

(*Ariel*, 2007), *Diccionario de Términos de Seguros* (*Ariel*, 2003). They belong to, what we have called, "CORP18", a corpus of eighteen specialised bilingual paper dictionaries of high academic standing published by *Ariel* and *Gestión 2000* (*Grupo Planeta* media group), two prestigious publishing houses. This main corpus is divided in two subcorpora, namely SUBCORPG2 (*Gestión 2000*) and SUBCORP14, also known as the *Alicante Dictionaries* (Mateo, 2018), published by *Ariel*.

The three dictionaries selected for this study make up SUBCORP3: they share the same author and they also feature the highest intratextual error repetition rate within SUBCORP14. Figure 1 illustrates how CORP18 is divided into different subcorpora.

Figure 1: *CORP18 subcorpora*



In this paper, SUBCORP3 and SUBCORP14 will be compared in terms of error frequency and error repetition and reproduction. Results for error frequency in both subcorpora will be then compared to results obtained from SUBCORPG2, a group of

dictionaries which, unlike works from SUBCORP14, feature a simple structure and were elaborated by freelance authors. We will prove that there is a relationship between the complexity of the works, the type of authors involved and error frequency.

Appendix 1 presents information about the eighteen dictionaries, including *inter alia* the date of edition/re-edition, the publishing house, and the letter code assigned to each dictionary. The codes account for the initial letters of each title. Thus, the code for *Diccionario de Fiscalidad Internacional y Aduanas* is DFIA, DCI for *Diccionario de Comercio Internacional*, DTS for *Diccionario de Términos de Seguros*, etc.

2. Formal Correctness and Typographical Error Reproduction in Specialised Texts

It is unquestionable that the printing press made possible the advent of modern lexicography. In Hanks' (2012) words: "… dictionaries as products for widespread general use only became available because of the rapid reproduction of identical copies that printing made possible" (p. 25). In the pre-printing era, manuscripts were copied by hand and included copying errors, as Hanks also points out. However, while "rapid replication and massive dissemination of identical copies of a text –including large and complex texts such as dictionaries – became possible" (p. 25), it is clear that the printing also brought about new perspectives for error reproduction (Estrada, 2012, p. 110).

Gómez & Vargas (2004), co-authors of two dictionaries from SUBCORP14, declared that the purpose of their works was to provide linguistic mediators (translators, interpreters, writers) with a tool designed to improve the quality of writing, proofreading and

translation of specialised texts (p. 365). However, to meet this goal reference works should first exhibit a very high degree of formal correctness.

Tarp (2012) associated lack of quality and occurrence of mistakes with a series of specialised bilingual paper dictionaries generated by computer programmes and published throughout the period 1992-2012. As no trained lexicographer or subject-field expert was involved, those dictionaries were not "confidence-inspiring" according to Tarp (p. 119). Similarly, Biel (2008) claimed that after the communist system fell in Poland, a number of English-Polish poor-quality legal dictionaries appeared. They were prepared "in a hurry without sufficient advisory teams and reviews by lawyers" and typos in those works were "not infrequent" (p. 29). Finally, Gelpí (2007) distinguished between "institutional" and "independent" lexicographers, and attributed less reliable, lower-quality works to the latter (p. 4).

As indicated in Rodríguez-Rubio (2018, p. 77), Robert Hooper (editor of medical dictionaries) stated in 1839 that typos could be transmitted and eventually affect content:

> *It is too little the custom in the present day to take any notice of typographical errors; the consequence of which is, that they are transmitted from one edition to another, till they come at last to affect the meaning, and are rendered permanent. (p. vii).*

Cárdenas (2005) also illustrates this point when he refers to section 1967 of the Spanish Civil Code. This section includes a reference to "the above mentioned three paragraphs", while there are four paragraphs (not three). According to Cárdenas (citing Luis

Díez-Picazo y Ponce de León) this mistake caused an irresolvable doctrinal debate (p. 103). Although the Code has been in force since 1889, the text has yet to be corrected.

Certain typos can go beyond the limits of the works in which they appear. Iamartino (2017) referred to the so-called "ghost words" as:

> *… the result of a mistake – often a spelling mistake in the manuscript handed over to the printer, or a typo – that comes to be included in a dictionary and is later taken for granted and copied verbatim by successive generations of lexicographers. (pp. 64-65)*[1]

Error reproduction can also occur when there is a typo in a prominent position, such as a title. For instance, in DTCF there is an error appearing in the title of the inside covers: *A TERMINOLOGICAL DICTIONARY OF THE PHARMACEUTICAL SCIENCIES** (pp. iii, v). This typo appears again in section "References" of Mateo (2018). The same error also appears in the Virtual Library of the Spanish Royal Academy of Pharmacy (*Real Academia Nacional de Farmacia, RANF*). In this case, the misprint is indicated by means of [sic].

2.1. Technology, Lexicographical Data Reuse and Quality

Beall (2005) stresses the importance of the detection and correction of typographical errors occurring in the data and the metadata of digital databases (p. 6). The author declares that the problem of typographical errors is probably underrated in the online environment: "[Information searchers] likely assume that the search results represent a complete and accurate retrieval based on their search criteria, when in fact, dirty data may be causing some

relevant objects to be excluded" (pp. 4-5). Moreover, Oliveira, Rodrigues & Henriques (2005) manifest that misspellings constitute one of the problems to be solved from the perspective of data quality (p. 12), and according to Hettiarachchi, Attygalle, Hettiarachchi & Ebisuya (2013) typing mistakes contribute to data degradation in databases (p. 96).

Typos occur not only in paper dictionaries, but also as part of the noisy data appearing in digitised versions of paper dictionaries (Zajic, Maxwell, Doermann, Rodrigues & Bloodgood, 2011; Bloodgood & Strauss, 2016). Neither is electronic lexicography free from error reproduction, as lexicographical projects use (and sometimes share) lexicographical databases that may contain mistakes. Even born-digital reference works and dictionary writing systems (DWS) may feature typographical errors (Tavast, Langemets, Kallas & Koppel, 2018). More specifically, misspelling and mistake detection is relevant for corpus query systems (CQS), tools and methods (Rayson, 2015; Kallas, Koeva, Kosem, Langemets & Tiberius, 2019). The interconnection paradigm is the perfect breeding ground for the propagation of errors, for data are reusable and extensively reused[2]. Atkins & Rundell (2008) point out that some DWS "allow you to create 'templates', or generic entries, for common entry-types, containing ready-made configurations of structural elements which can be re-used whenever needed" (pp. 115-116). Mistakes, therefore, are bound to be made and reproduced through different means, whether manual (e.g. typing) or automated (e.g. copy formatting).

Technology improves correctness in texts. In 2017, Tarp, Fisker & Sepstrup presented *Write Assistant*, an information tool designed to assist writers in a second language, feeding on a corpus and on several

digital dictionaries. The authors claimed that they did not use Internet as a source of data because it contains "too much 'noise' in the form of unedited texts and misspelled words", whereas *Write Assistant* required high-quality data (pp. 501-502). The authors stated that "even existing corpora have to be further 'cleaned' and items with less than ten occurrences deleted as a means to avoid as many spelling mistakes as possible" (p. 502). The rationale behind this is that dictionaries contain high-quality data. As we will see, this is not always the case, as far as formal quality is concerned.

Computer technology has yet to solve the problem of typographical error detection and correction, especially regarding erroneous words that can be valid in another context (i.e. *real-word errors*, see below). Dictionary writing systems "help to deliver higher levels of quality, accuracy, and internal consistency" (Atkins & Rundell, 2008, p. 117), including spellcheckers that "minimize the risk of typos" (p. 116). However, it is a fact that spellcheckers overlook mistakes or generate their own, and, ironically, even though computers are expected to improve the quality of dictionaries, their use may discourage formal correctness in dictionary making. As Landau (2001) noted:

> *Management only sees computers as time-savers and expects fewer people to do jobs in the same time or less than in precomputer days. As computers become capable of performing tasks ever more quickly, they also generate the need to perform more tasks (…) In fact, to compensate for the time needed to perform the increasing number of tasks and still preserve tight schedules, some other stage, such as a proofreading stage, may be accelerated or skipped entirely, thus jeopardizing the quality of the book in another dimension. (p. 400)*

3. Precedents on Formal Error Categorisation

For this study, we have categorised typographical errors based on studies from psycholinguistics and natural language processing (NLP), as shown in Rodríguez-Rubio and Fernández-Quesada (2020). Our referent from psycholinguistics (Wells, 1916) is much earlier than the one from the NLP field (Damerau, 1964). Despite the time span, in both works the same four basic categories of mistypings (first case) and misspellings (second case) were established: letter omission, addition, substitution, and transposition.

3.1. Precedents from Psycholinguistics

Wells (1916) studied the psychomotor mechanisms intervening in typing operations. To him we owe the basic categorisation of typing errors: "The errors fall naturally into four sorts,—omissions, substitutions, transpositions (metatheses) and additions" (p. 59).

Rumelhart & Norman (1982) described four categories of typing errors: *transposition errors* (*becuase* for *because*), *doubling errors*, consisting of the repetition of the wrong letter (*scholl* for *school*), *alternation reversal errors* as a variant of the former (*thses* for *these*), and *other errors* (pp. 4-5)[3].

3.2. Precedents from Natural Language Processing (NLP)

Miller & Friedman (1957) and Blair (1960) were some of the first studies developing an algorithm for automatically detecting and correcting spelling errors. To Damerau (partially based on Blair) we owe the four basic error-generating operations (missing letter, extra letter, wrong letter, transposed letter) (1964, p. 171). As we said, Wells had already established these categories in 1916.

Mitton (1987) tackled errors implying a correct but invalid word: "A checker that detects errors simply by looking up words in a dictionary will obviously fail to spot errors that happen to match dictionary words, such as 'wether' for 'whether.' I call these 'real-word errors.'" (p. 496) The author established three types of real-word errors: *wrong-word error* (*know* for *now*), *wrong-form-of-word error* (*thing* for *things*, *use* for *used*), and *word-division error* (*miss dress* for *mistress*) (pp. 497-98). The first type involved the substitution of the valid word for a non-related word, whereas in the second type the wrong word was a form of the valid word.

Kukich (1992) did a thorough analysis of automatic error detection techniques. The author developed Mitton's (1987) range of real-word errors and established the following error generation mechanisms (1992, p. 412): "simple typos" (*from* for *form*), "cognitive or phonetic lapses" (*there* for *their*), "syntactic or grammatical mistakes, including the use of the wrong inflected form" (*arrives* for *arrive*), "wrong function form" (*his* for *her*), "semantic anomalies" (*minuets* for *minutes*), "insertions or deletions of whole words" (*the system has been operating system for*…), and "improper spacing" (*ad here* for *adhere*).

4. Ambiguity as a Limitation to Formal Error Categorisation

As Luelsdorff (1986) has stated: "Error data is often 'noisy', theoretically intractable, in the sense that it frequently admits of a host of mutually contradictory classifications." (p. 53) Literature from the fields of psycholinguistics and NLP has extensively covered this issue (Logan, 1999; Pollock & Zamora, 1984).

Ambiguity can also affect the classification of errors as misspellings or mistypings. In theory, a misspelling is a genuine

error caused by ignorance, whereas a mistyping is a mistake. "Performance errors are simply due to mechanical or neuro-motor problems (typographical errors, 'slips of the pen'), whereas competence errors reflect ignorance about language rules or misconceptions about the domain." (Véronis, 1988) In other words, orthographical errors are cognitive errors while typographical errors are motoric errors (van Berkel & De Smedt, 1988). However, as Mitton (1996) states: "Studies of uncorrected typos face the same data-collection problems as studies of slips of the pen; it is easy enough to collect errors from keyboarded text, but it is impossible to separate the typos from the misspellings." (p. 88) Damerau (1964) focuses on the detection and correction of *spelling errors*, but his description of errors corresponds to what we consider mistypings: "These are the errors one would expect as a result of misreading, hitting a key twice, or letting the eye move faster than the hand." (p. 171) According to Min, Wilson & Moon (2000), the category *spelling errors* covers *typographical errors*, *orthographical errors* and *scanning errors* (p. 1). Finally, Peterson (1980) claims that spelling errors can be narrowed down to: "author ignorance", "typographical errors on typing", and "transmission and storage errors" (e.g. optical character recognition) (p. 677).

In our study, we assumed that we were dealing with typographical errors, as the authors and proofreaders of the dictionaries were supposed to have a sound linguistic knowledge.

5. Materials

This paper analyses three specialised bilingual paper dictionaries in the English-Spanish language pair shown in Table 1. They share the same author and editor, they belong to related fields, and they also

feature the highest intratextual error repetition rate within SUBCORP14, that is why they compose SUBCORP3.

Table 1: *Information about SUBCORP3 dictionaries*

Dict. title/code	Authorship	Length (pages)	Series
Diccionario de Fiscalidad Internacional y Aduanas DFIA	Castro, 2009	1912	*Ariel Economía* (Economy)
Diccionario de Comercio Internacional DCI	Alcaraz, Castro, 2007	1144	*Ariel Derecho* (Law)
Diccionario de Términos de Seguros DTS	Castro (Dir. Alcaraz), 2003	793	*Ariel Derecho* (Law)
		3849	

As previously mentioned, SUBCORP3 is part of SUBCORP14, a collection of fourteen works known in metalexicographical circles as the *Alicante Dictionaries*. This collection opened with *Diccionario de Términos Jurídicos*, which "soon became a milestone in Spanish specialised lexicography" (Mateo, 2018, p. 422). Elaborated by institutional authors, SUBCORP14 works often feature the co-authorship formula. They are linked to "The Academic and Professional English" research group (University of Alicante) and to the IULMA ("Inter-University Institute of Applied Modern

Languages") of the Community of Valencia. The fact that a leading legal consulting firm sponsored DFIA and that the then Spanish Minister of Science signed the preface reinforces the institutional character of this particular work. The same applies to DTCF, which was sponsored by the above referred Spanish Royal Academy of Pharmacy (*Real Academia Nacional de Farmacia, RANF*).

In contrast, SUBCORPG2 dictionaries were made by freelance individual authors. As previously mentioned, SUBCORP3 works will be compared in terms of error frequency with the other works composing SUBCORP14, and also with SUBCORPG2.

6. Methods

SUBCORP3 was scrutinised from beginning to end (page by page). Not only were the bodies (English-Spanish/Spanish-English) analysed, but also the hyperstructure or megastructure of the dictionaries (i.e. cover pages, introductions, etc.). The results of this paper only include those errors found in the bodies. We used homogeneous error detection and classification criteria for all dictionaries composing CORP18.

In this paper, we focus on the following error categories: non-word errors (letter omission, addition-repetition, substitution, or transposition), and real-word errors (word omission, addition-repetition, substitution, or transposition).

6.1. Solving Ambiguity

The manual revision allowed us to solve a large number of ambiguities. Other times we had to admit several corrections for the wrong term. Here is an example of ambiguous error in SUBCORP14 and the method used for solving the ambiguity (in this case, the

consultation of the source text): *The Turkish government has and agreed to retain a 1% golden share of the monopoly*, in the subentry for *golden share* in DTBA, p. 163, and in DTBO, p. 140. The last paragraph of an article entitled "Minister's fall drives Turkish rebound" (July 2001) found in the BBC NEWS website reads: "In May, Mr Oksuz was instrumental in halting the planned sale of over 45% of Turk Telekom and agreed to retain a 1% 'golden' share of the landline monopoly to help ease some of those fears." Supposing that this quote inspired the illustrative sentence in DTBA and DTBO, the coordinating conjunction *and* preceding *agreed* is incorrect[4].

Another method for solving ambiguity is the consultation of cross-references in related subentries. Sometimes, the ambiguity is unsolvable. For instance, in the subentry for *efectuar* of DCI (p. 833) there are two possible solutions for the number disagreement error found in *les rogamos nos lo notifique*, namely *les rogamos nos lo notifiquen*, and *le rogamos nos lo notifique*.

We based our error categorisation on the effects appearing in the erroneous terms, not on their ultimate causes. For instance, an idiomatic but invalid word in a specific context fell under the corresponding real-word error category, irrespective of the mechanism having presumably operated. Thus:

- *form* (for *from*) fell under the category "intralingual real-word error", instead of "transposition non-word error".
- *annual* (for *anual*) was categorised as an interlingual real-word error, instead of an addition non-word error, and so on.

6.2. Non-word Errors

In SUBCORP3, four categories of non-word errors were established:

1. Omission of one or more letters (e.g. *withholing* for *withholding*).

2. Addition of one or more letters. We distinguished between:

    a) Repetition of one or more letters

        - Repetition of a single letter (e.g. *eearner* for *earner*);

        - Addition of letter to a homogeneous digraph (e.g. *betweeen* for *between*);

        - Repetition of syllable or group of letters (e.g. *substantiatiating* for *substantiating*).

    b) Other letter additions (e.g. *Custroms* for *Customs*).

3. Substitution of one letter (e.g. *economiv* for *economic*).

4. Transposition of one or more letters (e.g. *agaisnt* for *against*).

Although our error categorisation does not focus on the causes of the mistakes (e.g. the psychomotor mechanisms having presumably operated), we are aware that repetitions may reveal "anticipation" or "perseveration" of a letter both within or beyond the word limit (Logan, 1999; Lashley, 1951). Among the non-word errors that may have been caused by interlexical anticipation in SUBCORP3, we find: *ingound goods* and *Unites States*. Among the non-word errors presumably caused by perseveration, we have *fully liably* (interlexical) and *policyownery* (intralexical). Among the examples of anticipation and perseveration in SUBCORP14, we find: *foor for thought*, *commond bond*, *market maket*, *retirada inmediatada*, and *correo terrestreo*.

An important psycholinguistic aspect that concerns non-word errors is the possible relationship between the position of the erroneous letter in the word and the *primacy*, *recency* and *bathtub* effects. These phenomena, widely recognised in cognitive sciences, convey the idea that the initial and final positions in a series of items are more salient than the middle positions. Consequently, the elements at the ends are better remembered. If the ends of words are more prominent, it would be expected that the proofreader would detect and correct errors at the initial and final positions to a greater degree, to the detriment of middle position errors[5]. Our results confirm that tendency, as 88% of the non-word errors occurred at middle positions. Out of the remaining 12%, only 2.3% correspond to first-letter position errors, and 9.7% to final-letter position errors.

6.3. Real-word Errors

Following Mitton (1987), we subclassified substitution real-word errors as: *wrong-word error* and *wrong-form error*.

For real-word errors, we used the same four basic categories as for non-word errors:

1. Omission of one or more words (e.g. *The system of temporary does not cover…*).

2. Addition of one or more words, divided into "Repetition of one or more words" (e.g. *error of <u>of</u> fact*) and "Other word additions" (e.g. *rejection of <u>on</u> an offer-in-compromise*). When whole expressions or phrases were repeated, each repeated word was computed.

3. Substitution of one word, divided into:

a) Substitution of word (*wrong-word error*), subdivided into intralingual substitution [e.g. *fiscal authority* (ENG) for *fiscal austerity* (ENG)] and interlingual [e.g. *dividen* (SPA) for *dividend* (ENG)].

b) Modification of inflection (*wrong-form error*), subdivided into gender disagreement, number disagreement, and other modifications. The latter includes different types of errors: adjective for adverb, subject pronoun for possessive pronoun, etc. (e.g. *you* for *your*).

4. Transposition of one or more words (e.g. *a mi leal y saber entender* for *a mi leal saber y entender*). This kind of error was not included in our results, the number of cases being negligible.

6.4. Intratextual/Intertextual Errors and Repeated/Similar Errors

We observed repeated errors and similar errors in one or more dictionaries. For example, in DFIA, we found *taxr*, *taxs*, *tx*, *ta*, *tax tax* (for *tax*), *taxtion* (for *taxation*), whereas *accomodation* was found in DFIA, DTDH (x 2), and DTCF, and *acommodation* was found in DTBA, and DTTO.

Error repetition was indicated by means of the sign "=", whereas we used "~" for similar errors. Since equality is more specific than similarity, repeated errors show a higher indentation level, also related with the distinction between repetition and reproduction of errors that will be explained in the next subsection. In Table 2, the errors in pages 1240-1241, 1241, and 943 are equal. However, the errors in pages 1240-1241 and 1241 are "more equal", as they appear in the same sentences.

Table 2: *Indentation levels in repeated/similar errors*

Representation of the entry content	Page DFIA	Comments
income tax withholding (V. *withholing* of tax at source*)	585	It should read "*withholding*"
~ **withholding allowance** (◊ *He claimed a withholding* allowance for him and…*)	1240-1241	Similar error with the same underlying term
= **withholding exemption** (◊ *He claimed a withholding* exemption for him and…*)	1241	In the same sentence (higher indentation level)
= **remove the obligation to withhold or the witholding* obligation**	943	In a different sentence (lower indentation level)

6.5. Mechanisms of Typographical Error Reproduction

We considered that an error was repeated (intratextually or intertextually) when the same erroneous item appeared at another position in the texts, there being no relationship between the different occurrences (beyond the mere relationship of repetition). We considered that an error was reproduced when it was repeated and there was a relationship between the occurrences of the repeated error. Finally, we considered that an error persisted when it was reproduced over time (whether in different editions of the same dictionary or in other dictionaries).

Table 3 shows examples of repeated errors and reproduced errors in SUBCORP3.

Table 3: *Error repetition versus error reproduction (SUBCORP3)*

Representation of the entry content	Page	Comments	Dict. code
ERROR REPETITION			
fiscal issue (◊ *The new government will have to addresss*…*) = **permanent residence addresss***	477 818	It should read "*address*" Error repeated, but not reproduced (occurrences are not related to one another)	DFIA
ERROR REPRODUCTION			
limited retention (–*retains for its own acount**–) = **retention limit** (–*retains for its own acount**–)	256 396	It should read "*account*" Error repeated and reproduced (occurrences are related to one another)	DTS

We could not verify error persistence in a particular dictionary, as we did not carry out a diachronic analysis of the re-editions of each work. We did observe persistence in different dictionaries edited over time. Table 4 shows errors having persisted in SUBCORP14. The date of edition (or impression) appears below the dictionary code.

Table 4: *Error persistence in SUBCORP14*

Representation of the entry content	Page	Comments	Dict. code (year)
ERROR PERSISTENCE			
cut-throat competition (competividad* feroz)	198	It should read "competitividad"	DTCF (2011)
= **cut-throat competition** (competividad* feroz)	259		DTEFC (2014)
golden share (◊ *The Turkish government has and* agreed to retain a 1 % golden share of the monopoly*)	140	It should read "*… has agreed to…*"	DTBO (2003)
= **golden share** (◊ *The Turkish government has and* agreed to retain a 1 % golden share of the monopoly*)	163		DTBA (2009)

The mechanisms of error reproduction (either intratextually or intertextually) in SUBCORP14 fall within two main categories: a) Error reproduction in related subentries (equivalent or homologous subentries); and b) Error reproduction in non-related subentries (notably through the use of the same illustrative sentence).

a) Error Reproduction in Related Subentries

The lemmas in equivalent subentries (e.g. *endowment assurance|endowment insurance*) convey the same idea in the same section of the dictionary. In turn, the lemmas in homologous subentries convey the same idea, each one in its corresponding section (e.g. *crop-hail insurance|seguro de cosechas*).

Table 5 shows errors reproduced in related subentries in SUBCORP14.

Table 5: *Error reproduction (related subentries)*

Representation of the entry content	Page	Comments	Dict. code
IN EQUIVALENT SUBENTRIES			
endowment assurance (–*stated in the the* policy*–) = **endowment insurance** (–*stated in the the* policy*–)	137-138 138	Repetition of definite article	DTS
IN HOMOLOGOUS SUBENTRIES			
preferential tax treament* sectors (FISC sectores con tratamiento fiscal preferente o privilegiado) = **sectores con tratamiento fiscal preferente o privilegiado** (TAXN preferential tax treament* sectors)	847 1832	It should read "**treatment**"	DFIA

b) Error Reproduction in Non-related Subentries

As indicated above, this mechanism is often related to the use of the same illustrative sentence in two or more subentries[6].

Table 6 shows errors having been reproduced in non-related subentries in SUBCORP3.

Table 6: *Error reproduction (non-related subentries)*

Representation of the entry content	Page	Comments	Dict. code
IN NON-RELATED SUBENTRIES			
address[1] (◊ *Once identified, collection shorfalls* should be property* addressed*) = **collection shortfall** (◊ *Once identified, collection shorfalls* should be property* addressed*)	26 187	It should read "*shortfalls*" and "*properly*", respectively	DFIA

7. Results

Table 7 shows the frequency of errors in each SUBCORP3 dictionary. It indicates whether the error occurs every page, every ten pages, etc. It is the result of dividing the number of pages (1,912 in DFIA, 1,144 in DCI, and 793 in DTS) by the corresponding error incidence (error repetitions included).

The error frequency in DFIA (1,912 divided by 861) is overtly higher than the frequency of the other two works composing SUBCORP3 (3.4 times higher than DCI and three times higher than DTS).

Table 7: *Frequency of errors in SUBCORP3*

ERROR CATEGORY	FREQ. DFIA	FREQ. DCI	FREQ. DTS
Non-word error	**3.57**	**13.78**	**11.17**
Omission	7.80	29.33	23.32
Addition	11.38	40.86	39.65
Repetition	19.51	81.71	56.64
Other addition	27.31	81.71	132.17
Substitution of letter	24.83	104.00	66.08
Transposition	41.57	228.80	158.60
Real-word error	**5.88**	**16.82**	**16.18**
Omission	42.49	71.50	264.33
Addition	15.06	44.00	21.43
Repetition	16.07	63.56	46.65
Other addition	239.00	143.00	39.65
Substitution	12.50	44.00	88.11
Substitution of word (*wrong-word*)	17.07	143.00	158.60
Intralingual	32.41	163.43	396.50
Interlingual	36.08	1144.00	264.33
Modif. of inflection (*wrong-form*)	46.63	63.56	198.25
Gender disagreement	318.67	381.33	0.00
Number disagreement	91.05	95.33	198.25
Other modification	136.57	381.33	0.00
All categories	**2.22**	**7.58**	**6.61**

Table 8 shows the frequency of errors in SUBCORP3, compared to SUBCORP14. It results from dividing the number of pages (3,849 in SUBCORP3, and 11,996 in SUBCORP14) by the corresponding error incidence.

The frequency of SUBCORP3 (3,849 divided by 1,132) is lower than the frequency of SUBCORP14. The frequency of SUBCORPG2 is approximately half the frequency of the other two subcorpora. It is worth reminding that the dictionaries composing SUBCORPG2 are different from those making up SUBCORP14, as the former feature a simpler structure and were elaborated by individual freelancers (not by institutional authors or co-authors).

Table 8: *Frequency of errors in SUBCORP3/SUBCORP14/SUBCORPG2*

ERROR CATEGORY	FREQ. SCORP3	FREQ. SCORP14	FREQ. SCORPG2
Non-word error	**5.58**	**5.35**	**15.69**
Omission	12.10	11.07	30.26
Addition	17.82	21.27	57.77
Repetition	30.55	52.38	254.20
Other addition	42.77	35.81	74.76
Substitution of letter	38.49	31.82	74.76
Transposition	68.73	54.78	0.00
Real-word error	**8.71**	**6.49**	**11.05**
Omission	60.14	69.74	70.61
Addition	20.26	20.79	158.88
Repetition	24.99	26.14	423.67
Other addition	106.92	101.66	254.20
Substitution	20.47	10.93	14.28
Subst. of word (*wrong-word*)	30.79	19.04	24.92

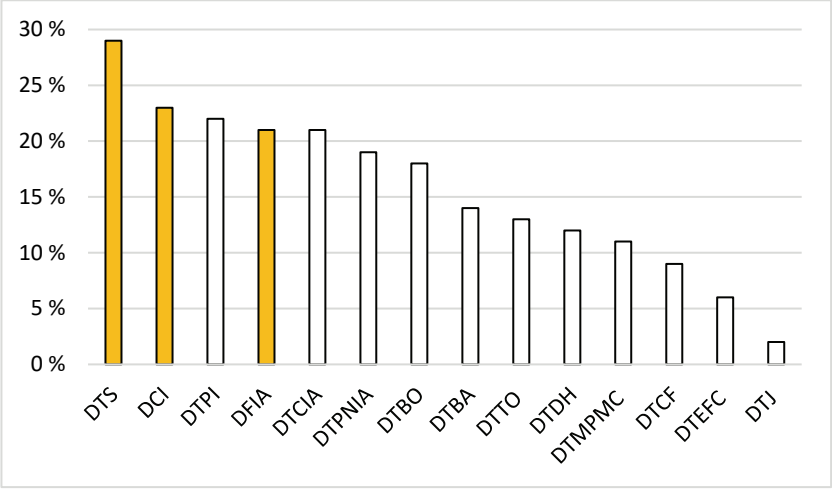
Intralingual	56.60	48.37	43.83
Interlingual	67.53	31.40	57.77
Modif. of inflection (*wrong-form*)	61.10	25.63	33.45
Gender disagreement	427.67	118.77	127.10
Number disagreement	104.03	44.59	84.73
Other modification	226.41	122.41	97.77
All categories	**3.40**	**2.93**	**6.48**

In SUBCORP3, omission non-word errors accounted for 46% of the total number of non-word errors. In the case of SUBCORP14, the proportion of omission non-word errors was 48%. In SUBCORP3, omission and addition-repetition errors all together accounted for 77% of the total number of non-word errors (the proportion being 73% in SUBCORP14).

As shown in Rodríguez-Rubio & Fernández-Quesada (2020), not only is error frequency in DFIA considerably higher than in DTS and DCI, but also DFIA holds the third position in the SUBCORP14 error frequency ranking (whereas DCI and DTS occupy the penultimate and antepenultimate position, respectively).

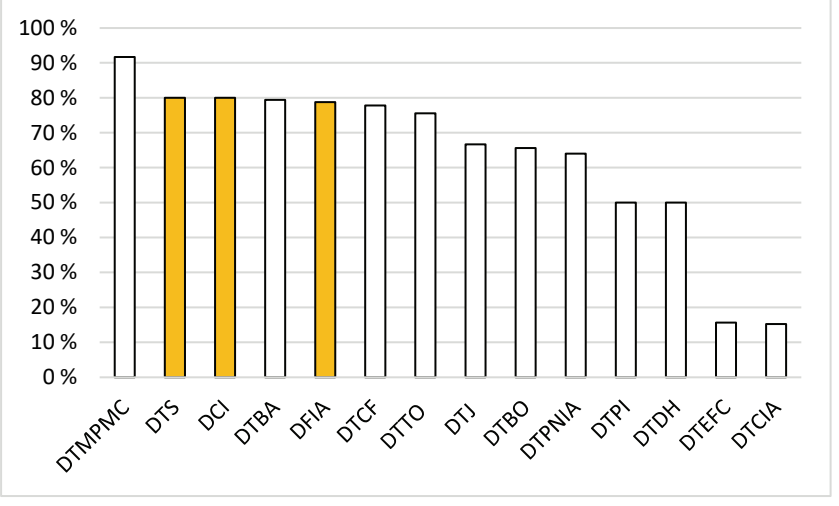
In Figure 2, we observe that the works composing SUBCORP3 hold the first positions in the SUBCORP14 intratextual error repetition ranking (including both non-word errors and real-word errors). DTS, DCI, and DFIA occupy the first, second, and fourth position, respectively. The error repetition rate in SUBCORP3 (22%) is significantly higher than the one found in SUBCORP14 (14%).

Figure 2: *Error repetition % (SUBCORP3/SUBCORP14)*



SUBCORP3 works also occupy the first positions in terms of the proportion of repeated errors involving reproduction, as Figure 3 shows.

Figure 3: *Reproduction % of repeated errors (SUBCORP3/SUBCORP14)*



8. Discussion

In a global context where data have an increasing value and arouse more and more interest, quality is of paramount importance. Our research is relevant from the perspective of data quality and data cleansing, considering that error detection is "a natural first step" in data analysis (Heidari, McGrath, Ilyas & Rekatsinas, 2019, p. 1).

Ambiguity is a limitation to typographical error categorisation. In our study, we classified errors according to the apparent effect observed in the erroneous terms (i.e. omission, addition-repetition, substitution, or transposition). We used context to solve the ambiguities that arose during the classification process. We did not focus on but partially referred to errors presumably caused by anticipation and perseveration. These phenomena reveal interesting psychomotor aspects that may be subject to future research.

New research perspectives may also follow the lead of two aspects based on our quantitative results: 1) the relative position of non-word errors in words; and 2) the most frequent type of repetition in non-word errors. As stated before, first-letter errors only account for 2.3% of non-word errors in SUBCORP14, whereas 9.7% of non-word errors appeared in last letter positions. A deeper analysis could reveal if our findings confirm a widely-observed psycholinguistic phenomenon: end positions in words are more salient than middle positions. Regarding repetition non-word errors, we found that the most frequent type was the addition of a letter to a homogeneous digraph (*betweeen* for *between*), which accounted for 42% of the total number of repetition errors in SUBCORP14 (56% in SUBCORP3). This suggests that it is harder for the proofreader to

spot errors involving the repetition of a letter in a word whose correct form includes a repetition of that letter.

The high error frequencies found and the error reproduction patterns show that errors are not only widespread, but also systemic, revealing what we may call a *negative consistency*, that is, an error consistency.

It seems clear that for the making of the *Alicante Dictionaries* text templates were used. While the use of template entries in dictionaries and databases can improve productivity and consistency in the compilation process (Atkins & Rundell, 2008), typos in those templates will be reproduced unless corrected.

MacKellar wrote in 1893 that "imperfection clings to humanity", and therefore it is a "vain hope" that texts "could be produced perfectly free from errors" (p. 136). Anyone undertaking a long and tedious work like the making of a dictionary "must be guilty of the perpetration of mistakes, blunders, and errors in his copying, however careful he may be" (Wallis, 1920, p. lxxii). However, what we discuss here is not that the dictionaries under study should contain no formal defect, but rather if the error frequency and error patterns found in them are reasonable, if they trespass a *tolerance threshold* (Michiels, 1996) for typographical errors, a concept not yet quantified in literature. Once reproduced, errors can persist through new editions of a dictionary or the reuse of the source text in other reference works. This is of paramount importance to a wide range of language professionals concerned with formal correctness, such as authors, proofreaders, and editors.

As previously stated, scholars have paired low quality with: a) automatically generated dictionaries in which no trained

lexicographer or subject-field expert participated (Tarp, 2012); b) insufficient advisory teams and reviews (Biel, 2008); and c) non-institutional authors (Gelpí, 2007). Dictionaries compiled by institutional lexicographers who are also experts in the fields (like the *Alicante* authors) should, therefore, be expected to feature a higher quality. However, in our study, we have observed that the works by institutional authors and a first-line publishing house as *Ariel* (the *Alicante Dictionaries*) feature a higher frequency of errors than the dictionaries written by freelance authors. The complexity of the *Alicante Dictionaries* may partly account for that discrepancy.

We know from first-hand that the proofreading of dictionaries is a difficult task. Many factors must have interfered with the revision of the works under study: economic and financial constraints, operative and management complexities, etc. Still, the high frequency of errors observed, together with error reproduction (intratextually or intertextually) bespeak deficiencies as far as proofreading is concerned. Mateo (author or co-author of several *Alicante* dictionaries) states that Spanish bilingual paper dictionaries from the field of Economics usually comply with quality requirements that e-dictionaries do not always meet (2014, p. 45). Being prominent exponents of Spanish specialised paper lexicography, the *Alicante Dictionaries* should be expected to exude formal quality.

9. Conclusions

A model of typographical errors was described for SUBCORP3 (*Diccionario de Fiscalidad Internacional y Aduanas*, *Diccionario de Comercio Internacional*, *Diccionario de Términos de Seguros*). We recorded repeated errors and similar errors, both intratextually (in a

particular work), and intertextually (in several SUBCORP3 dictionaries). SUBCORP14, the group of fourteen dictionaries known as the *Alicante Dictionaries* (Mateo, 2018), also featured repeated and similar errors. In both subcorpora, error reproduction operated through: a) related subentries (equivalent or homologous subentries); and b) non-related subentries (typically by using the same illustrative sentences). Typographical error reproduction can affect both paper and online dictionaries, and our findings help to keep errors at bay and to improve the quality of reference works in the future. Typographical errors can linger through time and pass from one edition to another. We could not confirm error persistence within dictionaries, as we did not carry out a diachronic study of the various editions of the works. However, error persistence can be observed among dictionaries that were edited at different moments.

SUBCORP3 holds the first positions in the SUBCORP14 intratextual error repetition ranking, as well as in terms of the proportion of repeated errors that involve reproduction. On the other hand, *Diccionario de Fiscalidad Internacional y Aduanas* shows an error frequency markedly higher than the other two dictionaries composing SUBCORP3 (3.4 higher than *Diccionario de Comercio Internacional* and three times higher than *Diccionario de Términos de Seguros*). In accordance with the academic standing and the prestige of the publisher, and considering the high error frequency observed, we propose that *Diccionario de Fiscalidad Internacional y Aduanas* be revised.

Comparing the error frequency in SUBCORP3/SUBCORP14 to the frequency found in SUBCORPG2 (a subcorpus of dictionaries featuring a simpler structure), we observed that the error frequency was directly proportional to the complexity of the works.

Complexity should not justify a much higher error frequency, as the more complex the work, the more means should be devoted to error detection and correction. Moreover, in our opinion institutional authors should be prepared to assign more resources to those tasks than freelance authors do.

Having in mind Gouws' (2011) guiding idea that dictionaries in a particular period should always be better than previous works (p. 19), our study offers an added value to lexicographers who wish to perfect formal correctness in their dictionaries, and, more precisely, to fight against typographical error reproduction, so important in terms of image and quality perception. In practice, our data showcase the types and subtypes of error that a lexicographer may more frequently encounter. For instance, our results suggest that they should be prepared to find abundant omission non-word errors, which might lead the way for the necessary corrective actions. Further research is required in other lexicographical corpora, in order to ascertain if this tendency could be regarded as a universal benchmark. Moreover, the fact that there is technology capable of detecting many typographical errors does not involve that it will be used correctly or sufficiently. Only practical studies can determine to what extent typographical errors are corrected in dictionaries.

Notes

1. Reverend W. W. Skeat coined the expression *ghost word* in 1886 (Read, 1978), a category from which, as Skeat (1887) himself suggested, real-word errors must be excluded (pp. 352, 356).

2. Two recent examples of shared lexicographical information and interoperability are the EKILEX DWS (Tavast et al., 2018) and the ELEXIS

Project (Declerck, McCrae, Navigli, Zaytseva & Wissik, 2018). Kallas et al. (2019) elaborated on the reuse or integration of lexicographical data in the framework of the latter (p. 46).

3. Some examples of doubling errors in DFIA: *atrraction* for *attraction*, in the subentry for *force of attraction* (p. 485); *cuurent* for *current*, in *tax legislation* (p. 1101).

4. See the journalistic article in: http://bit.ly/2PrAowz

5. The idea is not new. Wheatley (1893) says:

> *One reason why misprints are overlooked is that every word is a sort of pictorial object to the eye. We do not spell the word, but we guess what it is by the first and last letters and its length, so that a wrong letter in the body of the word is easily overlooked. (p. 101)*

6. Leroyer (2018) states that one of "the formal requirements for the quality of good lexicographic examples" is that they do not include corpus noise and typos (pp. 445-446).

References

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Beall, J. (2005). Metadata and Data Quality Problems in the Digital Library. *Journal of Digital Information, 6*(3), 1-20. https://bit.ly/3dqvSbZ

Biel, L. (2008). Legal Terminology in Translation Practice: Dictionaries, Googling or Discussion Forums? *Skase Journal of Translation and Interpretation, 3*, 22-38. http://bit.ly/2PKGZCm

Blair, C. R. (1960). A Program for Correcting Spelling Errors. *Information and Control, 3*(1), 60-67. https://doi.org/10.1016/S0019-9958(60)90272-2

Bloodgood, M., & Strauss, B. (2016). Data Cleaning for XML Electronic Dictionaries via Statistical Anomaly Detection. *Proceedings of the 2016 IEEE 10th International Conference on Semantic Computing (ICSC)*, 79-86. Laguna Hills, CA, USA. https://doi.org/10.1109/ICSC.2016.38

Cárdenas, C. (2005). Las erratas en la publicación de las normas legales. *Themis, 51*, 97-114. http://bit.ly/2kBEW7y

Damerau, F. J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM, 7*(3), 171-176. https://doi.org/10.1145/363958.363994

Declerck, T., McCrae, J., Navigli, R., Zaytseva, K., & Wissik, T. (2018). ELEXIS – European Lexicographic Infrastructure: Contributions to and from the Linguistic Linked Open Data. *GLOBALEX 2018 Workshop: Lexicography and WordNets*, 18-23. http://bit.ly/2w3keTB

Estrada, A. (2012). De errores y erratas. Cómo corregir y normalizar un texto académico. *Normas – Revista de Estudios Lingüísticos Hispánicos 2*, 109-123. https://doi.org/10.7203/Normas.2.4660

Gelpí, C. (2007). Reliability of Online Bilingual Dictionaries. In H. Gottlieb & J. E. Mogensen (Eds.), *Dictionary Visions, Research and Practice* (pp. 3-12). Amsterdam: John Benjamins. https://doi.org/10.1075/tlrp.10.03gel

Gómez, A., & Vargas, Ch. (2004). Aspectos metodológicos para la elaboración de diccionarios especializados bilingües destinados al traductor. In L. González & P. Hernuñez (Eds.), *Las palabras del traductor. Actas del II Congreso "El español, lengua de traducción"* (pp. 365-398). Toledo: Esletra. http://bit.ly/2lVMkKW

Gouws, R. H. (2011). Learning, Unlearning and Innovation. In P. A. Fuertes-Olivera & H. Bergenholtz (Eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography* (paperback edition 2013) (pp. 17-29). London: Bloomsbury.

Hanks, P. (2012). Lexicography and Technology in the Renaissance and Now. *Contributions to the EFNIL 2012 Conference, Budapest*. http://bit.ly/2TrTVOt

Heidari, A., McGrath, J., Ilyas, I. F., & Rekatsinas, T. (2019). HoloDetect: Few-shot Learning for Error Detection. *2019 International Conference on Management of Data (SIGMOD'19), Amsterdam, Netherlands*. Association for Computing Machinery. https://doi.org/10.1145/3299869.3319888

Hettiarachchi, G. P., Attygalle, D., Hettiarachchi, D. S., & Ebisuya, A. (2013). A Generic Statistical Machine Learning and Data Mining Framework for Record Classification and Linkage. *International Journal of Intelligent Information Processing (IJIIP), 4*(2), 96-106. https://doi.org/10.4156/ijiip.vol4.issue2.10

Hooper, R. (1839). *Lexicon Medicum; or Medical Dictionary*. London: Longman. http://bit.ly/36B9UhI

Iamartino, G. (2017). Lexicography, or the Gentle Art of Making Mistakes. *Altre Modernità (Numero speciale — Errors: Communication and its Discontents)*, 48-78. http://bit.ly/2SW0NVk

Kallas, J., Koeva, S., Kosem, I., Langemets, M., & Tiberius, C. (2019). Lexicographic Practices in Europe: A Survey of User Needs. ELEXIS – European Lexicographic Infrastructure, Deliverable D1.1. https://bit.ly/2WJwSlC

Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM* Computing *Surveys, 24*(4), 377-439. https://doi.org/10.1145/146370.146380

Landau, S. I. (2001). *Dictionaries: The Art and Craft of Lexicography*. 2nd edition. Cambridge: Cambridge University Press.

Lashley, K. S. (1951). The Problem of Serial Order in Behavior. In L. A. Jeffress (Ed.), *Cerebral Mechanisms in Behavior* (pp. 112-146). New York: Wiley. https://bit.ly/2zdgvUM

Leroyer, P. (2018). The Oenolex Wine Dictionary. In P. A. Fuertes-Olivera (Ed.), *The Routledge Handbook of Lexicography* (pp. 438-454). Abingdon: Routledge. https://doi.org/10.4324/9781315104942-28

Logan, F. A. (1999). Errors in Copy Typewriting. *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1760-1773. https://doi.org/10.1037/0096-1523.25.6.1760

Luelsdorff, P. (1986). *Constraints on Error Variables in Grammar: Bilingual Misspelling Orthographies*. Philadelphia: John Benjamins. https://doi.org/10.1075/z.25

MacKellar, T. (1893). *The American Printer: A Manual of Typography*. Philadelphia: MacKellar, Smiths & Jordan Foundry. http://bit.ly/2FxNqTS

Mateo, J. (2014). Lexicographical and Translation Issues in the Inclusion of English Financial Neonyms in Spanish Bilingual Dictionaries of Economics on Paper. *Hermes – Journal of Language and Communication in Business, 52*, 41-58. http://bit.ly/2BL4qEl

Mateo, J. (2018). The *Alicante Dictionaries*. In P. A. Fuertes-Olivera (Ed.), *The Routledge Handbook of Lexicography* (pp. 421-437). Abingdon: Routledge. https://doi.org/10.4324/9781315104942-27

Michiels, A. (1996). Electric Words: Dictionaries, Computers, and Meanings. Review of Book *Electric Words: Dictionaries, Computers, and*

Meanings, by Y. A. Wilks, B. M. Slator, & L. M. Guthrie. *Computational Linguistics, 22*(3), 435-440. http://bit.ly/2TnhKbt

Miller, G. A., & Friedman, E. A. (1957). The Reconstruction of Mutilated English Texts. *Information and Control, 1*, 38-55. https://doi.org/10.1016/S0019-9958(57)90061-X

Min, K., Wilson, W., & Moon, Y.-J. (2000). Typographical and Orthographical Spelling Error Correction. *2nd International Conference on Language Resources and Evaluation 221*. http://bit.ly/2PjgN2h

Mitton, R. (1987). Spelling Checkers, Spelling Correctors and the Misspellings of Poor Spellers. *Information Processing and Management, 23*(5), 495-505. https://doi.org/10.1016/0306-4573(87)90116-6

Mitton, R. (1996). *English Spelling and the Computer*. London: Birkbeck ePrints. http://bit.ly/2UevCpB

Oliveira, P., Rodrigues, F., & Henriques, P. (2005). A Formal Definition of Data Quality Problems. *2005 International Conference on Information Quality (ICIQ)*. https://bit.ly/2UweZnG

Peterson, J. L. (1980). Computer Programs for Detecting and Correcting Spelling Errors. *Communications of the ACM, 23*(12), 676-687. https://doi.org/10.1145/359038.359041

Pollock, J. J., & Zamora, A. (1984). Automatic Spelling Correction in Scientific and Scholarly Text. *Communications of the ACM, 27*(4), 358-368. https://doi.org/10.1145/358027.358048

Rayson, P. (2015). Computational Tools and Methods for Corpus Compilation and Analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 32-49). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.003

Read, A. W. (1978). The Sources of Ghost Words in English. *Word, 29*(2), 95-104. https://doi.org/10.1080/00437956.1978.11435650

Rodríguez-Rubio Mediavilla, S. (2018). Análisis cuantitativo de erratas del *Diccionario Terminológico de las Ciencias Farmacéuticas Inglés-Español/Spanish-English* (Ariel, 2007). *Panace@, 19*(47), 76-88. http://bit.ly/35noh9s

Rodríguez-Rubio, S., & Fernández-Quesada, N. (2020). Towards Accuracy: A Model for the Analysis of Typographical Errors in Specialised Bilingual Dictionaries. Two Case Studies. *Lexikos, 30*, 386-415. https://bit.ly/38txJNj

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a Skilled Typist: A Study of Skilled Cognitive-motor Performance. *Cognitive Science, 6*, 1-36. https://doi.org/10.1207/s15516709cog0601_1

Skeat, W. W. (1887). Report upon "Ghost-words", or Words Which Have no Real Existence. *Transactions of the Philological Society 1885-1887*, 350-374. http://bit.ly/2PFC0CR

Tarp, S. (2012). Specialised Lexicography: 20 Years in Slow Motion. *Ibérica, 24*, 117-128. http://bit.ly/30V4nRY

Tarp, S., Fisker, K., & Sepstrup, P. (2017). L2 Writing Assistants and Context-aware Dictionaries: New Challenges to Lexicography. *Lexikos, 27*, 494-521. https://doi.org/10.5788/27-1-1412

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. *Proceedings of the XVIII EURALEX International* Congress*: Lexicography in Global Contexts. Ljubljana, Slovenia, 2018*, 749-761. https://bit.ly/2Ut3MnL

Van Berkel, B., & De Smedt, K. (1988). Triphone Analysis: A Combined Method for the Correction of Orthographical and Typographical Errors. *Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLP)*, 77-83. https://doi.org/10.3115/974235.974250

Véronis, J. (1988). Morphosyntactic Correction in Natural Language Interfaces. *COLING '88 Proceedings of the 12th Conference on Computational Linguistics, 2*, 708-713. https://doi.org/10.3115/991719.991782

Wallis, E. A. (1920). *An Egyptian Hieroglyphic Dictionary (Vol. 1).* London: John Murray. http://bit.ly/2ToX6qq

Wells, F. L. (1916). On the Psychomotor Mechanisms of Typewriting. *The American Journal of Psychology, 27*(1), 47-70. https://doi.org/10.2307/1412853

Wheatley H. B. (1893). *Literary Blunders: A Chapter in the "*History of Human Error*"*. London: Elliot Stock. https://bit.ly/3csDhX5

Zajic, D., Maxwell, M., Doermann, D., Rodrigues, P., & Bloodgood, M. (2011). Correcting Errors in Digital Lexicographic Resources Using a Dictionary Manipulation Language. In I. Kosem & K. Kosem (Eds.), *Electronic Lexicography in the 21st Century: New Applications for New Users (Proceedings of eLex 2011)* (pp. 297-301). Ljubljana: Trojina, Institute for Applied Slovene Studies. http://bit.ly/38T7AUu

Appendix 1. List of CORP18 Dictionaries

N.B. The first fourteen dictionaries conform SUBCORP14, corresponding to the *Alicante Dictionaries*. SUBCORP3 works are marked in grey. The last four dictionaries constitute the subcorpus by the publishing house *Gestión 2000*.

Title (author-s, date of edition/re-edition)	Code	Publisher
1. *Diccionario de Términos de la Banca* (Mateo, 2009)	DTBA	*Ariel*
2. *Diccionario de Términos de la Bolsa* (Mateo, 2003, Dir. Alcaraz)	DTBO	*Ariel*
3. *Diccionario de Términos del Calzado e Industrias Afines* (Alcaraz, Hughes, Mateo, Vargas, Gómez, 2006)	DTCIA	*Ariel*
4. *Diccionario Terminológico de las Ciencias Farmacéuticas/A Terminological Dictionary of the Pharmaceutical Sciences* (Domínguez-Gil, Alcaraz, Martínez, 2007, 2011 impression)	DTCF	*Ariel*
5. *Diccionario de Comercio Internacional* (Alcaraz, Castro, 2007)	DCI	*Ariel*
6. *Diccionario de Términos de Derechos Humanos/A Dictionary of Human Rights* (Campos, 2008, Dir. Alcaraz)	DTDH	*Ariel*
7. *Diccionario de Términos Económicos, Financieros y Comerciales/A Dictionary of Economic, Financial and Commercial Terms* (Alcaraz, Hughes, Mateo, 2012, 6th ed., 2014 impression)	DTEFC	*Ariel*

8. *Diccionario de Fiscalidad Internacional y Aduanas* (Castro, 2009)	DFIA	*Ariel*
9. *Diccionario de Términos Jurídicos/A Dictionary of Legal Terms* (Alcaraz, Hughes, Campos, 2012, 11th ed., 2014 impression)	DTJ	*Ariel*
10. *Diccionario de Términos de Marketing, Publicidad y Medios de Comunicación* (Alcaraz, Hughes, Campos, 2005, 2nd ed.)	DTMPMC	*Ariel*
11. *Diccionario de Términos de la Piedra Natural e Industrias Afines* (Alcaraz, Hughes, Mateo, Vargas, Gómez, 2005)	DTPNIA	*Ariel*
12. *Diccionario de Términos de la Propiedad Inmobiliaria* (Campos, 2003, Dir. Alcaraz)	DTPI	*Ariel*
13. *Diccionario de Términos de Seguros* (Castro, 2003, Dir. Alcaraz)	DTS	*Ariel*
14. *Diccionario de Términos de Turismo y de Ocio* (Alcaraz, Hughes, Campos, Pina, Alesón, 2006, 2nd ed.)	DTTO	*Ariel*
15. *Diccionario Económico, Contable, Comercial y Financiero* (Sanz, 2002)	DECCF	*Gestión 2000*
16. *Diccionario de Economía y Empresa* (Miles, 2002)	DEE	*Gestión 2000*
17. *Diccionario Jurídico* (Ramírez, 2003)	DJ	*Gestión 2000*
18. *Diccionario Inglés de Publicidad y Marketing* (Parra, 2000)	DIPM	*Gestión 2000*

Appendix 2. Examples of Typographical Errors in SUBCORP3/SUBCORP14

A. Intratextual Non-word Errors in SUBCORP3

OMISSION NON-WORD ERRORS

Representation of the entry content	Page DFIA	Comments
income tax withholding (V. *withholing* of tax at source*)	585	It should read "*withholding*"
~ **withholding allowance** (◊ *He claimed a withholding* allowance…*)	1240-1241	Similar error (same underlying term). Variation 1
~ **relieve of the obligation to withhold or of the wihholding* obligation**	936	Variation 2

ADDITION NON-WORD ERRORS

Representation of the entry content	Page DFIA	Comments
audit likelihood (probabibilidad* de ser ser [sic] sometido a una auditoría, posibilidad de ser auditado)	91	It should read "probabilidad". The second error is a repetition real-word error ("ser")
= **chances of audit** (probabibilidad* de ser sometido a una auditoría, probabibilidad* de ser auditado)	163	
~ **probabilidad de ser auditado** (S. *posibibilidad* de ser sometido a una auditoría*)	1742	It should read "*posibilidad*". Similar error (different underlying terms)

SUBSTITUTION NON-WORD ERRORS

Representation of the entry content	Page DFIA	Comments
application for an extension of time to file the tax return (solicitud de una prórraga* del plazo de presentación de la declaración)	69	It should read "prórroga"
= **solicitud de una ampliación del plazo de presentación de la declaración** (S. *solicitud de una prórraga* del plazo de presentación de la declaración*)	1854	In a homologous subentry

TRANSPOSITION NON-WORD ERRORS

Representation of the entry content	Page DFIA	Comments
demandar a alguien (LAW bring a lawsuit agaisnt*… institute procedings [sic] against…)	1426	It should read "against". Omission non-word error in "procedings"
= **interponer un pleito** (LAW bring a lawsuit agaisnt*, bring a suit agaisnt*… institute procedings [sic] against…)	1606	In an equivalent subentry

B. Intratextual Real-word Errors in SUBCORP3

OMISSION REAL-WORD ERRORS

Representation of the entry content	Page DFIA	Comments
abandonment of a claim (equivale a *relinquishment* [sic] *a claim*)	3	Omission of preposition ("*relinquishment of a claim*")
= **abandono de una reclamación** (relinquishment [sic] a claim)	1258	In a homologous subentry

ADDITION REAL-WORD ERRORS

Representation of the entry content	Page DCI	Comments
***swap* en distintas divisas y con diferentes tipos de interés** (S. *permuta financiera en distintas divisas y con diferentes diferentes* tipos de interés*)	1096	Word repetition ("*diferentes*")
~ **cross currency-rate swaps** (permuta financiera/*swap* en distintas divisas y con distintos* diferentes tipos de interés)	154	Other word addition ("distintos")

SUBSTITUTION REAL-WORD ERRORS

N.B. Wrong-word errors are indicated by means of "WWE", whereas we use "WFE" for wrong-form errors:

Representation of the entry content	Page DFIA	Comments
intangible assets (también llamados *inmaterial*/invisible assets*)	609	WWE (interlingual). It should read "*immaterial*"
= **invisible assets** (también llamados *inmaterial*/invisible assets*)	632	In the same sentence
deferment account limit (◊ *If you* deferment account limit is…*)	306	WFE (other modifications). It should read "*your*"
~ **sham corporation** (◊ *You must prove that your* are…*)	1002	It should read "*you*"

C. Intertextual Errors in SUBCORP3/SUBCORP14

ERRORS IN SUBCORP3

Representation of the entry content	Page	Comments
	DFIA	
abandono de la instancia (S. *desestimiento* de una demanda jurídica, desestimiento* de un recurso*)	1258	It should read "*desistimiento*". It could be a genuine spelling error
= **DESESTIMIENTO***	1440	Top of page term
	DCI	
abandonment, abdnt (desestimiento*)	3	
= **desestimiento***	812	
	DTS	
abandonment, abdnt (desestimiento*)	3	
= **abandono** (S. *desestimiento**)	511	

ERRORS IN SUBCORP14

Representation of the entry content	Page	Comments
	DFIA	
burden of … is [placed] on … (◊ *The burden of proving fraud is on the goverment**)	133	It should read "*government*". Omission of letter "n"
= **burden of … rests on …** (◊ *The burden of proving fraud rests on the goverment**)	133-134	
~ **lessen tax barriers** (◊ *Their govenment* has decided…*)	658	Omission of letter "r"
	DTS	
goverment* insurance	185	
= **seguro gubernamental** (goverment* insurance)	757	
	DTTO	
British Rail/Railways (*–goverment*-funded company–*)	70	
= **Railtrack** (*–goverment*-funded agency–*)	301	
	DTEFC	
orden ministerial (governement*/ministerial…)	1258	Addition of letter "e"

D. Errors in SUBCORPG2

Representation of the entry content	Page	Comments
	DJ	
evasion (*v. tax evasión**)	301	It should read "*evasion*". Interlingual substitution error
~ **final decisión* rule**	311	It should read "**decision**"
~ **invasión* of privacy**	341	It should read "**invasion**"
	DIPM	
Terminal markets (Mercados de futuros en el* que…)	329	It should read "los". Number disagreement error
Reverse motion (… que se* luego se reconstruye…).	276	Word repetition ("se")
	DEE	
foreign currency deposit (depósito en moneda extrajera*)	206	It should read "extranjera". Letter omission error
	DECCF	
packaking* and containers	79	It should read "**packaging**". Letter substitution error

First version received: May, 2020
Final version accepted: October, 2020