http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Towards Accuracy: A Model for the Analysis of Typographical Errors in Specialised Bilingual Dictionaries. Two Case Studies

Santiago Rodríguez-Rubio, *Linguist, Seville, Spain (santirrm@hotmail.com)*
and
Nuria Fernández-Quesada, *Department of Philology and Translation, Pablo de Olavide University, Seville, Spain (nferque@upo.es)*

Abstract: This paper presents the results of research on typographical error analysis in two specialised bilingual paper dictionaries: *Diccionario de términos económicos, financieros y comerciales/ A Dictionary of Economic, Financial and Commercial Terms* (*Ariel,* 2012), and *Diccionario de términos jurídicos/A Dictionary of Legal Terms* (*Ariel,* 2012). A model of errors is described, including similar errors and errors that are repeated both intratextually and intertextually. The error frequency in *A Dictionary of Economic, Financial and Commercial Terms* is higher than the average error frequency in a reference corpus of fourteen dictionaries (mainly first editions). This indicates that repeated editions do not always guarantee a higher level of formal correctness. Our results also show that a high frequency of errors does not necessarily entail a high intratextual error repetition rate. On the other hand, we establish a relationship between typographical errors and the access function in dictionaries, as that kind of error can interfere with access to accurate lexicographical information and data retrieval (especially when they occur in lemmas or sublemmas).

Keywords: DATA ACCESS, ENGLISH FOR SPECIFIC PURPOSES (ESP), NON-WORD ERROR, REAL-WORD ERROR, SPANISH BILINGUAL P-LEXICOGRAPHY, TYPOGRAPHICAL ERROR DETECTION

Opsomming: Op weg na akkuraatheid: 'n Model vir die analise van tipografiese foute in gespesialiseerde tweetalige woordeboeke. Twee gevallestudies. In hierdie artikel word die resultate van navorsing oor tipografiese foutanalise in twee gespesialiseerde tweetalige gedrukte woordeboeke voorgestel: *Diccionario de términos económicos, financieros y comerciales/A Dictionary of Economic, Financial and Commercial Terms* (*Ariel*, 2012), en *Diccionario de términos jurídicos/A Dictionary of Legal Terms* (*Ariel*, 2012). 'n Foutmodel beskryf gelyksoortige foute asook foute wat intratekstueel en intertekstueel herhaal word. Die foutfrekwensie in *A Dictionary of Economic, Financial and Commercial Terms* is hoër as die gemiddelde foutfrekwensie in 'n verwysingskorpus van veertien woordeboeke (hoofsaaklik eerste uitgawes). Dit dui daarop dat opeenvolgende uitgawes nie altyd 'n hoër vlak van formele korrektheid waarborg nie. Ons resultate toon ook aan dat 'n hoë frekwensie foute nie noodwendig 'n hoë intratekstuele foutherhalingsyfer tot gevolg het nie. Daarteenoor het ons vasgestel dat daar 'n verband tussen tipografiese foute en die toeganklikheidsfunksie in woordeboeke bestaan aangesien hierdie tipe foute toegang

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

tot akkurate leksikografiese inligting en data-onttrekking (veral wanneer hulle in lemmas of sublemmas voorkom) kan belemmer.

Sleutelwoorde: DATATOEGANG, ENGELS VIR SPESIFIEKE DOELEINDES (ESD), NIE-WOORD-FOUT, WARE WOORD-FOUT, SPAANSE TWEETALIGE P-LEKSIKOGRAFIE, OPSPORING VAN TIPOGRAFIESE FOUTE

1. Introduction

The starting point of this paper is the assumption that a dictionary that has gone through numerous reeditions will feature a very high degree of formal correctness and a relatively low frequency of typographical errors. In order to test this assumption, we have analysed two specialised bilingual paper dictionaries published by *Editorial Ariel* in the fields of Economy and Law (*A Dictionary of Economic, Financial and Commercial Terms*, and *A Dictionary of Legal Terms*). These two works, hereinafter "SUBCORP2", are part of a series of dictionaries known in international circles as "The *Alicante Dictionaries*" (Mateo 2018).[1] Information about these fourteen titles, hereinafter "CORP14", is provided in the Appendix 1. The Spanish title of each work received a code. Thus, *A Dictionary of Economic, Financial and Commercial Terms* was coded as "DTEFC", and *A Dictionary of Legal Terms* as "DTJ". The total length of CORP14 is 11,996 pages. In this paper, we will compare the frequency of errors and the intratextual error repetition rate in SUBCORP2 and CORP14.

Our typographical error classification is based on different studies from the fields of psycholinguistics and natural language processing (NLP). The author who first referred to the four basic categories of typing errors was Wells (1916: 59). In the NLP field, Damerau (1964: 171) defined four main categories of misspellings, being the same categories previously established by Wells. We have adopted those four categories while focusing on two types of typographical errors: (1) non-word errors (letter omission, addition-repetition, substitution, transposition); and (2) real-word errors (word omission, addition-repetition, substitution, transposition).[2]

Peterson (1986: 633-634) first addressed the detection and correction of errors involving the substitution of a grammatical word for another correct word (*horse* for *house*). Subsequently, Mitton (1987: 496-497) would explicitly distinguish between "non-word errors" and "real-word errors". Kukich (1992: 412) significantly developed the study of real-word errors, establishing several error generation mechanisms, such as "simple typos", "syntactic or grammatical mistakes" (including wrong inflected forms), and "insertions or deletions of whole words".

The main aim of our study is to present a model that not only classifies errors, but also establishes connections among them, one of the most notorious connections being error repetition. For instance, errors included in a particular sentence may reappear in the same sentence in another position in the same

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

dictionary or in another dictionary. Our research, therefore, goes beyond the mere counting of errors. It is simply not possible to detect all typographical errors and formal defects in long and complex texts. As Ren and Perrault (1992: 413) stated and this still applies nowadays: "No program is capable of detecting *every* error and capable of always suggesting *the* right correction." Not even a consolidated lexical database as WordNet is free from mistakes, including spelling errors (Horák and Rambousek 2018b: 1024). Still, we believe that all lexicographical errors should be corrected, for dictionaries are: (1) primary translation tools, and (2) influential in standardising the language. Moreover, mistakes can be a valuable source of information as far as ontologies and corpus lexicography are concerned (Domínguez Vázquez et al. 2018: 848).

Typographical error detection cuts across a wide range of areas in lexicography and terminology, including automatic data extraction. For example, the STyrLogism Project, based on the semi-automatic extraction of possible neologisms, used wordlists from dictionaries and corpora and excluded non-words and typographical errors (Stemle et al. 2019: 539-540). Also, Sassolini et al. (2019: 613) refer to the digitisation of *Grande Dizionario della Lingua Italiana* (Battaglia 1961–2002), during which manual and automatic techniques were used to identify spelling errors in the automatically extracted lemmas.

However, typographical error detection can be a tricky task. Vosse (1992: 112) provided the following example: "*Did you actually see the the error in this sentence?*" Landau (2001: 396) claimed that it is normal for a first edition of a dictionary to include "numerous errors". This statement immediately poses two questions: (1) what does "numerous" mean; and (2) what does "errors" mean. In our research, we addressed those questions by clearly defining the categories and subcategories of errors, and by quantifying those errors. On the other hand, Landau referred to errors found in first editions. DTJ, after eleven editions, features a much lower frequency of errors than DTEFC, which has gone through six editions. However, the frequency of typographical errors in DTEFC is higher than the one found in CORP14, mainly made up of first editions (see Table 6 in "5. Results"). This could be an indicator of severity, a concept that presents some limitations when classifying typographical errors.

2. Typographical errors and data retrieval in dictionaries

The access function is one of the main functions established in lexicography (Gouws 1996: 19). It is linked, among other aspects, to the accuracy of the data being accessed, and to how that information can be retrieved. Therefore, accessibility is a key aspect when dealing with both electronic and printed dictionaries. Lew and De Schryver (2014: 347) claim that digital dictionaries imply more frequent and quicker consultation compared to paper dictionaries. They also state (2014: 350) that one of the advantages of online dictionaries is "easier access to the lexical resources", as they are not subject to the constraints of a "fixed macrostructural organization", and information can be accessed through multiple access routes. However, not everything in the garden is rosy, as far as

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

e-lexicography is concerned. For instance, Fuertes-Olivera (2014: 35) claims that the *information overload* linked to e-lexicography may lead users to abandon the consultation (*information death*) or feel anxiety "as they are unsure of the reliability and quality of the data encountered (*information stress*)."

The focus must be placed on the satisfaction of the users' needs. Based on Lew (2008), Fuertes-Olivera and Niño-Amo (2013: 171) refer to *accessology* as "a new discipline that demands empirical data and theoretical considerations (…) with the aim of understanding how users really access information sources in order to retrieve the information they need as quickly and successfully as possible."

According to Landau (2001: 383), lexicographical database systems "provide separate fields for each component of the dictionary entry, so that one can access just those fields and none other." Dziemianko (2018: 667) conveys a similar idea: "Electronic dictionaries facilitate both outer and inner access, that is finding the right entry and the desired information within the entry (Bergenholtz and Gouws 2007: 243)." Dziemianko (2018: 668-669) states that many online dictionaries incorporate advanced matching functions that suggest a range of correct forms when the user introduces a misspelt search term, there being "plenty of room for improving the accuracy of the suggestions …" Deksne et al. (2013: 421) presented the *Tilde Dictionary Browser* (TDB), a browsing environment targeting language learners and teachers, translators, and other users, with the aim of maximising "the likelihood of providing users with a useful result even when searched items do not have a direct match in the dictionary due to misspellings, inflected forms, multi-word items or phrase fragments …" Similarly, Lew (2013: 21) states that modern e-dictionaries incorporate features such as the "did you mean" function, which corrects some misspellings, and the "suggest-as-you-type" facility. For the latter to work properly, the initial characters of the searched word must have been correctly entered, otherwise the system will not recognise them. This is yet another reason why reference works should avoid typographical errors as much as possible.

Töpel (2014) made a thorough review of studies on the use of e-dictionaries between 1993 and 2012. The author referred to a survey conducted by Lemnitzer (2001), the objective of which was to ascertain the reasons why searches in e-dictionaries were unsuccessful. According to Töpel (2014: 27), 62% of the 149,830 accesses contemplated in an initial phase did not succeed due to misspellings in the search words, among other factors. During a second phase, Lemnitzer's allowed the search function to recognize mistakes, and the rate of unsuccessful searches was reduced to 54%. Töpel (2014: 31) also referred to a survey carried out by Bergenholtz and Johnsen in 2005, where the authors found problems with searches due to "the misspelling of words (…), mistakenly writing words as separate words or as one word, incorrect word forms", and other aspects.

So far, we have referred to situations where e-dictionary users type wrong search terms, and the software detects the errors and suggests solutions. Therefore, it has been assumed that the user made the error, and that the text of the dictionary was correctly spelt. However, let us put it the other way round: if

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

users search through automatic means a lemma or sublemma that is actually misspelt in the dictionary, they will not find the corresponding item, unless exactly the same erroneous form appearing in the lemma or sublemma is typed, which is unlikely. In a paper dictionary like DTBA we find the following errors in correlative sublemmas: "**debt finacing***", "**debt finaced* buy-out**". In DTCIA, we find "**leather measurment* systemas***" and "**length mesaurement***". Should those errors occur in an e-dictionary, the user would not be able to access the desired information. The accuracy of the source text is essential for the automatic retrieval of lexicographical information. Koppel et al. (2019: 776) declared that mistakes or typos from the source texts were some of the problems that arose in Sõnaveeb, a portal displaying authentic corpus sentences automatically retrieved from Sketch Engine for Language Learning (SkELL). Koppel et al. (2019: 775) claimed that the occurrence of errors was normal, as they used sentences not previously revised by a lexicographer: "Dictionary users are accustomed to the fact that all data presented in a dictionary are controlled and edited by a lexicographer, and are hence correct." As we will see, the intervention of lexicographers does not necessarily entail a high degree of formal correctness.

3. Severity and typographical errors in dictionaries

Typographical errors have what we may call "the ability to find their way to the published text". Or, as Wheatley (1893: 101) put it: "The curious point is that a misprint which has passed through proof and revise unnoticed by reader and author will often be detected immediately the perfected book is placed in the author's hands." The author noted that a slight misprint such as the transposition of a letter could convey a meaning opposite to the intended one, as in "*unite*" for "*untie*" (1893: 149).

From a lexicographical perspective, Landau (2001: 396) manifested: "Making a dictionary is like painting a bridge: by the time one coat of paint has been applied, the bridge is in need of another. Just so, before a dictionary has been published one should start making plans for its revision." Johnson (1785: 15) said in relation to his own work: "to pursue perfection was, like the first inhabitants of Arcadia, to chase the sun, which, when they had reached the hill where he seemed to rest, was still beheld at the same distance from them." The fact that dictionaries will always be imperfect not being under discussion, let us focus on the severity of typographical errors found in them.

Prinsloo (2016: 235) declares that spelling errors in dictionaries are serious mistakes "since dictionaries are often used to check spelling." By way of example, the author refers to several letter substitution and accent errors found in *The Oxford Junior Primary Dictionary for Southern Africa* (Goodwill et al. 1991): *masadi* (for *mosadi*), *Dobokwane* (for *Dibokwane*), *mogatsa* (for *mogatša*), and *Mosupologo* (for *Mošupologo*). Similarly, Iamartino (2017: 64-65) states that the introduction of so-called "ghost words" (spelling mistakes or typos) in dictionaries is "a real blunder."

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Given the complexity of the typographical errors found in the dictionaries under study, it is not easy to classify them in terms of severity. In the following subsections, we refer to some aspects that could serve as indicators as to whether a particular typographical error is more severe than others.

3.1 Sequences of errors within a particular entry (or within a particular sentence)

In DTEFC, the entry for "**unenforceable**" (p. 832) reads: *inexigible, inapplicable*, inejecutable, que nose* suede* hacer cujmplir** (for *inexigible, inaplicable, inejecutable, que no se puede hacer cumplir*). The erroneous term *suede* entails an additional problem, namely ambiguity (in Spanish, *suede* could be corrected as *puede* but also as *suele*). The context of the entry indicates that the correct term is *puede*, but the user has to take the trouble to solve the ambiguity all the same.

3.2 Substitution real-word errors conveying a sense opposite to the intended one

In DTEFC (p. 756) and DTJ (p. 528), the subentry for "**slowing-down of economic activity**" reads: *contratación* de la actividad* for *contracción de la actividad*.

In DFIA, the subentry for "**TI relief**" (p. 1151) reads: *TI with partial* relief* for *TI with total relief*. Finally, in DTS the entry for "**desidia**" (p. 587) reads: *debida negligencia** for *debida diligencia*.[3]

3.3 Non-word errors involving a long edit distance or ambiguity

The erroneous term *sientros** (for *siniestros*) involves a significant edit distance with regard to the intended word: one transposition of letter "n" and two additions (letter "i" and letter "s"). Subentry for "**outstanding claims**", DTEFC (p. 597).

On the other hand, *puelen* ser sinónimos* contains the same ambiguity mentioned in "3.1 Sequences of errors …" (*pueden* or *suelen*), but in this case it cannot be solved by resorting to the context of the entry. Subentry for "**allocation**", DTJ (p. 45).

3.4 Intratextual or intertextual repetition of errors

The erroneous term *agreeement* appears twelve times in three CORP14 dictionaries, more precisely in DTEFC, DCI, and DFIA (see distribution in "C. Intertextual errors in SUBCORP2/CORP14" in the Appendix 2). The erroneous term *comission* appears twenty-one times in six dictionaries (DTBA, DTBO, DTCF, DTDH, DTS, and DTPI). The erroneous term *commision* appears thirteen times in five dictionaries (DTBA, DTBO, DTCF, DTPNIA, and DTS). We believe that a typographical error being repeated a significant number of times in several

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

dictionaries implies a higher severity, compared to an error being repeated fewer times in a single dictionary, or not being repeated at all.

3.5 Errors in lemmas or sublemmas

Lemmas and sublemmas are prominent items both in paper and electronic dictionaries. Therefore, the occurrence of typographical errors in those positions may be a hint of severity. In CORP14, we found an error (non-word or real-word) in lemmas or sublemmas every 31 pages. The dictionary featuring a higher frequency of errors in those positions was DTPNIA (one error every 14 pages), whereas the work featuring a lower frequency was DTPI (one error every 364 pages). SUBCORP2 figures were: one error every 26 pages in DTEFC and one error every 178 pages in DTJ. Some errors in lemmas/sublemmas occurred intertextually (e.g. "**Finantial* Instrument Exchange**" was found both in DTBA and DTBO).

As indicated above, errors in lemmas or sublemmas are especially important in e-dictionaries, as far as data retrieval is concerned.

4. Materials and methods

4.1 Materials

For this paper, we chose two specialised bilingual paper dictionaries from, what we have called, "CORP14". As previously stated, CORP14 corresponds to the *Alicante Dictionaries*, a group of fourteen English–Spanish/Spanish–English dictionaries having great relevance in Spanish specialised bilingual lexicography and English for Specific Purposes academia.[4] These works are linked to the IULMA ("Inter-University Institute of Applied Modern Languages" of the Community of Valencia). Fuertes-Olivera (2018: 8) referred to the *Alicante Dictionaries* in the following terms:

> These dictionaries stand out as lexicographic milestones in Spanish-speaking countries and high-quality bilingual (English–Spanish/Spanish–English) specialized dictionaries covering different areas, domains, and sub-domains. They are innovative in several aspects that are difficult to find in paper specialized bilingual dictionaries.

Within CORP14, we built SUBCORP2 around two dictionaries: *Diccionario de términos económicos, financieros y comerciales/A Dictionary of Economic, Financial and Commercial Terms* (*Ariel*, 2012); and *Diccionario de términos jurídicos/A Dictionary of Legal Terms* (*Ariel*, 2012). Not only do they feature similar authorship teams and belong to related fields, but they are also the root of the *Alicante Dictionaries*. DTJ and DTEFC first appeared in 1993 and 1996, respectively. These are also the two CORP14 works with the highest number of re-editions.

Table 1 presents information on SUBCORP2:

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Table 1: Authorship, length, and collection of SUBCORP2

Dictionary code	Authorship/Date	Length (pages)	Edition number	Collection
DTEFC	Enrique Alcaraz Varó, Brian Hughes, José Mateo Martínez, 2012 (2014 printing)	1,440	6	*Ariel Economía* (Economy)
DTJ	Enrique Alcaraz Varó, Brian Hughes, Miguel Ángel Campos Pardillos, 2012 (2014 printing)	1,071	11	*Ariel Derecho* (Law)
		2,511		

4.2 Methods

SUBCORP2 was manually examined page by page, following a linear method (from the beginning to the end of the works), and using the same error detection and classification criteria. All parts of the works were analysed, from the bodies (English–Spanish/Spanish–English) to the front pages, forewords, and introductions. In the results presented here, only the errors found in the bodies are included. The elements of the bodies that could not be examined in a homogeneous way (e.g. unnoticeable errors prima facie, such as the ones found in cross-references) were excluded.

The data compilation stage started in 2016. It took approximately three months for SUBCORP2, and twelve months for CORP14. Typographical errors of different kinds were detected and classified (errors dealing with punctuation marks, cross-references, bold type, italics, spacing, etc.). In this paper, we give an account of two error categories especially relevant in quantitative and qualitative terms:[5]

— Non-word errors (letter omission, addition-repetition, substitution, or transposition).
— Real-word errors (word omission, addition-repetition, substitution, or transposition).

Our typology is not based on the psychomotor mechanisms having presumably operated, but on the apparent effects observed in the erroneous words.

4.2.1 Non-word errors

The expression "non-word errors" refers to typographical errors (or spelling errors, in some cases) implying an idiomatically incorrect term (Ahmed et al. 2009: 39).

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

They convey no meaning in any language or in any context. Non-word errors are typically due to human mistakes, and they are usually more easily detectable than real-word errors, regardless of the means used (whether manual or automatic).

As previously mentioned, non-word errors were classified according to Wells (1916: 59) and Damerau (1964: 171). The latter established four main spelling error categories: (1) Substitution of one letter; (2) Omission of one letter; (3) Addition of one letter; and (4) Transposition of two adjacent letters.

In SUBCORP2, we established the following classification of non-word errors:

1. Omission of one or more letters (e.g. *banrupt* for *bankrupt*).
2. Addition of one or more letters, divided in "Repetition of one or more letters", and "Other letter additions" (e.g. *methjod* for *method*). Three main repetition types were described: (a) Repetition of a single letter (e.g. *workker* for *worker*); (b) Addition of letter to a homogeneous digraph (e.g. *agreeement* for *agreement*); and (c) Repetition of syllable or group of letters (e.g. *misdememeanours* for *misdemeanours*).
3. Substitution of one letter (e.g. *wothdraw* for *withdraw*).
4. Transposition of one or more letters, not necessarily adjacent (e.g. *agreeemnt* for *agreement*).

4.2.2 Real-word errors

The expression "real-word errors" refers to typographical errors (or spelling errors) implying an idiomatically correct term, albeit invalid from the contextual point of view. These errors can also be referred to as "context-dependent errors". In some cases, the error may imply the omission of a contextually valid term, or the occurrence of an idiomatically correct word from another language. Real-word errors may be human or machine errors.

In our study, substitution real-word errors were classified according to the distinction made by Mitton (1987: 497-498) between "wrong-word error" (*know* for *now*) and "wrong-form-of-word error" (*was* for *is*, *thing* for *things*, *use* for *used*). In the first type, the erroneous word is different from the valid one. In the second type, the erroneous word is a derivative of the valid one.

The same four basic error categories used for non-word errors were applied to real-word errors, resulting in the following classification of real-word errors for SUBCORP2:

1. Omission of one or more words (e.g. *business to settled**).
2. Addition of one or more words, divided in "Repetition of one or more words" (e.g. *cada una una** *de las doce ciudades*), and "Other word additions" (e.g. *The immigrants were provided them with food* for *The immigrants were provided with food*). In repeated phrases, every repetition was computed (e.g. in *absolute grounds for refusal for refusal*, two repetitions were computed).

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

3. Substitution. This category was divided in:
   (a) Substitution of word (wrong-word error). In turn, divided into intralingual substitution [e.g. *to close human beings* (ENG) for *to clone human beings* (ENG)] and interlingual substitution [e.g. *fondo para contingencies* (ENG) for *fondo para contingencias* (SPA)].
   (b) Modification of inflection (wrong-form error). In turn, divided into gender disagreement, number disagreement, and other modifications. In the latter, different errors were included: adjective for noun, past participle for infinitive form, etc. (e.g. *there is concerned* for *there is concern*).
4. Other real-word errors. This category is not included in our results, as the number of cases was negligible. These are usually word order or transposition errors (e.g. *esta alude teoría a una estrategia* for *esta teoría alude a una estrategia*, where the subject-verb order was inverted).

4.2.3 Repeated/similar errors from an intratextual/intertextual perspective

Error repetition and error similarity were depicted from a two-fold perspective: intratextually (in a particular dictionary), and intertextually (in several dictionaries). We recorded similar errors with the same underlying term, or with a different underlying term. Table 2 shows examples resulting from the combinations of the two paradigms ("Repeated/similar error" and "Intratextual/intertextual error"):

Table 2: "Repeated/similar error" and "Intratextual/intertextual error" paradigms

	Intratextual error (DTEFC)	**Intertextual error in SUBCORP2** (DTEFC + DTJ)	**Intertextual error in CORP14**
Repeated error	*activiación* x 2 (for *activación*)	*navagación* (for *navegación*)	*acount* (DTEFC, DFIA, DTS x 3, DTBA) (for *account*)
Similar error (same underlying term)	*shareholdrs'*, *shareholers'* (for *shareholders'*)	*inversors* (DTEFC), *invesoras* (DTJ)	*agreeement* (DTEFC x 2, DCI x 6, DFIA x 4), *agreeemnt* (DTEFC), *agrement* (DFIA), *disagreeement* (DTJ)
Similar error (different underlying term)	*fabriación*, *diposición*, *enajnación*, *delcaración*	*Bretña* (DTEFC), *Inglatera* (DTJ)	*progresssive*, *regresssive* (DFIA), *objetive*, *subjetive* (DTCF)

Similar errors with different underlying terms (last row in Table 2) may feature different connections. For instance, *fabriación*, *diposición*, *enajnación*, etc. are

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

erroneous terms in words displaying the ending "*-ción*", whereas *Bretña* and *Inglatera* are omission non-word errors referring to "Britain" and "England", respectively. A relationship of antonymy is found in *progresssive/regresssive* and *objetive/subjetive*.

Throughout the compilation stage, relations among various errors were established by means of prospective searches. During the data organisation stage, all CORP14 non-word errors and real-word errors were gathered in individual files, with a view to defining those relations in a more precise way. This was a key aspect, as originally there were fourteen Word files (one per dictionary), each one containing all the errors (non-word errors, real-word errors, and other errors) found in a particular work, so we lacked intertextual perspective.

The intratextual error repetition rate was calculated as follows: all instances of non-word errors were counted in a particular dictionary, followed by the counting of all repeated instances. By means of a simple rule of three, the non-word error repetition rate was calculated for that dictionary. The same applied to real-word errors, and a combined error repetition rate (non-word and real-word) was then calculated for that particular work. The same applied to the rest of dictionaries.

A repeated error was indicated by means of "=". A similar error was indicated by means of "~". The indentation level used for repeated errors was higher than the one for similar errors, as a repeated error features a more specific relation with regard to the reference item than a similar error does. Thus, in Table 3, the indentation level between the errors in pages 101-102 and 102 (featuring a repeated error) is higher than the level existing between the errors in pages 101-102 and 14 (featuring a similar error). Moreover, several "equality levels" were established. For instance, in Table 3, the errors in pages 193-194, 194 and 196 are equal, but the errors in pages 193-194 and 194 show a higher level of equality because they appear in the same sentence. Consequently, the indentation level between the errors in pages 193-194 and 194 is higher than the level existing between the errors in pages 193-194 and 196.

Table 3: Indentation levels and repeated/similar errors

Representation of the entry content	Page	Comments
DTS		
convertible term assurance, CTA (–*polichyolder**–)	101-102	It should read "*policyholder*"
= **convertible term insurance** (–*polichyolder**–)	102	Repeated error. Higher indentation level with regard to the reference item (error in page 101-102)

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

~ **adjustable life policy** (◊ *The policywoner*…*)	14	Similar error. It should read "*policyowner*". Lower indentation level with regard to the reference item (error in page 101-102)
hard sell (equivale a *hard presssure* selling*)	193-194	It should read "*pressure*"
= **hard selling** (equivale a *hard presssure* selling*)	194	Error repeated in the same sentence. Higher indentation level with regard to the reference item (error in page 193-194)
= **high presssure* selling**	196	Error repeated in a different position. Lower indentation level with regard to the reference item (error in page 193-194)

Cases were found of errors being reproduced in different subentries through an illustrative sentence. For example, the following gender disagreement error was found twice in DTCF: "**dieta equilibrada** (◊ *Un* dieta equilibrada es esencial para …*)", and "**equilibrado** (◊ *Un* dieta equilibrada es esencial para …*)" (p. 800 and 827, respectively).

The complex microstructure of the dictionaries under study may have hindered error detection to a certain extent. Many of their articles not only include different sections (typically the main entry, the semantic field, the translation, the exemplification, and cross-references), but also explanations within the translation section. There is no objection to be made regarding this way of presenting information, as it is definitely very instructive having the dictionary user in mind. See below two related DTCF subentries from the English–Spanish area, where several errors occur:

tenofovir disoproxil *n*: FÁRMACO tenofovir disoproxil; fármaco antirretrovírico/antirretroviral *–antiviral drug–*, perteneciente al grupo de los nucleósidos inhibidores de la transcriptasa *-nucleoside reverse transcriptase inhibitiors*–* que inhiben la acción de la transcriptasa inversa *–reverse transcriptase–* incorporándose al nuevo ADN *–DNA–* y evitando así la replicación del virus de la inmunodeficiencia adquirida [VIH] *–human immunodeficiency virus [HIV] replication–*; V. *antiretroviral, HIV, nucleoside reverse transcriptase inhibitiors**.

(p. 588)

zidovudine *n*: FÁRMACO zidovudina; fármaco antivírico/antiviral *–antiviral drug–*, también llamado *azidothymidine,* perteneciente al grupo de los nucleósidos inhibidores de la transcriptasa inversa *–nucleoside reverse transcriptase inhibitior** [sic]*–*, que inhibe la acción de esta enzima incorporándose al nuevo ADN *–DNA–* y evitando así la replicación del virus de la inmunodeficiencia humana [VIH] *–human immunodeficiency virus [HIV] replication–* … ◊ *Zidovudine was the first drug approved by …* ; V. *antiviral, HIV, nucleoside reverse transcriptase inhibitor*.

(p. 637)

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

For revision purposes, the implications of such a complex microstructure is that the proofreader will have to do a fine-grained job, as bilingual text is systematically intermingled and the spellchecker will probably lose track. A way of automatically addressing this problem would be to treat Spanish and English information separately within each article. As pointed out earlier, Landau (2001: 383) and Dziemianko (2018: 667) referred to the possibility of accessing different fields of a dictionary entry separately. However, in complex entries such as the ones of DTCF reproduced above, we doubt computer programs could discern if a part of a particular component (e.g. an explanation written in English within an entry section written in Spanish) actually includes a misspelling, as the spellchecker will indiscriminately mark as erroneous all words written in English, whether they are correctly spelt or not. The only alternative we see is that a fine-grained job is carried out during the preparation of the lexicographical database, so that the spellchecker will later know what language must be applied to each part of the section.

5. Results

Table 4 shows the incidence and frequency of errors in each SUBCORP2 dictionary. The frequency was calculated dividing the number of pages (1,440 in DTEFC, and 1,071 in DTJ) by the corresponding incidence (error repetitions included). Thus, the frequency indicates whether a particular error occurs every page, every two pages, etc. In DTEFC, a frequency of one error every 2.52 pages was found (resulting from dividing 1,440 by 571). In DTJ, a frequency of one error every 7.60 pages was found (resulting from dividing 1,071 by 141).

It is noteworthy that the frequency in DTEFC is three times higher than the one in DTJ.

Table 4: Incidence and frequency of errors in SUBCORP2

ERROR CATEGORY	INCID. DTEFC	INCID. DTJ	FREQ. DTEFC	FREQ. DTJ
Non-word error	**279**	**49**	**5.16**	**21.86**
Omission	126	16	11.43	66.94
Addition	73	15	19.73	71.40
Repetition	22	9	65.45	119.00
Other addit.	51	6	28.24	178.50
Substitution	58	13	24.83	82.38
Transposition	22	5	65.45	214.20

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

ERROR CATEGORY	INCID. DTEFC	INCID. DTJ	FREQ. DTEFC	FREQ. DTJ
Real-word error	**292**	**92**	**4.93**	**11.64**
Omission	11	2	130.91	535.50
Addition	46	27	31.30	39.67
Repetition	40	21	36.00	51.00
Other addit.	6	6	240.00	178.50
Substitution	235	63	6.13	17.00
Subst. of word (wrong-word)	159	29	9.06	36.93
Intralingual	44	19	32.73	56.37
Interlingual	115	10	12.52	107.10
Modif. of inflection (wrong-form)	76	34	18.95	31.50
Gender disagree.	27	11	53.33	97.36
Number disagree.	43	21	33.49	51.00
Other modif.	6	2	240.00	535.50
All categories	**571**	**141**	**2.52**	**7.60**

Table 5 shows the incidence and frequency of errors in SUBCORP2/CORP14. The frequency was calculated dividing the number of pages (2,511 in SUBCORP2, and 11,996 in CORP14) by the corresponding incidence. In SUBCORP2, a frequency of one error every 3.53 pages was found (resulting from dividing 2,511 by 712). In CORP14, a frequency of one error every 2.93 pages was found (resulting from dividing 11,996 by 4,091).

Table 5: Incidence and frequency of errors in SUBCORP2/CORP14

ERROR CATEGORY	INCID. SUBCORP2	FREQ. SUBCORP2	INCID. CORP14	FREQ. CORP14
Non-word error	**328**	**7.66**	**2,244**	**5.35**
Omission	142	17.68	1,084	11.07
Addition	88	28.53	564	21.27
Repetition	31	81.00	229	52.38
Other addit.	57	44.05	335	35.81
Substitution	71	35.37	377	31.82
Transposition	27	93.00	219	54.78

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

ERROR CATEGORY	INCID. SUBCORP2	FREQ. SUBCORP2	INCID. CORP14	FREQ. CORP14
Real-word error	**384**	**6.54**	**1,847**	**6.49**
Omission	13	193.15	172	69.74
Addition	73	34.40	577	20.79
Repetition	61	41.16	459	26.14
Other addit.	12	209.25	118	101.66
Substitution	298	8.43	1,098	10.93
Subst. of word (wrong-word)	188	13.36	630	19.04
Intralingual	63	39.86	248	48.37
Interlingual	125	20.09	382	31.40
Modif. of inflection (wrong-form)	110	22.83	468	25.63
Gender disagree.	38	66.08	101	118.77
Number disagree.	64	39.23	269	44.59
Other modif.	8	313.88	98	122.41
All categories	**712**	**3.53**	**4,091**	**2.93**

The most frequent errors in CORP14 were substitution real-word errors and omission non-word errors (one error every 10.93 pages and every 11.07 pages, respectively). The less frequent errors were omission real-word errors and transposition non-word errors (one error every 69.74 pages and every 54.78 pages, respectively).

If we compare Table 4 and Table 5, we see that the frequency of errors in DTEFC (one error every 2.52 pages) is higher than in CORP14 (one error every 2.93 pages). This aspect is relevant, as most of CORP14 works are first editions, whereas DTEFC is a sixth edition.

Table 6 shows the frequency of errors of each CORP14 dictionary according to each error subcategory. The most frequent error in each work is indicated in pink. SUBCORP2 works are indicated in blue. In the first column, the edition number of those dictionaries having been edited more than once is shown in parentheses. The last column shows the total frequency of errors in each dictionary, from higher frequency to lower frequency.

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Table 6: Itemised frequency of errors in CORP14

	NW OMISS.	NW ADDIT.	NW SUBST.	NW TRANSP.	RW OMISS.	RW ADDIT.	RW SUBST.	TOTAL FREQ.
DTCF	6.1	11.49	19.5	54.6	35.23	11.87	**4.99**	1.58
DTTO (2)	9.52	19.88	29.39	9.14	42.25	9.52	9.8	1.89
DFIA	7.8	11.38	24.83	41.57	42.49	15.06	12.5	2.22
DTPNIA	6.98	17.74	16.78	88.71	124.2	27	9.27	2.36
DTEFC (6)	11.43	19.73	24.83	65.45	130.91	31.3	6.13	2.52
DTBA	7.56	30.24	22.68	57.73	79.38	26.46	8.94	2.57
DTBO	10.32	37.31	18.65	97	69.29	13.47	11.02	2.72
DTDH	21.17	22.41	29.31	381	42.33	29.31	10.58	3.56
DTCIA	7.9	21.87	60.93	60.93	213.25	77.55	26.66	3.84
DTMPMC (2)	35.27	105.8	88.17	264.5	105.8	17.63	10.8	4.72
DTPI	28	364	121.33	182	36.4	26	16.55	5.6
DTS	23.32	39.65	66.08	158.6	264.33	21.43	88.11	6.61
DCI	29.33	40.86	104	228.8	71.5	44	44	7.58
DTJ (11)	66.94	71.4	82.38	214.2	**535.5**	39.67	17	7.6

In twelve (i.e. 85.7%) of the fourteen dictionaries, the most frequent errors were omission non-word errors or substitution real-word errors.

In CORP14, the most frequent subcategory of error was "substitution real-word error" in DTCF (one error every 4.99 pages of that work).[6] The less frequent subcategory of error was "omission real-word error" in DTJ (one error every 535.5 pages of that work). Both the highest frequency and the lowest one are marked in bold in Table 6.

The dictionaries featuring a higher frequency of errors were DTCF and DTTO (one error every 1.58 pages and every 1.89 pages, respectively). The dictionaries featuring a lower frequency of errors were DTJ and DCI (one error every 7.60 pages and every 7.58 pages, respectively). As previously stated, DTEFC is part of the CORP14 works featuring a higher frequency.

Figure 1 shows the intratextual error repetition rate (non-word and real-word combined) of each CORP14 dictionary. SUBCORP2 works appear in blue. The dictionaries featuring a lower error repetition rate are the ones having gone through more editions, namely DTJ and DTEFC. The other two works featuring more than one edition (DTTO and DTMPMC) are also part of the CORP14 dictionaries with a lower error repetition rate:

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

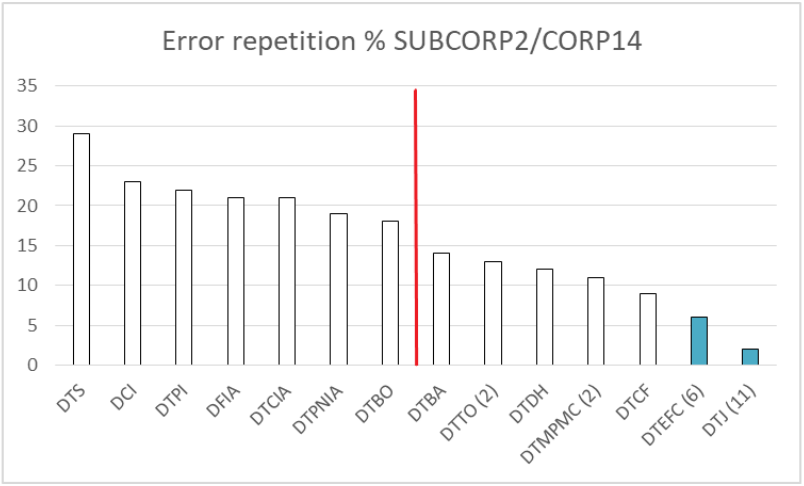

Figure 1: Intratextual error repetition rate in SUBCORP2/CORP14

Comparing Table 6 and Figure 1, it is noteworthy that the dictionaries featuring a higher frequency of errors do not necessarily match the dictionaries featuring a higher error repetition rate. The positions of DTCF, DTTO, DTEFC, DTPI, DTS, and DCI in the frequency ranking diverge from their respective positions in the error repetition ranking. This divergence is especially marked in DTCF, DTS, and DCI. On the other hand, DTJ occupies the last position in terms of both error frequency and error repetition rate, which is consistent with the fact that it has gone through eleven editions.

6. Discussion

Throughout more than three years of research, we found no study having deeply analysed typographical errors in dictionaries. Therefore, what we present here is a novel contribution to the analysis of formal correctness and typographical error detection in bilingual reference works. We have highlighted certain aspects that we consider relevant in terms of severity, among them intratextual or intertextual error repetition, and the occurrence of errors in lemmas or sublemmas. A more systematic analysis of the severity of typographical errors will be the subject of future work.

Access issues are key both in paper and in electronic dictionaries. We believe that formal correctness is particularly important for automatic accessing purposes as far as online dictionaries are concerned. Online dictionaries imply a quicker consultation compared to paper ones, and one of the advantages of the former is an easier access through multiple access routes (Lew and De Schryver 2014: 347, 350). Similarly, Gouws (2014: 175) points out that the article

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

structure in e-dictionaries should be different from that of a printed dictionary: "Data should rather be retrieved from different search zones constituting a multi-layered article structure with a variety of screen shots to present the relevant lexicographic data." Such advantages and innovations concerning access in online dictionaries are of paramount importance, but they rely to a certain extent on formal correctness, because a lemma will not be accessed (or will not be quickly accessed) if it is misspelt. Fuertes-Olivera et al. (2019: 79) state that information tools must procure an easy and quick conversion of lexicographical data into information on the part of users. In our opinion, typographical errors are not consistent with the notion of a quick and easy access to lexicographical information, as they generate noise that may eventually lead to frustration on the part of the user or even to information stress. Erroneous information does not seem to correlate with the idea of "reliability and quality of the data encountered" expressed in Fuertes-Olivera (2014: 35). Besides, from the perspective of a professional translator, the special relevance of a quick access to lexicographical information can be condensed in the saying "time is money".

The practical focus has been in our study from the very beginning. Both direct and indirect applications can be established. The direct application is obviously the correction of the works, or, at least, of the most relevant works. We offered our collaboration to the *Ariel* managing editor, to the director of the IULMA, and to the main lexicographers involved in the making of the *Alicante Dictionaries*. They said they were not interested in our offer, at least for the moment.

Let us now refer to some indirect applications. First of all, our error categorisation can help lexicographers and proofreaders have a clearer picture of the kinds of errors that they will encounter in dictionaries, which, in turn, will be beneficial for revision purposes. More specifically, we believe our research could contribute to perfecting quality control (QC) and quality assurance (QA) procedures in dictionary making. For instance, a list of frequent typographical errors in Spanish–English dictionaries could be elaborated, based on our findings. We observed error reproduction that could occur in any dictionary-making process. Some of the errors detected may relate to the use of common word processing functions, such as copy formatting. But our know-how could also be of interest for dictionary writing systems (DWS), as no software is wholly protected from the seemingly inextricable mechanisms of typographical error generation. The insights we provide could serve as a reference for certain error detection tasks, for example, when revising a lexicographical text where bilingual segments are interwoven in such a way that the normal functioning of the spellchecker is hindered. Under those conditions, a manual revision of the text is highly recommended. The lessons learned throughout our research could especially be of use when dealing with dictionaries that share the same textual sources (a breeding ground for error reproduction). Fuertes-Olivera et al. (2019: 80) refer to the erroneous idea that a new lexicographical project should not make use of previous works. Some pages below, the authors (citing

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Fuertes-Olivera 2016) manifest in relation to a dictionary portal of their own: "The lexicographic data is reusable, subject to a constant process of updating and can be used in conjunction with other tools …" (2019: 83). In this vein, we would recommend lexicographers to establish in their QC procedures mechanisms enabling them to trace those parts of the text (e.g. illustrative sentences) being reused at several locations in a particular dictionary or in other dictionaries. Those specifications, as well as any other specification regarding access issues, must be incorporated in the lexicographical database from the very beginning, in order to fully exploit the capabilities of electronic features. There is no use in a dictionary having a good search engine and a good user interface, if the element they feed on (i.e. the lexicographical database) is not properly designed or contains many typographical errors.

Typographical errors in dictionaries cannot be anticipated. However, based on our quantitative results, we established a number of patterns regarding errors. For instance, 75% of non-word errors occurred in English words, and 94% of non-word errors were found in words having six or more letters. Some of these aspects could help predicting where errors will appear to a certain extent, with a view to envisaging the corresponding corrective actions. A deeper analysis of the compiled errors could contribute to the field of automatic error detection-correction, as a number of word substitution errors found in the dictionaries were attributable to the spellchecker. After all, lexicography and NLP can interact to their mutual benefit (Horák and Rambousek 2018a: 179). Besides, there is room for improvement in spell checking features used in electronic dictionaries (Dziemianko 2018: 668-669). The importance of quality, data management, and data retrieval should be contextualised in the challenge that lexicography is facing in order to offer more user-friendly and accurate tools in the digital era.

7. Conclusions

The results suggest that a higher number of editions does not necessarily correlate with a lower frequency of typographical errors in dictionaries. The sixth edition of *A Dictionary of Economic, Financial and Commercial Terms* features a higher frequency of errors than CORP14, mainly made up of first editions. The fact that DTEFC is longer than CORP14 dictionaries and that it has been subject to several enlargements (with new errors occurring) could partly account for this apparent incongruence, but, in principle, a sixth edition should be expected to contain fewer errors than a first edition. In contrast, *A Dictionary of Legal Terms* shows the lowest frequency of errors found in CORP14 after eleven editions (see Table 6 in "5. Results"). These findings help to provide a base for future research, as all six and eleven editions of the SUBCORP2 dictionaries should be analysed from a diachronic perspective in order to determine if the same errors persist or if the frequency of errors change through several editions.

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Similarly, there is no apparent relationship between a high frequency of errors and a high intratextual error repetition rate, as exemplified by the data obtained from DTCF, DCI and DTS.

Contrariwise, there seems to be a relationship between the number of editions and the intratextual error repetition rate: the error repetition percentage is 2.8 times lower in SUBCORP2 (sixth and eleventh editions) than in CORP14 (mainly first editions). In order to establish a more precise correlation between the number of editions and a dictionary's error repetition rate, again, all editions of DTEFC and DTJ should be analysed.

Typographical error detection and formal correctness are essential aspects affecting the quality of dictionaries. Moreover, typographical errors can hinder data access, notably errors in lemmas or sublemmas in e-dictionaries. This may even have a negative impact on translators, as they usually work with tight deadlines, and for them time is precious. Finally, our model of typographical errors could contribute to expanding knowledge and to offer new perspectives on natural language processing areas such as machine learning or data extraction, with the aim of minimising the occurrence of errors in texts. Not only have we depicted a universe of authentic errors in reference works, but we have also established relationships and observed patterns among those errors, which might be relevant to future studies.

We recommend that *Diccionario de términos económicos, financieros y comerciales/A Dictionary of Economic, Financial and Commercial Terms* (*Ariel*, 2012) be revised on the following grounds. First, a dictionary is a reference work, and as such, it should evince a high degree of formal correctness. Let us not forget that a dictionary is relevant not only for text production, reception and translation, but also for revision and correction purposes (Fuertes-Olivera and Tarp 2008: 79; Fuertes-Olivera 2009: 22). Second, the typographical error frequency in DTEFC is three times higher than it is in *Diccionario de términos jurídicos/A Dictionary of Legal Terms* (*Ariel*, 2012). It is even higher than the overall frequency in the fourteen works of the reference corpus. Third, DTEFC contains a significantly high frequency of errors in lemmas and sublemmas (as compared to the reference corpus). Last, but not least, the DTEFC is one of the cornerstones of the *Alicante Dictionaries*, a series of renowned Spanish works having been recently covered in a prominent international work in the metalexicographical field. It is reasonable to assume that the users of these dictionaries (in Spain and abroad) would expect the work to contain as few errors as possible. The making of these dictionaries (as of any other first-rate dictionary) must have been complex and costly, but formal correctness is a basic requirement that cannot be ignored.

Endnotes

1. We do not deem it necessary to go any deeper into the importance of the Spanish language. Apart from being one of the most widely used languages in the world, its relevance for lexicography has been noted, for instance, in Mairal-Usón and Fuertes-Olivera (2016: 25).

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

2. "Non-word errors" and "real-word errors" are generic expressions that can refer both to typographical errors and to spelling errors. The errors found in the dictionaries under study were considered typographical errors (we assumed that the authors and proofreaders of the works knew the correct spelling).
3. The Spanish expression "*contratación de la actividad*" suggests the idea of "economic upswing", whereas "*contracción de la actividad*" refers to "economic downturn". On the other hand, "*negligencia*" (meaning "negligence") is the opposite to "*diligencia*" (meaning "diligence"). Moreover, the expression "*debida negligencia*" is a *contradictio in terminis*.
4. For further information about these dictionaries, see Mateo (2018). In section "Historical Perspective", Mateo (2018: 422) states that this group of dictionaries is made up of fourteen works, whereas in section "Core issues and topics" (423), we can read: "The *Alicante* specialised dictionaries consist of thirteen specialised dictionaries …" In section "References" (435-436), thirteen titles are listed, not including DCI. We think the correct number is fourteen (not thirteen), as DCI should definitely be considered part of the series.
5. It is worth noting that the subclassification of non-word errors and real-word errors shown here is an improvement on the one presented in Rodríguez-Rubio (2018: 80-81), where we used the same general categories ("non-word error" and "real-word error"), but less homogeneous subcategories.
6. See in Rodríguez-Rubio (2018) a complete analysis of typographical errors detected in DTCF (*Diccionario Terminológico de las Ciencias Farmacéuticas/A Terminological Dictionary of the Pharmaceutical Sciences*). The paper is written in Spanish. Please contact the author for further information.

References

Dictionaries

Battaglia, S. (Ed.). 1961–2002. *Grande Dizionario della Lingua Italiana*. Torino: UTET.

Goodwill, J.S. et al. 1991. *The Oxford Junior Primary Dictionary for Southern Africa with North Sotho, South Sotho, Setswana and Afrikaans Words.* Cape Town: Oxford University Press.

Johnson, S. 1785. *A Dictionary of the English Language (Vol. I)*. Sixth edition (Preface). London: J.F. and C. Rivington.

Other Literature

Ahmed, F., E.W. de Luca and A. Nürnberger. 2009. Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness. *Polibits* 40: 39-48.

Bergenholtz, H. and R. Gouws. 2007. The Access Process in Dictionaries for Fixed Expressions. *Lexicographica. International Annual for Lexicography* 23: 237-260.

Bergenholtz, H. and M. Johnsen. 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes — Journal of Language and Communication in Business* 34: 117-141.

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

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Deksne, D., I. Skadiņa and A. Vasiļjevs. 2013. The Modern Electronic Dictionary that Always Provides an Answer. Kosem, I., J. Kallas, P. Gantar, S. Krek, M. Langemets and M. Tuulik (Eds.). 2013. *Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia*: 421-434. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Domínguez Vázquez, M.J., C. Valcárcel and D. Lindemann. 2018. Multilingual Generation of Noun Valency Patterns for Extracting Syntactic-Semantical Knowledge from Corpora (MultiGenera). Čibej, Jaka, Vojko Gorjanc, Iztok Kosem and Simon Krek (Eds.). 2018. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, 17–21 July 2018, Ljubljana, Slovenia*: 847-854. Ljubljana: Ljubljana University Press, Faculty of Arts.

Dziemianko, A. 2018. Electronic Dictionaries. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*. Abingdon/New York: Routledge: 663-683.

Fuertes-Olivera, P.A. 2009. El *English–Spanish Accounting Dictionary*: un diccionario de internet para traductores. *puntoycoma, Boletín de los traductores españoles de las instituciones de la Unión Europea* 115-S: 22-28.

Fuertes-Olivera, P.A. 2014. Designing Online Dictionaries of Economics: Two Opposing Views. *Hermes — Journal of Language and Communication in Business* 52: 25-40.

Fuertes-Olivera, P.A. 2016. European Lexicography in the Era of the Internet: Present Situations and Future Trends. *Plenary Talk,* Beijing, 2 December 2016. Talk sponsored by the Commercial Press and the Chinese Association of Lexicography.

Fuertes-Olivera, P.A. 2018. Introduction: Lexicography in the Internet Era. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 1-15. Abingdon/New York: Routledge.

Fuertes-Olivera, P.A. and S. Tarp. 2008. La teoría [sic] Funcional de la Lexicografía y sus consecuencias para los diccionarios de economía del español. *Revista de Lexicografía* 14: 75-95.

Fuertes-Olivera, P.A. and M. Niño-Amo. 2013. Internet Dictionaries for Communicative and Cognitive Functions: *El Diccionario Inglés–Español de Contabilidad*. Fuertes-Olivera, P.A. and H. Bergenholtz (Eds.). 2013. *e-Lexicography: The Internet, Digital Initiatives and Lexicography:* 168-186. London/New York: Continuum.

Fuertes-Olivera, P.A., M. Niño-Amo and A. Sastre. 2019. Tecnología con fines lexicográficos: su aplicación en los Diccionarios Valladolid-UVa. *Revista Internacional de Lenguas Extranjeras* 10: 75-100.

Gouws, R.H. 1996. Bilingual Dictionaries and Communicative Equivalence for a Multilingual Society. *Lexikos* 6: 14-31.

Gouws, R.H. 2014. Article Structures: Moving from Printed to e-Dictionaries. *Lexikos* 24: 155-177.

Horák, A. and A. Rambousek. 2018a. Lexicography and Natural Language Processing. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 179-196. Abingdon/New York: Routledge.

Horák, A. and A. Rambousek. 2018b. Wordnet [sic] Consistency Checking via Crowdsourcing. Čibej, Jaka, Vojko Gorjanc, Iztok Kosem and Simon Krek (Eds.). 2018. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, 17–21 July 2018, Ljubljana, Slovenia*: 1023-1029. Ljubljana: Ljubljana University Press, Faculty of Arts.

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

Koppel, K., J. Kallas, M. Khokhlova, V. Suchomel, V. Baisa and J. Michelfeit. 2019. SkELL Corpora as a Part of the Language Portal Sõnaveeb: Problems and Perspectives. Kosem, I. et al. (Eds.). 2019.

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference, 1–3 October 2019, Sintra, Portugal: 763-782. Brno, Czech Republic: Lexical Computing CZ s.r.o.

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

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*. Second edition. Cambridge: Cambridge University Press.

Lemnitzer, L. 2001. Das Internet als Medium für die Wörterbuchbenutzungsforschung. Lemberg, I. et al. (Eds.). 2001. *Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher:* 247-254. Tübingen: Max Niemeyer.

Lew, R. 2008. Lexicographic Functions and Pedagogical Lexicography: Some Critical Notes on Sven Tarp's *Lexicography in the Borderland between Knowledge and Non-knowledge*. Iwan, K. and I. Korpaczewska (Eds.). 2008. *Przegląd Humanistyczny. Pedagogika. Politologia. Filologia*: 114-123. Szczecin: Szczecińska Szkoła Wyższa Collegium Balticum.

Lew, R. 2013. Online Dictionary skills. Kosem, I., J. Kallas, P. Gantar, S. Krek, M. Langemets and M. Tuulik (Eds.). 2013. *Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia*: 16-31. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Lew, R. and G.-M. de Schryver. 2014. Dictionary Users in the Digital Revolution. *International Journal of Lexicography* 27(4): 341-359.

Mairal-Usón, R. and Fuertes-Olivera, P.A. 2016. Recursos tecnológicos y digitales para la gestión del lenguaje científico en español. *Educación Médica* 17(2): 24-38.

Mateo, J. 2018. The *Alicante Dictionaries*. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 421-437. Abingdon/New York: Routledge.

Mitton, R. 1987. Spelling Checkers, Spelling Correctors and the Misspellings of Poor Spellers. *Information Processing and Management* 23(5): 495-505.

Peterson, J.L. 1986. A Note on Undetected Typing Errors. *Communications of the ACM* 29(7): 633-637.

Prinsloo, D.J. 2016. A Critical Analysis of Multilingual Dictionaries. *Lexikos* 26: 220-240.

Ren, X. and F. Perrault. 1992. The Typology of Unknown Words: An Experimental Study of Two Corpora. *Proceedings of the 14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, 23–28 August, 1992. Volume 1:* 408-414.

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

Sassolini, E., A.F. Khan, M. Biffi, M. Monachini and S. Montemagni. 2019. Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study. Kosem, I. et al. (Eds.). 2019. *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference, 1–3 October 2019, Sintra, Portugal*: 603-621. Brno, Czech Republic: Lexical Computing CZ s.r.o.

Stemle, E.W., A. Abel and V. Lyding. 2019. Language Varieties Meet One-Click Dictionary. Kosem, I. et al. (Eds.). 2019. *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference, 1–3 October* 2019, *Sintra, Portugal*: 537-546. Brno, Czech Republic: Lexical Computing CZ s.r.o.

Töpel, A. 2014. Review of Research into the Use of Electronic Dictionaries. Müller-Spitzer, C. (Ed.). 2014. *Using Online Dictionaries*: 13-54. Lexicographica Series Maior. Berlin/Boston: De Gruyter.

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Vosse, T. 1992. Detecting and Correcting Morpho-syntactic Errors in Real Texts. *Proceedings of the 3rd Conference on Applied Natural Language Processing, ANLC '92, Trento, Italy, 31 March–3 April 1992*: 111-118.

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

Wheatley, H.B. 1893. *Literary Blunders: A Chapter in the* "History of Human Error". London: Elliot Stock. Available online at: https://bit.ly/3csDhX5 (Last accessed on 7 October 2020).

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Appendix 1: List of CORP14 dictionaries

N.B. All works are English–Spanish/Spanish–English, published by *Editorial Ariel*. SUBCORP2 works appear in blue.

Title	Code	Authorship/Date	Length (pages)	*Ariel* collection
1. ***Diccionario de Términos de la Banca***	DTBA	José Mateo Martínez, 2009	635	Economy
2. ***Diccionario de Términos de la Bolsa***	DTBO	José Mateo Martínez (ed. Enrique Alcaraz Varó), 2003	485	Law
3. ***Diccionario de Términos del Calzado e Industrias Afines***	DTCIA	Enrique Alcaraz Varó, Brian Hughes, José Mateo Martínez, Chelo Vargas Sierra, Adelina Gómez González-Jover, 2006	853	Industry
4. ***Diccionario Terminológico de las Ciencias Farmacéuticas/A Terminological Dictionary of the Pharmaceutical Sciences***	DTCF	Alfonso Domínguez-Gil Hurlé, Enrique Alcaraz Varó, Raquel Martínez Motos, 2007 (2nd print. 2011)	1,092	Medical Sciences
5. ***Diccionario de comercio internacional***	DCI	José Castro Calvín (ed. Enrique Alcaraz Varó), 2007	1,144	Law
6. ***Diccionario de Términos de Derechos Humanos/ A Dictionary of Human Rights***	DTDH	Miguel Ángel Campos Pardillos (dir. Enrique Alcaraz Varó), 2008	381	Law
7. ***Diccionario de términos económicos, financieros y comerciales/A Dictionary of Economic, Financial and Commercial Terms***	DTEFC	Enrique Alcaraz Varó, Brian Hughes, José Mateo Martínez, 2012 (6th ed., 2nd print. 2014)	1,440	Economy
8. ***Diccionario de Fiscalidad Internacional y Aduanas***	DFIA	José Castro Calvín, 2009	1,912	Economy
9. ***Diccionario de términos jurídicos/A Dictionary of Legal Terms***	DTJ	Enrique Alcaraz Varó, Brian Hughes, Miguel Ángel Campos Pardillos, 2012 (11th ed., 2nd print. 2014)	1,071	Law

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Title	Code	Authorship/Date	Length (pages)	*Ariel* collection
10. ***Diccionario de términos de marketing, publicidad y medios de comunicación***	DTMPMC	Enrique Alcaraz Varó, Brian Hughes, Miguel Ángel Campos Pardillos, 2005 (2nd ed.)	529	Economy
11. ***Diccionario de Términos de la Piedra Natural e Industrias Afines***	DTPNIA	Enrique Alcaraz Varó, Brian Hughes, José Mateo Martínez, Chelo Vargas Sierra, Adelina Gómez González-Jover, 2005	621	Industry
12. ***Diccionario de Términos de la Propiedad Inmobiliaria***	DTPI	Miguel Ángel Campos Pardillos (ed. Enrique Alcaraz Varó), 2003	364	Law
13. ***Diccionario de Términos de Seguros***	DTS	José Castro Calvín (ed. Enrique Alcaraz Varó), 2003	793	Law
14. ***Diccionario de términos de turismo y de ocio***	DTTO	Enrique Alcaraz Varó, Brian Hughes, Miguel Ángel Campos Pardillos, Víctor Manuel Pina Medina, Mª Amparo Alesón Carbonell, 2006 (2nd ed.)	676	Tourism
			11,996	

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Appendix 2: Examples of typographical errors in SUBCORP2/CORP14

A. Intratextual non-word errors in SUBCORP2

OMISSION NON-WORD ERRORS

Representation of the entry content	Page	Comments
DTEFC		
propiedad neta o valor de una propiedad o sociedad (shareholdrs'* equity)	1312-313	It should read "shareholders'"
~ **valor patrimonial** (shareholers'* interest/equity)	1422	Similar error (same underlying term)
DTJ		
improcedencia (inapropriateness*, inoportuneness*)	852	Similar errors (different underlying terms). It should read "inappropriateness" and "inopportuneness", respectively

ADDITION NON-WORD ERRORS

Representation of the entry content	Page	Comments
DTEFC		
trigger level (nivel de activiación*)	819	Other addition. It should read "activación"
= **trigger price** (precio de activiación*)	820	

SUBSTITUTION NON-WORD ERRORS

Representation of the entry content	Page	Comments
DTEFC		
beneficio de inventario (… outcome of an inventary* on the estate …)	941	It should read "inventory". Possible interference of the Spanish term "inventario"
~ **ratio entre existencias y ventas** (invengory*-sales ratio)	1325	

TRANSPOSITION NON-WORD ERRORS

Representation of the entry content	Page	Comments
DTJ		
confinar (restrcti*)	725	It should read "restrict"

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

B. Intratextual real-word errors in SUBCORP2

OMISSION REAL-WORD ERRORS

Representation of the entry content	Page	Comments
DTEFC		
client account (opera en nombre su [sic] cliente)	193	Preposition missing ("en nombre de su cliente")
DTJ		
pendiente[1] (business to settled*)	942	It should read "to be settled"

ADDITION REAL-WORD ERRORS

Representation of the entry content	Page	Comments
DTJ		
motivos de denegación absolutos (absolute grounds for refusal for* refusal*)	910	Repetition of phrase

SUBSTITUTION REAL-WORD ERRORS

N.B. "WWE" is used for wrong-word errors, and "WFE" for wrong-form errors:

Representation of the entry content	Page	Comments
DTEFC		
balloon gas* gone up¡, the	97	WWE (intralingual). It should read "has". The exclamation mark is inverted
DTJ		
intellectual property (la propiedad intelectultual [sic] se divide en dos categories*)	328	WWE (interlingual). It should read "categorías". Besides, letter addition error in "intelectultual" (it should read "intelectual")
DTEFC		
globalizar[1] (◊ *En los primeros años del siglo XXI la economía está globalizar**)	1150	WFE (infinitive for past participle). It should read "*globalizada*"

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

C. Intertextual errors in SUBCORP2/CORP14

REAL-WORD ERRORS IN SUBCORP2

Representation of the entry content	Page	Comments
DTEFC		
appropriate intellectual property (piratear la propiedad intellectual*)	66	It should read in Spanish ("intelectual")
= **World Intellectual Property Organization, WIPO** (organización mundial para la defensa de la propiedad intellectual*)	857	
= **Organización Mundial para la Defensa de la Propiedad Intellectual*** (World Intellectual Property Organization, WIPO)	1259-260	There is a discrepancy in the use of initial letter upper case, compared to the previous example
= **derechos de la propiedad intellectual*, DPI** (intellectual property rights, IPR)	1048	
DTJ		
intellectual (intellectual*)	328	It should read in Spanish ("intelectual")

NON-WORD ERRORS IN CORP14

Representation of the entry content	Page	Comments
DTEFC		
coefficient (V. *agreeement* coefficient*)	199	It should read "*agreement*". Addition of letter to a homogeneous digraph
= **acuerdo sobre aumento de salarios según productividad** (annual improvement agreeement*)	889	In the previous lemma, there is an omission error in the corresponding Spanish word ("**acurdo* sintético de tipos de cambio de divisas a plazo**"). It should read "**acuerdo**"
~ **venta a plazos** (sale or agreeemnt*)	1427	Similar error (same underlying term). Transposition
DTJ		
disconformidad (disagreeement*)	783	It should read "disagreement"

http://lexikos.journals.ac.za; https://doi.org/10.5788/30-1-1606

Representation of the entry content	Page	Comments
DCI		
bilateral agreement (V. *tripartite agreeement**)	63-4	
= **bilateral contract** (V. *tripartite agreeement**)	64	In the same phrase
= **bilateral treaty** (V. *tripartite agreeement**)	64	
= **sector-specific trade agreeement***	532	
= **sectoral agreeement***	532	
= **sectoral trade agreeement***	532	
~ **treaty** (V. *international agreeements**)	606	It should read "*agreements*"
DFIA		
bilateral agreement (V. *tripartite agreeement**)	117	Same entry/phrase as DCI (63-4)
= **bilateral arrangement or contract** (V. *tripartite agreeement**)	117	
= **bilateral contract** (V. *tripartite agreeement**)	117	Same entry/phrase as DCI (64)
= **bilateral treaty** (V. *tripartite agreeement**)	118	Same entry/phrase as DCI (64)
~ **limited power of attorney** (V. *general agency agrement**)	666	Letter omission