http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

The Mental Lexicon in Lexicography: The *Diccionarios Valladolid-UVa*

Pedro A. Fuertes-Olivera, *Department of Afrikaans and Dutch, University of Stellenbosch, South Africa; International Centre for Lexicography, University of Valladolid, Spain; and Centre of Excellence in Language Technology, Ordbogen A/S, Odense, Denmark (pedro@emp.uva.es)*

Abstract: This article analyzes the possibility of making dictionaries that take into consideration the mental lexicon, i.e. words do not work in isolation; instead, they are dynamic constructs that are activated, stored, processed and retrieved gradually. For that, it proposes several general lexicographical and methodological ideas and illustrates them referring to their implementation in the *Diccionarios Valladolid-UVa*: (a) dictionary types are a thing of the past; (b) words are not only holistic products but also processes that are always on the move; consequently their descriptions in dictionaries must be as complete and precise as possible; (c) dictionaries must be equipped with dynamic search system, e.g. systems for allowing human and machine-users search and retrieve *a la carte*, e.g. in a speaking situation; (d) there must be a huge number of words and other data types for describing each meaning and usage of each lemma, thus favoring the creation of patterns and the learning process associated with Artificial Intelligence (AI); (e) designing and making online dictionaries is a cooperative process in which lexicographers and several types of experts must participate; (f) the main task of lexicographers is the preparation of lexicographical data, which can be used in many different forms, formats and usages, being the making of dictionaries one of them.

Keywords: E-LEXICOGRAPHY, ARTIFICIAL INTELLIGENCE, MENTAL LEXICON, WORDS AS PRODUCTS, WORDS AS PROCESSES, LEXICOGRAPHICAL METHODOLOGY

Opsomming: Die kognitiewe leksikon in die leksikografie: Die *Diccionarios Valladolid-UVa.* Hierdie artikel analiseer die moontlikheid om woordeboeke wat die kognitiewe leksikon in ag neem, te skep, m.a.w. woorde funksioneer nie in isolasie nie; inteendeel, hulle is dinamiese konsepte wat geleidelik geaktiveer, gestoor, geprosesseer en onttrek word. Met hierdie doel in gedagte word verskeie algemene leksikografiese en metodologiese idees aangebied en geïllustreer deur na hul toepassing in die *Diccionarios Valladolid-UVa* te verwys: (a) woordeboektipes behoort tot die verlede; (b) woorde is nie net holistiese prdukte nie, maar ook prosesse wat gedurig beweeg; gevolglik moet hul beskrywings in woordeboeke so volledig en presies moontlik wees; (c) woordeboeke moet toegerus word met dinamiese soekstelsels, bv. stelsels wat soektogte en onttrekkings deur menslike en masjien-gebruikers *a la carte* toelaat soos in 'n gespreksituasie; (d) daar moet 'n groot aantal woorde en ander datatipes vir die beskrywing van elke betekenis en gebruik van 'n lemma wees om sodoende die skep van patrone en die aanleerproses wat met Kunsmatige Intelligensie (KI) geassosieer word, te steun; (e) die ontwerp en skep van aanlyn woordeboeke is 'n koöperatiewe proses waaraan leksikograwe en verskeie soorte kundiges moet deelneem;

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

(f) die hooftaak van leksikograwe is die voorbereiding van leksikografiese data, wat in baie verskillende vorms, formate en toepassings gebruik kan word, met die skep van woordeboeke as een daarvan.

Sleutelwoorde: E-LEKSIKOGRAFIE, KUNSMATIGE INTELLIGENSIE, KOGNITIEWE LEKSIKON, WOORDE AS PRODUKTE, WOORDE AS PROSESSE, LEKSIKOGRAFIESE METODOLOGIE

0. Introduction

This paper revolves around two broad issues. The first one considers the lexicon from a cognitive point of view. Such an approach has shown its centrality, e.g. in Natural Language Processing (NLP), and its dynamicity, i.e. its functioning within dynamic networks that illustrate that words do not work in isolation; instead, their working somehow reproduces the modularity, parallelism, dynamicity and high connectivity of the human brain (Section 1, below). This means that words are, on the one hand, products, i.e. holistic entities, and, on the other hand, processes, i.e. entities on the move. As products, they are codified and can be stored in knowledge resources, e.g. dictionaries. As processes, they are always changing, modifying and/or adapting their meanings, forms and usages to different environments.

Research, e.g. Indefrey and Levelt (2004), has maintained that starting from meanings, the speaker initially activates *lemmata*, i.e. abstract lexical forms devoid of lexical information, and then *phonological forms*, i.e. sounds, syllable, phonemes. Such a process connects the first issue with the second one, which considers whether we can devise an ecosystem to support word finding, i.e. word access at the moment of speaking and writing, e.g. by relying on Computer Science to overcome the challenges this transformation poses for traditional lexicography (Section 2, below).

This paper, then, assumes that we can design and make dictionaries that facilitate the brain processes used when humans are employing languages, the so-called "mental-lexicon" (Aitchison 2003; see Section 1, below), and that Computer Science, especially Artificial Intelligence (AI) may help us in the abovementioned task. It illustrates them with ideas and data from the design and making of the *Diccionarios Valladolid-UVa* (Section 3). This lexicographical project has been discussed in lexicographical environments that are basically related with the user as reader (Tarp and Fuertes-Olivera 2016; Fuertes-Olivera 2019; Fuertes-Olivera and Tarp 2020; Fuertes-Olivera, Tarp and Sepstrup 2018; Fuertes-Olivera and Esandi-Baztan 2020). In this paper, I will focus on lexicographical decisions that aim at suiting the speaker/writer when they need products that can be easily converted into processes. This user can be a human or a machine, as I believe that the future of lexicography is in the preparation of lexicographical data that will be used in a myriad of usages, forms, formats and purposes (Fuertes-Olivera and Tarp 2020).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

1. The mental lexicon

Jackendoff (2002) claims that the mental lexicon is a kind of dictionary that, firstly, contains individual speakers' word representations, which are described in terms of their meanings, pronunciations, formal and functional characteristics, and so on; and, secondly, deals with how those words are activated, stored, processed, and retrieved by each speaker. This view assumes that the mental lexicon or *brain dictionary* is an individual's construct that is constantly changing and growing as new words are learned and old words are forgotten. As such, users of the brain dictionary activate their search gradually, i.e. depending on individual user needs and situations.

The above view opposes the approach mostly present in *traditional dictionaries,* i.e. repositories of words that have a physical or digital form. Traditional dictionaries also contain descriptions of the meaning, form, pronunciation, syntactic characteristics and so on of words but, to the best of my knowledge, most of them take for granted that their words are holistic entities and as such can only be activated, stored, processed and retrieved holistically, i.e. in a way that seems to be different from the modular, dynamic, parallel and highly connective system used by our brain (Emmorey and Fromkin 1988).

Research (Aitchison 2003) has also shown that the brain dictionary is not organized alphabetically. Instead, its organization seems to be connected with the existence of neural circuits. These are subjected to processes such as *spreading activation*, a concept proposed in semantic network theory (Forster and Chambers 1973; Marslen-Wilson 1987) to refer to a "hypothetical mental process that takes place when one of the nodes in the semantic network is activated" (Traxler 2012: 84).

Within the tenets of semantic network theory, three main approaches to the activation of the brain dictionary have been proposed: (a) priming; (b) neighborhood effects; and (c) frequency effects. Priming basically defends the activation of related words once a particular word is searched for; for instance, if we have "euro", priming will activate "European Union" (Hoey 2005). Neighborhood effects (Andrews 1989) refer to the activation of all similar "neighbors" of a target word, i.e. items that are highly confusable with the target word, e.g. for the word "game", its neighbors will be "came, dame, fame, lame, name, same, tame, gale, gape, gate, and gave," (*Wikipedia*, The Mental Lexicon). Frequency effects suggest that words that are frequent in an individual's language are recognized faster than words that are infrequent (Forster and Chambers 1973).

On the other hand, traditional dictionaries tend to list all their words alphabetically, which suits the reader but not the writer or speaker. Furthermore, dictionaries tend to offer their users complete descriptions of their words and do not usually include any system for helping their potential users to search only for what they need in a particular usage situation. In addition, they typically give full lexicographical status to single-word lexemes, i.e. they only offer a complete lexicographical description of words such as "bank" but not of

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

"extended units of meaning", i.e. phraseological expressions such as "cry all the way to the bank", although recent research has shown that they are crucial in language processing and must be, therefore, lemmatized (Rundell 2018; Fuertes-Olivera 2019).

In sum, the mental lexicon starts from concepts, i.e. meanings, whereas the traditional dictionary approach adopts a topological view, which starts from forms and basically corresponds to off-line processing deliberately searching in a lexical resource, e.g. a dictionary. The next section offers a brief description of some attempts aiming at reconciling the working of our mental lexicon with the design and making of novel reference resources that might be more in line with how our brain works.

2. Computer science and lexicography

In the field of lexicography, Computer Science has mostly focused on the design of Natural Language Processing (NLP) tools which can facilitate information extraction, information retrieval, named-entity recognition, parsing, chunking, part-of-speech tagging, word sense disambiguation, and so on. The tools help in "the representation of linguistically expressible knowledge in language understanding, the use of knowledge for several sorts of commonsense reasoning, and knowledge accumulation" (Espinosa-Anke 2017: 4).

Projects such as *FrameNet* (Ruppenhofer et al. 2018), typically identified as *knowledge resources*, are examples of the joint work of computer scientists and lexicographers. Such resources are useful "because they can store meanings of words and phrases, relations of any kind (e.g. syntactic, syntagmatic, semantic or ontologic) holding among them, and also descriptions about entities or commonsense facts" (Espinosa-Anke 2017: 3).

Knowledge resources are classified into three categories: *structured knowledge resources*, e.g. dictionaries or knowledge bases, *unstructured knowledge resources*, e.g. statistical models derived from text corpora, or *semi-structured knowledge resources*, e.g. *Wikipedi*a (Hovy et al. 2013). In what follow, I will focus on structured knowledge resources, i.e. manually-crafted resources, because they are considered to represent knowledge at the highest level of quality. Hence, it can be hypothesized that if these are the highest quality resources, they will offer the best potentialities for overcoming some of the challenges NLP is facing, in particular those concerned with meaning extraction and elimination of contextual ambiguity.

Research, e.g. Clark et al. (2012); Hovy et al. (2013); Espinosa-Anke (2017), classifies structured knowledge resources into three types: (a) lexicographical resources; (b) lexical databases and thesauri; (c) knowledge bases. Lexicographical resources such as dictionaries are typically human readable. They consist of a list of words and their associated *senses* or *meanings,* usually accompanied by an array of lexicographical data, i.e. data that have been prepared by a human lexicographer with the aim of offering a (complete) descrip-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

tion of the meaning and usage of the *lemmas*, i.e. the entry words of dictionaries. For the purpose of this article, it is adequate to indicate that the data included in dictionaries are typically accessed and retrieved holistically, that such a characteristic does not suit the working of our brain and, consequently, this lexicographical method must be changed assuming that lexicographers aim at improving these knowledge resources by using NLP methodology in lexicography.

Lexical databases and thesauri, e.g. *Roget's Thesaurus* (see Kirkpatrick 1987) and *WordNet* (Miller 1995; Fellbaum 1998) represent *senses* by grouping them into sets of (cognitive) semantic relations, called *synsets* in *WordNet*). They employ an onomasiological ordering, i.e. the lemmas are arranged by their meanings and consequently users access them through their semantics. This means that word forms such as "bank", which can have several semantic relations, are found in several sections of the lexical databases and thesauri. For instance, the *Roget's Thesaurus* (Kirkpatrick, 1987) uses a topical distribution, e.g. "bank" is connected semantically to *height*, *support*, *obliquity*, *edge*, *laterality*, *land*, *store*, *lending* and *treasury*. Research (Zock and Bilac 2004; Zock and Schwab 2008) has shown that such resources, especially online lexical databases and thesauri, are especially useful for production purposes. In this paper, I will also add some comments to the above reflections and will hypothesize that the basics of these resources, i.e. semantic relations such as synonymy, antonymy, hyponymy and meronym are especially relevant for meaning extraction and disambiguation, thus increasing the potential use of NLP for lexicographical purposes. For instance, they can help the formation of network hubs in the human brains assuming that lexicographers can create some kind of pattern between the meaning and the semantic relations of a particular lemma (see 4, below).

Ré et al. (2014: 1) define a knowledge base (KB) as "a relational database together with inference rules, with information extracted from documents and structured sources". Espinosa-Anke (2017: 8) adds that in general, "we expect KBs to be graph-like data structures where each node represent an entity or concept (e.g. *Nintendo* or *hope*), and where edges between nodes may express *WordNet*-like semantic relations, but also ontologic relations such as *is-based-in* or *is-CEO-in*". They are especially useful for representing knowledge in a network form, especially in terminological knowledge bases (TKBs), which are in the forefront of research in this field. This can be achieved by giving definitions a reticular form, which consists of two stages, "first, to de-contextualise the terms and, second, to retain only the contexts that can be used to code knowledge in a network form" (Condamines 2018: 338).

Condamines (2018: 343) also looks at the future of TKBs and makes three observations, two of which are relevant for this paper. Firstly, TKBs are being substituted by *ontologies*, i.e. explicit specifications of conceptualizations (Gruber 1993; Roussey et al. 2018). Ontologies are being constructed by applying machine learning methods on very large corpora, usually the entire Web, "in order to spot new patterns and new triplets". She adds that with machine-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

learning methods, the main aim "is not to build a precise representation of the knowledge but, rather, to detect enough regularities to assume that some couples of terms have a constant and relevant relationship. In these cases, the most important application is to improve information retrieval". Secondly, the use of natural language processing tools, which are being used more and more, has emphasized that the so-called *term*, i.e. the linguistic representation of a concept, can be used as a key for entering specialized texts. Such a key may rely on patterns (a top-down method) or distributional contexts (a bottom-up method).

The above paragraphs have shown a mounting interest in several fields. In the next two sections, I will discuss possible ways of reinforcing this connection by defending a relationship between some lexicographical ideas and practice with Artificial Intelligence methods.

3. Lexicographical philosophy for designing and making online structured knowledge resources: The *Diccionarios Valladolid-UVa*

This section describes some of the main lexicographical ideas underlying the design and making of the *Diccionarios Valladolid-UVa*. For space reasons, I will only focus on very general principles that may influence the use of IA methods in lexicography (see Section 4). The first idea is explained in terms of the tenets of the *function theory of lexicography* (e.g. Bergenholtz and Tarp 2003; Tarp 2008; Fuertes-Olivera and Tarp 2014). Dictionaries are information tools dealing with "things", "facts" and "languages". The advent of online lexicography has reinforced this idea whose practical application is that *lexicographers do not need to design and make different dictionary types*, i.e. monolingual, bilingual, general, specialized, abridged, semi-abridged, learner's and so on. In the digital environment, lexicographers can, and in my opinion should, deal with all the words they can find, describe them in the most precise way, and adapt and store them in Dictionary Writing Systems (DWS) that facilitate different types of searches and retrieval. For practical purposes, this idea implies considering three specific decisions: (a) selection of the headword or lemma list; (b) selection of the empirical sources; (c) selection of the data types to be included in the DWS as well as its *grammar* i.e. specifications about the structure of the dictionary (Kilgarriff 2006), and *homepage* with specific search and retrieval systems.

Following current practice in lexicography, I consider the selection of the headword or lemma list to be an *ongoing process*, i.e. a process that is never finished. As such, lexicographers must decide on the method for selecting the initial lemma list and its continuous amplification. Since the advent of the *Cobuild Dictionary* (Sinclair 1987), lexicographers have mostly defended a corpus-based approach to headword selection, i.e. the words to be included must be *basically* extracted from corpora on the basis of their frequency and/or keyness. My proposal is different: the selection is a process that needs taking into consideration its inception and continuous development. Its initial stage aims at selecting the words that users *really* look up, as research has discovered that

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

many of the words lemmatized in existing dictionaries — some researchers claim that almost 80%; see Bergenholtz and Norddahl 2014 — have never been looked up (Trap-Jensen et al. 2014). The *Diccionarios Valladolid-UVa* have followed this methodology and initially selected two lists of single-word lemmas, one for English and one for Spanish. The initial headword lists of the *Diccionarios Valladolid-UVa* were selected at Ordbogen A/S headquarters, a Danish language technology company with whom we have been designing and making our lexicographical projects since 2014 (Fuertes-Olivera 2019).

The Danish company used big data analytics for around two months. The process comprised several stages and was based on an analysis of around one million daily searches in several dictionaries, e.g. an English–Spanish/Spanish–English dictionary, an English–German/German–English dictionary, an English monolingual dictionary, a Spanish monolingual dictionary, and so on. Around 80% of the searches could be matched, i.e. they could be interpreted with the aim of identifying the most popular dictionary articles in both languages. After two months of work with the logfiles of the searches — they amounted to more than 60 million logfiles — IT staff at Ordbogen A/S were able to produce the above-mentioned lists, each comprising around 20,000 single words. These are the words most searched for the period under analysis. The editor of the project systematized them and decided on their amplification, i.e. the process used for adding more lemmas to the initial lemma list. From now on, I will only refer to the Spanish list and the Spanish dictionary of the project.

Systematization means that all the members of the lists must be converted into a *unit of inclusion*, e.g. a lemma in traditional lexicography. Following standard practice, the editor *initially* converted the list into 16,678 single-word lemmas and these were included in the DWS in their canonical form, e.g. the infinitive of the verb, but adapted to an online process of searching (see Section 4, below). In January 2022, the project had completed the lexicographical description of around 10,000 of the initial single-word lemmas. This resulted in around 60,000 meanings or senses (around 6 meanings per lemma). This means that polysemous words are abundant and need some special treatment for making them adequate for disambiguating purposes (see Section 4 below).

The quantity of the meanings included offers some clues on the general philosophy of the project. For illustrative purposes, I will compare the lexicographical data of 25 single-word lemmas with their treatment in the *Diccionario de la Lengua Española* (RAE), which is the dictionary designed and made by the Royal Spanish Academy:

1. *ábaco* (abacus);
2. *abajo* (down); (downstairs);
3. *abalorio* (glass bead);
4. *abanderado* (standard-bearer), (champion), (linesman);
5. *abanderar* (register);
6. *abandonado* (deserted), (abandoned);

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

7. *abandonar* (leave), (abandon), (desert), (give up), (withdraw), (pull out), (resign), (retire), and so on;
8. *abanico* (fan), (range);
9. *abaratar* (reduce), (lower), (cut) and so on;
10. *abarcar* (cover), (cope with), (embrace), (circle), (take in);
11. *abastamiento* (provisions);
12. *abastecer* (supply);
13. *abastecimiento* (supply);
14. *abasto* (supply), (basic provisions);
15. *abatir* (shoot down), (bring down), (knock down), (pull down), (demolish), (fell), (cut down), (bow), (lower), and so on;
16. *abdicación* (abdication);
17. *abdomen* (abdomen);
18. *abecedario* (alphabet);
19. *abeja* (bee);
20. *abejorro* (bumblebee);
21. *aberración* (aberration);
22. *abertura* (opening), (hole), (slit);
23. *abeto* (fir);
24. *abiertamente* (openly);
25. *abierto* (open), (undone), (split), (openminded), and so on.

The comparison only aims at illustrating some of the key differences between the two lexicographical projects. For space reasons, I will only focus on differences that may be connected with the possible use of AI in lexicography (see Section 4). These lemmas have 153 meanings (around six meanings per lemma) in the *Diccionarios Valladolid-UVa* and 114 meanings (around four and a half meanings) in the *Diccionario de la Lengua Española* (RAE). This difference is relevant and will be explained below.

Amplification is also an *on-going process*. It is initially concerned with the words and expressions that are related with the lemmas of the initial lemma list. In the *Diccionarios Valladolid-UVa*, an expression or "extended unit of meaning" (Rundell 2018) is a linguistic unit formed by two or more orthographical words that expresses a concept and is used as a unit within a sentence. Such a unit is converted into an "extended-unit-of-meaning-lemma" and included in the lemma list if it is still in use, e.g. by being in around 5% of the Google minitexts used as sources (see below) and in four out of seven existing dictionaries that we also look up during the process of compilation: *Diccionario de la Lengua Española* (RAE); *Diccionario del Español Actual* (Seco et al. 2011); *Diccionario Español–Inglés* (Collins); *Diccionarios.com; Lexico Spanish* (Oxford); *SpanishDict*; and *WordReference (Spanish; Spanish–English)*.

The lemmatization of expressions is based on the tenets of *semantic network theory* (see Section 1, above). This theory affirms that humans *mostly* use meaning networks in their daily linguistic interactions. Hence, all the expressions that can be identified during the process of description of the initial lemmas are

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

lemmatized in the *Diccionarios Valladolid-UVa.* For instance, we have lemmatized *pájaro bobo* (penguin or tropical bird), which was found when I was describing the adjective *bobo* (stupid). Spanish dictionaries typically include *pájaro bobo* as an expression at the end of the dictionary articles for *pájaro* (bird) or as a meaning of the adjective *bobo*. The lemmatization of *pájaro bobo* facilitates searching and retrieval, as will be explained below (see Section 4).

One or more of the 10,000 single-word lemmas already completed are present in around 30,000 "extended-unit-of-meaning-lemmas" (i.e. each single-word lemma is in around 3 extended-units-of-meaning-lemmas). Their lexicographical description has amounted to around 40,000 more meanings, (around 1,25 meaning per expression). For instance, the abovementioned 25 lemmas are part in one or more 83 new extended-unit-of meaning-lemmas (e.g. *en abierto*; *el que mucho abarca poco aprieta*) also included in the *Diccionarios Valladolid-UVa.* The *Diccionario de la Lengua Española* (RAE) only lemmatizes single-word lemmas, and consequently there are no extended-unit-of meaning-lemmas in this dictionary, which nests them at the bottom of a dictionary article, usually accompanied with definition and, sometimes, some grammar information. Of the 25 words under analysis, the *Diccionario de la Lengua Española* includes 49 expressions (e.g. *echar abajo*), i.e. almost half of those included in the *Diccionarios Valladolid-UVa*. As before, I will comment on the different numbers between both dictionaries below.

The above figures illustrate an interesting difference between the "single-word-lemma" and the "extended-unit-of-meaning-lemma": extended units of meaning tend to be monosemic entities, and this tendency increases when the number of words forming part of the expression also increases. In other words, the use of extended-unit-of-meaning-lemmas tend to eliminate polysemy and, hence, meaning ambiguity. It seems evident that the larger the number of extended units of meaning included in the structured knowledge resource the less meaning ambiguity in it.

By "related words" I mean the words that stem from the initial single-word-lemmas due to grammar rules. In Spanish, these *basically* affect some nouns, adjectives, adverbs, and verbs. For instance, *abanderado* is a male noun and its related word is *abanderada* (female noun). In traditional Spanish dictionaries such as the *Diccionario de la Lengua Española*, this process of amplification only exists for lemmatizing some manner adverbs, i.e. they are formed by adding *-mente* to the base of an adjective, e.g. *abiertamente*. For the rest of related words, Spanish dictionaries use constructs such as *abanderado, ra* that do not exist in real linguistic interactions (Fuertes-Olivera and Tarp 2022) or do not lemmatize them at all. For instance, the related words of the verbs *abanderar, abandonar, abaratar, abastecer* and *abatir* (they are *abanderarse, abandonarse, abaratarse, abastecerse, and abatirse;* they are reflexive or pronominal verbs) and the related word of the adjective *abierto* (i.e. a noun, which is nominalized by putting an article before it, e.g. *un abierto, el abierto, unos abiertos, los abiertos*) are not lemmatized in the *Diccionario de la Lengua Española.*

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

However, in the *Diccionarios Valladolid-UVa,* this process of amplification is totally active and works with nouns, adjectives, adverbs and verbs. For instance, of the abovementioned 25 single-word lemmas we have included 11 more single-word lemmas: *abanderado* (adjective), *abanderada* (noun), *abastos* (plural noun), *abejorra* (noun), *abierto* (noun), *abierta* (noun), *abanderarse, abandonarse, abaratarse, abastecerse,* and *abatirse* (reflexive or pronominal verbs). These 11 lemmas contain 35 meanings (around 3 meanings per lemma).

The application of this amplification policy means that the *Diccionarios Valladolid-UVa* not only contains a much larger stock of lemmas and meanings but also that it is much more useful for NLP as all relevant word strings, no matter how many words they contain, are lemmas and are described in full. In other words, amplification also offers some clues on another lexicographical idea that underlies the design and making of the *Diccionarios Valladolid-UVa: the lexicographical process must be as complete and precise as possible*. The rationale for such a philosophy is twofold: (a) it offers a better description of the language and (b) it facilitates searching and retrieving. Hence, it might be better prepared for using NLP tools, as I will show below (see Section 4). This idea, which is also the philosophy of semi-structured knowledge resources such as *Wikipedia*, eliminates the traditional conception of dictionaries as finished products, subjected to the publication of different editions, and limited, in one way or another, to a particular topic, variety, user's needs, situation, and so on. In sum, the making of dictionaries is a never-ending process that must *constantly* calibrate amplification and the finding of adequate empirical sources.

In today's world, I think that the Internet is the best empirical source for lexicographical work. In other words, the internet is a *lexicographical corpus*, defined by Fuertes-Olivera (2012: 51) as "any collection of texts where lexicographers can find inspiration for completing the dictionary structures they need when making a dictionary". Going a step further, I add that such a lexicographical corpus not only is adequate for making dictionaries but also for any knowledge resources that can be imagined. Consequently, time is ripe for using the Internet to understand the meaning and usage of a particular word or expression in a way that reduces, even eliminates if possible, the "creation and maintenance effort". In this project, we use "Google minitexts", i.e. the two to three lines Google retrieves when making a particular search, for an initial analysis of the meaning and usage of lemmas (Tarp and Fuertes-Olivera 2016). If we find relevant information in them, we click on the homepage and analyze the text or part of it. With this method, it takes around 15 minutes for finding out relevant meanings and linguistic characteristics of most lemmas, especially of extended-unit-of-meaning-lemmas and single-word-lemmas that have up to 7 different meanings, i.e. around 85% of all the lemmas described so far. For instance, only 6 out of 36 single-word lemmas (i.e. the 25 initial lemmas and the 11 created by amplification) contain more than 7 meanings (16%).

The "Google-minitext" method does not properly work with lemmas that have a lot of meanings, e.g. the adjective *abierto* (opened) has 22 meanings and the verb *hacer* (make, do) has 55 meanings. In such a situation, which currently amounts to around 15% of the lemmas finished so far, we use a "guided search method". It consists in searching in the Internet if the meaning(s) previously found in the consulted dictionaries can be confirmed, i.e. are still used. This method implies the construction of "search strings" formed by the lemma (in quotation marks) plus two or three keywords extracted from the definitions found in the consulted dictionaries. For instance, the search string "abierto" + billete (ticket) + vuelta (return) retrieves more than 3 million hits. Just in the first twenty we can easily confirm the meaning of *abierto* referring to a ticket whose return date is not fixed yet. Such a meaning is a figurative or metaphorical extension of its base meaning. These results were obtained with several different browsers, which indicated that this meaning of *abierto* is still in daily use and that the results are not affected by the search history or cookies of the browser.

The "guide search method" explains the third main lexicographical idea behind the design and making of this dictionary. It can be summarized by saying that *all existing dictionaries, encyclopedias, glossaries as well as grammar books, usage books and the like, should be consulted for inspiration, but not for copying and pasting*.

Finally, the selection of the data types or lexicographical data to be included is basically a cooperative process. *Cooperation is, then, another important idea underlying the Diccionarios Valladolid-UVa.* Cooperation implies the joint work of lexicographers, IT people and experts, e.g. in web design tools. All of them must jointly decide the number of data types they need for describing each lemma and the characteristics of the DWS which must be used. Existing Spanish dictionaries usually use between three and six different data types. The *Diccionario de la Lengua Española*, for instance, always has etymology, abbreviations for indicating part of speech of the lemma and a definition, usually a short one or a synonym. In addition, for many lemmas, it also has expressions (if there are) and one or two clause or sentence examples. For instance, for *abejorro*, the dictionary includes its origin (it comes from "De abeja"), three meanings, two of them described with a short sentence and one with a synonym, and the abbreviation "m", for "masculine" (Example 1):

> *abejorro*
> *De abeja*
>
> **1.** m. Insecto himenóptero, semejante a la abeja pero más grande, de cuerpo velludo, generalmente negro y con bandas amarillas, que produce un zumbido al volar y vive en enjambres poco numerosos.
>
> **2.** m. **escarabajo sanjuanero.**
>
> **3.** m. Persona de conversación pesada y molesta.

Example (1): *abejorro* in the *Diccionario de la Lengua Española* (RAE)

In the *Diccionarios Valladolid-Uva* there are up to 25 possible data types for each Spanish lemma, being the typical one described with fourteen data types: part of speech; inflections, meaning, antonym, synonym, related words, phrase sentence, example sentence, diastratic and/or diaphasic mark (for lemma, antonym, synonym and related words) and diatopic mark for meaning. In addition, some lemmas also have ten more data types: (a) a photo, e.g. for animals, plants and objects; (b) alternative inflections and orthography; (c) part of the conjugation of a verb, (d) proscriptive notes, which are used for recommending between options, e.g. orthographic options, (e) link to a conjugation table, e.g. a verb; (f) link to an external text, e.g. *Wikipedia*; (g) grammar note, (h) usage note, (i) phrase note, which explains the syntactic pattern of an extended-unit-of-meaning-lemma, and (j) synonymy note, which explains possible specific uses of a synonym, e.g. it is only used in Argentina. Example (2) shows *abejorro*, as it is now in the DWS of the *Diccionarios Valladolid-UVa*:

abejorro
noun
<un abejorro, el abejorro, unos abejorros, los abejorros>

meanings

1. insecto parecido a la abeja perteneciente a la familia de los ápidos; tiene el cuerpo más gordo y puede llegar a los 3 centímetros de largo; tiene el cuerpo cubierto de vello oscuro y una trompa muy desarrollada, que emite un zumbido intenso al volar; vive en enjambres poco numerosos debajo del musgo o de las piedras; en el enjambre solo hay una hembra, que es la que fecunda; se alimenta del polen y néctar de las flores

Synonyms for this meaning:

— *abejarrón*
— *abejón*
— *Bombus* <formal>

Phrase sentences for this meaning

— cámaras capaces de captar el vuelo del abejorro con un nivel de detalle espectacular
— diferencia entre el abejorro y la abeja carpintera
— las picaduras de abejorros
— los abejorros, que son bien gorditos y peludos
— si una flor apetitosa está solitaria o concurrida por otros abejorros

Example sentences for this meaning

— El abejorro de tierra o Bombus terrestris, es uno de los tipos de abejorros más empleados en la agricultura intensiva, debido a su alto nivel de polinización.
— El pelo grueso actúa como aislante, manteniendo al abejorro a una temperatura adecuada.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

Related words for this meaning

- abeja
- abejorro carpintero
- abejorro común
- abejorro cuco

Photo for this meaning:

2. insecto de la familia de los escarabajos; tiene el cuerpo de color marrón oscuro y élitros pardos; puede llegar a los 3 centímetros de largo; roe las hojas de las plantas cuando es adulto y sus raíces en estado de larva; emite un zumbido intenso al volar

Synonyms for this meaning

escarabajo sanjuanero
Melolontha melolonta <formal>

Phrase sentences for this meaning

- el caparazón pardo del abejorro
- los abejorros que decidieron abandonar los árboles en los que se encontraban para invadir prados, jardines y herbazales
- los élitros del abejorro
- los huevos del abejorro bajo los pastos o el césped

Example sentences for this meaning

- Las antenas de estos abejorros se caracterizan por poseer laminillas terminales, capaces de plegarse como varillas de un abanico y formar una maza.

Related words for this meaning:

- abejorro carpintero
- abejorro común
- abejorro cuco

Photo for this meaning

3. en sentido figurado, hombre cuya conversación resulta aburrida, pesada y causa molestia <informal>

Phrase sentences for this meaning

- al abejorro que no aguanta nadie
- mejor ser abejorro que mosca cojonera
- que es un abejorro y un pesado
- un abejorro dando la tabarra

Example sentences for this meaning

- El vecino es un abejorro, como te vea te enrolla hablando de cosas que no te importan.

Related words for this meaning

- abejorra

4. en sentido figurado, persona (hombre o mujer) cuya conversación resulta aburrida, pesada y causa molestia <informal>

Phrase sentences for this meaning:

- no quedar con esos abejorros, sus conversaciones son demasiado cargantes
- se largó en cuanto empezamos a hablar de abejorros

Example sentence for this meaning

- Son como los abejorros: no callan ni debajo del agua.

Example (2): *abejorro* in the DWS of the *Diccionarios Valladolid-UVa*

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

Comparing examples (1) and (2) offer several conclusions that are relevant for the use of AI in lexicography (see Section 4):

— In the *Diccionarios Valladolid UVa* there are around 400 words for describing *abejorro* lexicographically, whereas the *Diccionario de la Lengua Española* uses fewer than 50 words, i.e. the *Diccionarios Valladolid-UVa* uses almost 12 times more words for describing *abejorro* than the *Diccionario de la Lengua Española*.
— The *Diccionarios Valladolid-UVa* also uses photos for describing physical meanings, e.g. animals in the lemma *abejorro*.
— The *Diccionarios Valladolid-UVa* does not use abbreviations.
— The *Diccionarios Valladolid-UVa* offers a very precise description of meanings and forms, e.g. each definition of each lemma goes with inflections, part of speech, semantic relations, varieties and so on. In other words, most of the lexicographical data are attached to each specific definition.
— Each meaning is independently described.

Using such a large quantity of lexicographical data for describing each lemma influences the design and characteristics of the DWS used for compiling the knowledge resource. In the *Diccionarios Valladolid-UVa*, we are working with an in-house DWS designed by the joint work of IT people at Ordbogen A/S and the editor of the project (Fuertes-Olivera 2019). The DWS of the *Spanish* part of the *Diccionarios Valladolid-UVa* has 30 slots: 25 of these contain the lexicographical data previously commented (see example 2, above). In addition, there are two slots for ordering lemmas and meanings, one slot for internal communication, one slot for administrative purposes, e.g. knowing who has been working in the description of the lemma, and one slot for internal searching, e.g. for searching for "figurative meanings".

4. Using Artificial Intelligence in lexicography

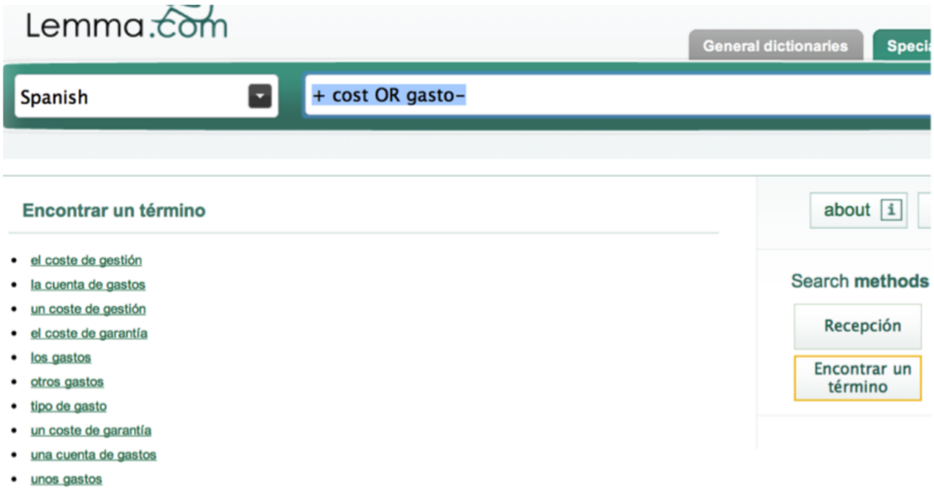
Artificial Intelligence is a wide-ranging branch of Computer Science concerned with building smart machines capable of performing tasks that typically require human intelligence. Russel and Norvig (2010: viii) claim that AI is "the study of agents that receive percepts from the environments and perform actions". In the last five years, there have been several proposals for using AI in lexicography. Plakhotniuk (2018), for example, claims that the collaboration of AI and e-lexicography basically concerns two aspects: (a) improving the data extracted from existing dictionaries and (b) eliminating constraints, e.g. editorial constraints, for digitalizing printed sources. In this section, I will focus on the first aspect and will maintain that the improvement needs not only more lexicographical data (e.g. 12 times more words for describing the word *abejorro* in the *Diccionarios Valladolid-UVa* than in *Diccionario de la Lengua Española,* see examples 1 and 2, above), but also better created, systematized and ordered,

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

e.g. by trying to reproduce the way the mental lexicon works (see Section 1, above). This means the adoption of three main methodological approaches to dictionary making.

Firstly, our data are prepared for adopting the "closure criteria", which mean that "everything wtwhat [sic] occurs on the right side of a dictionary must be listed on the left side of the same dictionary" (Dembitz et al. 2005: 1). In other words, it is easy to create a list of word types extracted basically from definitions and phrase and example clauses. Furthermore, such a huge number of lexicographical data for each meaning of each lemma is in line with the so-called "middle ground", i.e. working with big data and good data (Hovy et al. 2013), and allow the creation of "multiple alignment, i.e. treating words in context and comparing their contextual usage metrically" (Dembitz et al. 2005: 2).

Secondly, the Spanish dictionary must be equipped with a search system which will allow users to retrieve *a la carte*, i.e. different data in different situations and for different users. The system will offer users the search button ENCONTRAR UN TÉRMINO (FIND A TERM). This button will allow "users who are uncertain of the exact form of the term to be searched for, or who want to explore the data of a particular term field, to generate their own searches and search strategies by using Boolean operators" (Fuertes-Olivera and Leroyer 2014). For instance, using the search string "+ cost OR gasto-" (Figure 1), users retrieve a series of texts, all of which are clickable and adequate for retrieving the dictionary article in which such texts are, e.g. being part of the phrase or example sentences.



+ cost OR gasto-

Encontrar un término

- el coste de gestión
- la cuenta de gastos
- un coste de gestión
- el coste de garantía
- los gastos
- otros gastos
- tipo de gasto
- un coste de garantía
- una cuenta de gastos
- unos gastos

Figure 1: Retrieving texts when searching + cost OR gasto-

Thirdly, the data must be formalized for showing "consistency of inner semantic relations" (Plakhotniuk 2018: 78). These are adequate for creating patterns for machine learning, whose aim, according to Condamines (2018: 343) "is not to build a precise representation of the knowledge but, rather, to detect enough regularities" which will allow us to find constant and relevant relationships, e.g. for eliminating meaning ambiguity and reproducing the processes associated with how words are activated, stored, processed and retrieved by speakers, who never use them in isolation but in contexts. For this, the data stored in the DWS of the *Diccionarios Valadolid-UVA* contains the following:

1. Inflections for nouns and adjectives and conjugations for verbs, both for single-word lemmas and unit-of-meaning-lemmas. These will allow users to retrieve data in any form. For instance, the search engine of the *Diccionario de la Lengua Española* does not work with search strings such as "habríamos querido" (a perfect conditional form of *querer*). This search string will be found in the *Diccionarios Valladolid-UVa*. In other words, users do not need to know the canonical form of the lemma for searching.

2. Very precise definitions (Fuertes-Olivera and Esandi-Baztan 2020); they group semantically similar senses, thus allowing the search engine to search for strings such as that of Figure 1, e.g. the search string "*serpiente* (serpent) + *venenosa* (poisonous) + *americanismo* (Americanism*)*" will retrieve all the poisonous serpents that are living in South America, whereas the string "*serpiente* (serpent) - *venenosa* (poisonous) + *americanismo* (Americanism*)*" will retrieve those that are not poisonous. Furthermore, all definitions are self-sufficient, i.e. neither recursive definitions nor synonyms are used for defining each meaning of each lemma. For instance, the second definition of *abejorro* in examples (1) and (2) refers to the same animal; in the *Diccionario de la Lengua Española,* users are given a synonym and linked to a different dictionary article, whereas in the *Diccionarios Valladolid-UVa* the meaning is precise and users have all what they need in the dictionary article. In a similar vein, every time a word related with the lemma is used in the definition, this word is also defined, typically after formulae such as "que es" (that is) or "es decir" (i.e.). For instance, the lemma *avicultura* (aviculture), which is used in the definition of the adjective *avícola* (poultry), is also defined in *avícola* after "que es" (that is) (example 3):

   avícola
   referido a o relacionado con la avicultura, que es una técnica, actividad, etc. que se ocupa de la cría de aves y el aprovechamiento de sus productos

Example (3): Definition of *avicultura* in the entry for *avícola*

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

3. semantic networks between definitions and semantic relations, especially with synonyms and, less frequently, antonyms and related words (see example 2). This means that definitions *explicitly* differentiate between similar meanings, e.g. between literal and figurative meanings, animal, things or human beings functioning as actors, and so on. Each of these meanings always goes with up to three synonyms and/or antonyms. The synonyms are replaceable, e.g. in all the phrase and example clauses used in the dictionary article. For instance, in the DWS of the *Diccionarios Valladolid-UVa*, there are three meanings for the Spanish verb *aullar*. In the *Diccionario de la Lengua Española*, this verb only has one meaning and its description is recursive "dar aullidos". Example (4) shows the three meanings and its semantic relations in the *Diccionarios Valladolid-UVa*:

    1. emitir un animal sonidos agudos, tristes y prolongados (animals emit high-pitched, sad and long sounds)

       synonym:
       a. gemir (whine)
       b. mugir (moo, bellow)

    2. emitir una persona sonidos agudos, tristes y prolongados (persons emit shrill, sad and long sounds)

       synonym:
       a. gritar (shout)
       b. vociferar (yell)

    3. en sentido figurado, emitir una cosa sonidos agudos, tristes y prolongados (figuratively, something emits high, intense and long sounds)

       synonym:

       a. bramar (roar)
       b. ulular (howl)

Example (4): Creation of a semantic network between definitions and semantic relations

In example (4), there are three meanings: two literal (the default criterion in Spanish dictionaries) and one figurative, being the actor of the process the main difference between the three meanings: they are respectively an animal, a person, and an object or abstract actor such as the wind. This difference is reinforced with the selection of synonyms *gemir* (whine) and *mugir* (moo and bellow) for animals, *gritar* (shout) and *vociferar* (yell) for human beings, and *bramar* (roar) and *ulular* (howl) for objects or abstract things such as the wind. For instance, *el viento aulla* (the wind howls) is correct Spanish, whereas *el viento vocifera* (the wind yells) is nonsensical and never used (for instance, I did not find any hit of "viento vocifera" in Google Books. Spanish 2019 (Google Books Ngram Viewer). In sum, these semantic networks are useful for meaning disambiguation (and also for creating different dictionary types, e.g. a dictionary of synonyms containing the meanings and its antonyms and synonyms).

4. similarities, e.g. those formed by the gender of a noun and its reference to a man, woman or person, e.g. the meaning 3 of *abejorro* (example 2, above) starts with *hombre* (man), whereas the meaning 4 does it with *persona* (person), i.e. one refers to a male (and it also has its counterpart *abejorra* (woman); see the discussion on related words above), whereas the other meaning is generic and refers to human beings in general (some other generic also include institutions, organizations, companies, countries, etc. (Fuertes-Olivera and Tarp 2022).
5. phrase and sentence examples (see example 2, above) for each meaning of each lemma. There are *always* from three to six of them for all content words and expressions and between one and two for function words and expressions. They illustrate grammar, usages, e.g. indicating contractions ("del"), singular and plural forms ("abejorro" and "abejorros"), and meanings, e.g. the phrase and example sentences of the meaning one of example (2) confirm the four main attributes of the meaning of this insect: "the insect lives among flowers", "the insect is fatty", "it is dark brown with yellow lines", and "these insects are used in intensive farming".
6. photos for all material beings, objects and things, e.g. animals and instruments. These are not only very useful for describing their meanings in a perfect way but also for differentiating material meanings from abstract ones, most of which are figurative. For instance, in *autopista* (motorway), the DWS contains two meanings: the literal one ("a highway designed for fast traffic, with controlled entrance and exit and so on") goes with a photo of a motorway, whereas the figurative meaning ("an easy way to achieve something without much work") goes without photo but with the indication that this meaning is figurative.

In sum, the data types are all perfectly formalized, standardized and adequate for (a) proposing an interdisciplinary approach to dictionary making, one which meets the needs "of creators of intellectual information systems and dictionaries for humans and machine-based users" (Plakhotniuk 2018: 78), and (b) training the system and hence allowing AI methods reproduce our mental lexicon.

5. Conclusion

This paper has analyzed the possibility of making dictionaries that take into consideration the mental lexicon, i.e. words do not work in isolation; instead,

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

they are dynamic constructs that are activated, stored, processed and retrieved gradually. This possibility demands the design and making of dictionaries that are very different from the static structured knowledge resources that now exist. In my view, these dictionaries of the future demand new lexicographical thinking, especially one that analyzes the possibility of using AI for solving complex problems such as disambiguating meanings and allowing users search in speaking situations. My proposal, which is illustrated with the *Diccionarios Valladolid-UVa*, is based on several general ideas and specific lexicographical practice, all of which view AI as an adequate methodology for designing and making the dictionary of the future:

— *Lexicographers do not need to design and make different dictionary types*, i.e. existing dictionary typologies do not suit AI as humans do not segment their brains into the categories typically used in today's lexicographical work.

— *The lexicographical process must be as complete and precise as possible*, e.g. with the inclusion of photos for describing material objects, inflections and conjugated forms, and so on. This favors searching and retrieving assuming the "closure criteria" and searching and retrieving *a la carte*, i.e. many different possibilities of searching and retrieving.

— *All existing dictionaries, encyclopedias, glossaries as well as grammar books, usage books and the like, should be consulted for inspiration, but not for copying and pasting*, e.g. using the Web as a lexicographical corpus facilitates the process of compilation in around 85% of the lemmas, offers real language use and allows lexicographers to equip their meaning descriptions with phrase and example clauses that help disambiguate meaning and create multiple alignments. Existing resources can help complete description by facilitating the use of "guided searches", which must be employed in very specific situations, e.g. when we have to describe highly polysemous lemmas.

— *Cooperation is a must and no adequate structured knowledge resource can be implemented without the joint work of lexicographers and experts, e.g. IT and web experts*, i.e. dictionaries are no longer the realm of linguists and their making is much more than describing the grammar and meaning of isolated words.

— *Words should also be considered processes that are always on the move*, e.g. as they can have different forms and meaning, we need systems that allow users retrieve them in different usages, forms, formats and purposes.

— *The Dictionary Writing System must be an in-house system created for specific lexicographical projects and equipped for favoring the creation of patterns,* e.g. those formed with semantic networks, *and the working of words in contexts*, e.g. a large number of phrase and example clauses adequate for AI methodology.

— *All the lexicographical work must be formalized and standardized*, e.g. adequate for human and machine-based users.

Acknowledgments

Thanks are due to the editor of the issue and to two anonymous reviewers, whose comments and suggestions have greatly improved the first draft of the article.

References

Aitchison, J. 2003. *Words in the Mind: An Introduction to the Mental Lexicon.* Oxford: Blackwell.

Andrews, S. 1989. Frequency and Neighborhood Effects on Lexical Access: Activation or Search? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15(5): 802-814. doi:10.1037/0278-7393.15.5.802.

Bergenholtz, H. and B. Norddahl. 2014. The Ideal Number of Lemmas in an Ideal Accounting Dictionary. *Hermes, Journal of Language and Communication in Business* 53: 143-150. doi:10.7146/hjlcb.v27i53.20988.

Bergenholtz, H. and S. Tarp. 2003. Two Opposing Theories: On H.E. Wiegand's Recent Discovery of Lexicographic Functions. *Hermes, Journal of Linguistics* 31: 171-196. 10.7146/hjlcb.v16i31.25743.

Clark, M., Y. Kim, U. Kruschwitz, D. Song, D. Albakour, S. Dignum, U. Cerviño Beresi, M. Fasli and A. de Roeck. 2012. Automatically Structuring Domain Knowledge from Text: An Overview of Current Research. *Information Processing and Management* 48(3): 552-568. https://doi.org/10.1016/j.ipm.2011.07.002.

Collins. *Diccionario Español–Inglés.* https://www.collinsdictionary.com/es/diccionario/ingles-espanol [Accessed May 5, 2022].

Condamines, A. 2018. Terminological Knowledge Bases. Pedro A. Fuertes-Olivera (Ed.). 2018. *The Routledge Handbook of Lexicography*: 335-349. London: Routledge.

Dembitz, Š., Lj. Jojić and J. Pavlek. 2005. Artificial Intelligence in Lexicography: Croatian Encyclopaedic Dictionary Example. *The 16th International DAAAM Symposium: Intelligent Manufacturing & Automation: Focus on Young Researchers and Scientists, University of Rijeka, 19–22 October 2005, Opatija, Croatia: 1-2.* https://bib.irb.hr/datoteka/231539.Dembitz-jojic-pavlek.pdf [Accessed May 5, 2022].

Diccionarios.com. *Diccionario Gratuito: Español.* https://www.diccionarios.com/ [Accessed May 5, 2022].

Emmorey, K.D. and V.A. Fromkin. 1988. The Mental Lexicon. Newmeyer, F.J. (Ed.). 1988. *Linguistics: The Cambridge Survey*: 124-149. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621062.006.

Espinosa-Anke, L. 2017. *Knowledge Acquisition in the Information Age: The Interplay between Lexicography and Natural Language Processing*. [Ph.D. Dissertation]. Barcelona: Universitat Pompeu Fabra. http://hdl.handle.net/10803/404985 [May 5, 2022].

Fellbaum, C. (Ed.). 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Forster, K.I and S.M. Chambers. 1973. Lexical Access and Naming Time. *Journal of Verbal Learning and Verbal Behavior* 12(6): 627-635. https://doi.org/10.1016/S0022-5371(73)80042-8.

Fuertes-Olivera, Pedro A. 2012. Lexicography and the Internet as a (Re-)source. *Lexicographica* 28: 49-70. https://doi.org/10.1515/lexi.2012-0005.

Fuertes-Olivera, Pedro A. 2019. Designing and Making Commercially Driven Integrated Dictionary Portals: The *Diccionarios Valladolid-UVa*. *Lexicography* 6(1): 21-41. https://doi.org/10.1007/s40607-019-00056-8.

Fuertes-Olivera, Pedro A. and M.A. Esandi-Baztan. 2020. Integrating Terminological Resources in Dictionary Portals: The Case of the *Diccionarios Valladolid-UVa*. *Lexikos* 30: 90-110. https://doi.org/10.5788/30-1-1598.

Fuertes-Olivera, Pedro A. and P. Leroyer. 2014. User-generated Exploratory Search Routes: ENCONTRAR UN TÉRMINO in the Accounting Dictionaries. Ruppenhofer, J. and G. Faass (Eds.). 2014. *Proceedings of the 12th Edition of the Konvens Conference*: 86-95. Hildesheim: University of Hildesheim. http://opus.bsz-bw.de/ubhi/volltexte/2014/289/pdf/konvens_proceedings.pdf.

Fuertes-Olivera, Pedro A. and S. Tarp. 2014. *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110349023.

Fuertes-Olivera, Pedro A. and S. Tarp. 2020. A Window to the Future: Proposal for a Lexicographically-assisted Writing Assistant. *Lexicographica* 36: 257-286. https://doi.org/10.1515/lex-2020-0014.

Fuertes-Olivera, Pedro A. and S. Tarp. 2022. Critical Lexicography at Work: Reflections and Proposals for Eliminating the Gender Bias in General Dictionaries of Spanish. *Lexikos* 32(2): 105-132.

Fuertes-Olivera, Pedro A, S. Tarp and P. Sepstrup. 2018. New Insights in the Design and Compilation of Digital Bilingual Lexicographical Products: The Case of the *Diccionarios Valladolid-UVa*. *Lexikos* 28: 152-176. https://doi.org/10.5788/28-1-1460.

Google Books Ngram Viewer. https://books.google.com/ngrams [Accessed: May 5, 2022].

Gruber, T.R. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5(2): 199-220. https://doi.org/10.1006/KNAC.1993.1008.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London/New York: Routledge. https://doi.org/10.4324/9780203327630.

Hovy, E., R. Navigli and S. P. Ponzetto. 2013. Collaboratively Built Semi-structured Content and Artificial Intelligence: The Story so Far. *Artificial Intelligence* 194: 2-27. https://doi.org/10.1016/j.artint.2012.10.002.

Indefrey, P. and Wilhem J.M. Levelt. 2004. The Spatial and Temporal Signatures of Word Production Components. *Cognition* 92(1–2): 101-144. 10.1016/j.cognition.2002.06.001.

Jackendoff, R. 2002. *Foundations of Language. Brain, Meaning, Grammar, Evolution.* Oxford: Oxford University Press.

Kilgarriff, A. 2006. Word from the Chair. G.-M. de Schryver (Ed.). 2006. *DWS. Proceedings of the Fourth International Workshop on Dictionary Writing Systems, 5 September 2006, Turin, Italy (Pre-EURALEX 2006):* 7. Pretoria: (SF)^2. https://tshwanedje.com/publications/dws2006.pdf.

Kirkpatrick, B. (Ed.). 1987. *The Authorized Roget's Thesaurus of English Words and Phrases.* London: Penguin.

Marslen-Wilson, W.D. 1987. Functional Parallelism in Spoken Word-recognition. *Cognition* 25(1–2): 71-102. https://doi.org/10.1016/0010-0277(87)90005-9.

Miller, G.A. 1995. *WordNet*: A Lexical Database for English. *Communications of the ACM* 38(11): 39-41. https://doi.org/10.1145/219717.219748.

Oxford Lexico Spanish. https://www.lexico.com/ [Accessed: May 5, 2022].

Plakhotniuk, Ye. 2018. Lexicography: From Art to Science (Paradigmatic Perspective). *Studia Philologica* 11: 74-80. http://nbuv.gov.ua/UJRN/stfil_2018_11_14.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1712 (Article)

Ré, C., A.A. Sadeghian, Z. Shan, J. Shin, F. Wang, S. Wu and Ce Zhang. 2014. Feature Engineering for Knowledge Base Construction. *arXiv* preprint *arXiv*:1407.6439. https://arxiv.org/pdf/1407.6439.pdf.

RAE (Real Academia Española). *Diccionario de la Lengua Española*. [Accessed: May 7, 2022. https://www.rae.es/.

Roussey, C., N. Hernandez and H. Zargayouna. 2018. Domain Ontologies. Pedro A. Fuertes-Olivera (Ed.). 2018. *The Routledge Handbook of Lexicography*: 217-234. London: Routledge.

Rundell, M. 2018. Searching for Extended Units of Meaning — and What To Do When You Find Them. *Lexicography* 5: 5-21. https://doi.org/10.1007/s40607-018-0042-1.

Ruppenhofer, J., H.C. Boas and C.F. Baker. 2018. FrameNet. Fuertes-Olivera, Pedro A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 383-398. London: Routledge.

Russel, Stuart J. and P. Norvig. 2010. *Artificial Intelligence. A Modern Approach.* 3rd edition. Boston: Prentice Hall.

Seco, M., O. Andrés and G. Ramos. 2011. *Diccionario del español actual*. 2nd edition. Madrid: Aguilar.

Sinclair, J. (Ed.). 1987. *Collins COBUILD English Language Dictionar*y. London: Collins ELT.

SpanishDict. https://www.spanishdict.com/. [Accessed: May 5, 2022].

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-knowledge.* Tübingen: Niemeyer. https://doi.org/10.1515/9783484970434.

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

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

Traxler, M.J. 2012. *Introduction to Psycholinguistics. Understanding Language Science.* Oxford: Wiley-Blackwell.

***Wikipedia.* The Mental Lexicon.** https://en.wikipedia.org/wiki/Mental_lexicon [Accessed: January 5, 2022].

***WordNet:* An Electronic Database.** https://wordnet.princeton.edu/ [Accessed: January 22, 2022].

WordReference. https://www.wordreference.com/ [Accessed: May 5, 2022].

Zock, M. and S. Bilac. 2004. Word Lookup on the Basis of Associations: from an Idea to a Roadmap. *ElectricDict '04: Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries. Association for Computational Linguistics, Geneva, Switzerland, 29 August 2004:* 29-35. https://dl.acm.org/doi/10.5555/1610042.1610048.

Zock, M. and D. Schwab. 2008. Lexical Access Based on Underspecified Input. *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008):* 9-17. Manchester, UK: Coling 2008 Organizing Committee. https://www.aclweb.org/anthology/W08-1902.pdf.