http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

Structuring a Collection of Lexicographic Data for Different User and Usage Situations

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: According to Fuertes-Olivera and Tarp (2020), lexicography is being currently shaped by three related tendencies: (a) the increasing use of disruptive technologies; (b) the necessity of finding new business models that can finance new lexicographic projects; and (c) the existence of growing competition from other information sources, e.g. Google. These trends have a particular influence on specialized dictionaries, defined here as tools that cover areas outside general cultural knowledge and its corresponding Language for General Purpose (LGP). This article adopts the view that lexicographers can deal with the abovementioned tendencies by preparing structured collections of lexicographic data with details that can be easily converted into information and stored in DWSs that allow multiple combinations and possible retrievals. This option is suitable for the tendencies, as it uses adequate technologies, e.g. ways of profiling or individualizing searches, defends a new business model based on the creation of lexicographic data that can feed many different tools, e.g. *Write Assistant* (Fuertes-Olivera and Tarp 2020), and offers better and more precise information than that of Google and other information tools.

Keywords: LEXICOGRAPHIC DATA, TECHNOLOGY, SPECIALIZED LANGUAGE, SPECIALIZED LEXICOGRAPHY, E-LEXICOGRAPHY

Opsomming: Die strukturering van 'n versameling leksikografiese data vir verskillende gebruikers- en gebruiksituasies. Volgens Fuertes-Olivera en Tarp (2020) word die leksikografie tans gevorm deur drie verwante tendense: (a) die toenemende gebruik van ontwrigtende tegnologieë; (b) die noodsaaklikheid om nuwe sakemodelle wat nuwe leksikografiese projekte kan finansier, te vind; en (c) die voorkoms van groeiende kompetisie van ander inligtingsbronne, bv. Google. Hierdie tendense het 'n bepaalde invloed op gespesialiseerde woordeboeke, wat hier gedefinieer word as hulpmiddels wat gebiede buite die algemene kulturele kennis en ooreenstemmende Taal vir Algemene Doeleindes (TAD) dek. Hierdie artikel steun die beskouing dat leksikograwe bogenoemde tendense kan hanteer deur gestruktureerde versamelings leksikografiese data met detail voor te berei wat maklik in inligting omskep kan word en in woordeboekskryfstelsels (WSSs) geberg kan word wat veelvoudige kombinasies en moontlike afvoere moontlik maak. Hierdie opsie is toepaslik vir die tendense, aangesien dit voldoende tegnologieë gebruik, bv. metodes vir profielsamestelling of geïndividualiseerde soektogte, steun bied aan 'n nuwe sakemodel wat gebaseer is op die skep van leksikografiese data wat baie verskillende hulpmiddels kan voed, bv. *Write Assistant* (Fuertes-Olivera en Tarp 2020), en beter en meer presiese inligting as dié van Google en ander inligtingshulpmiddels kan aanbied.

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

Sleutelwoorde: LEKSIKOGRAFIESE DATA, TEGNOLOGIE, GESPESIALISEERDE TAAL, GESPESIALISEERDE LEKSIKOGRAFIE, E-LEKSIKOGRAFIE

1. Introduction

A review of recent publications on lexicography and terminology reveals that we are in the middle of a *Cambrian explosion* (Fuertes-Olivera 2016), a concept typically used for referring to situations that are similar to what occurred around 540 million years ago, when new life forms began to multiply and everything changed in a way that made life completely different (Siegele 2014). For instance, Fuertes-Olivera and Tarp (2020) mention three recent tendencies that are shaping lexicographers' work and point to the future. The first tendency is the increasing use of disruptive technologies in almost all aspects of the discipline. The second is the growing obsolescence of the business model which, by and large, has financed the compilation of reference works for the last five hundred years. The third tendency is growing competition from other information sources, especially Google. This article addresses how these trends can affect the creation of online lexical resources for specialized languages, which is the topic of this study. By doing so, I present my solution and see it in line with Gouws's main academic endeavour, which has focused on adapting theory to practice and practice to theory (Gouws 2011), and being always "open" to new ideas and methods (Domínguez Vázquez and Gouws 2023).

Firstly, we will review the concept of specialized language, and especially that of specialized online dictionaries (section 2). Next, we will focus on the concept of *lexicographic data,* which we envisage as central for the future of lexicography (section 3). In section 4, we will illustrate a proposal for structuring lexicographic data in novel ways with the aim of offering alternatives that might lead to a rebirth of lexicography. For reasons of space, we will concentrate on different consultation possibilities, especially those that can be connected with the creation of specialized information tools from the same Dictionary Writing System (DWS). A final conclusion will summarise the main points discussed and will reflect on future developments in online lexicography.

2. Specialized languages

Cabré (1998: 118-121) indicates that specialized languages are sub-assemblies of the common language with their own rules and specific units. They are not easily understood by non-specialists, have to be consciously learned, and undergo change, variation and other processes that are also found in the common language, i.e. the unmarked linguistic code all humans learn and use, for instance, in face-to-face conversations and similar daily encounters. The above broad definition has typically led researchers to focus on the rules and units that are restricted, i.e. not used in the common language, and have created ref-

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

erence works that only (or mostly) deal with these rules and units.

Regarding the different specialized rules and units described so far, we will focus on lexical units, i.e. single words or chains of words that form the basic elements of the vocabulary of specialized languages. These represent the backbone of specialized dictionaries, i.e. reference tools that cover areas outside general cultural knowledge and its corresponding Language for General Purpose (LGP). Hence, they include disciplines such as technology, natural and social sciences, health sciences and humanities, to name just a few. In addition, they should not be considered a very specific type of dictionary with rigid and well-established characteristics, but rather a series of lexicographic works with different characteristics, sizes and terms, such as "glossary", "terminological database", "lexicon", "technical dictionary", "encyclopaedia", "vocabulary", "domain ontology", "knowledge bank", and so on (Fuertes-Olivera and Tarp 2014: 7-8). Under this general definition, specialized dictionaries describe the various specialized languages and substances of these disciplines to provide direct, specific access to their cognitive elements.

Such works may be either in paper format or online, but some are only paper whereas others are only digital. Regarding specialized online dictionaries, Fuertes-Olivera (2016: 227-228) has commented on two findings that are shaping the discipline. The first is that only a limited number of these are really different from their printed counterparts, that is, in terms of the use they make of internet technologies, e.g. for favouring individualisation (i.e. customisation or profiling); this is a defining factor when differences between printed and online dictionaries are analysed. The *Accounting Dictionaries* (Fuertes-Olivera and Niño Amo 2018) to a certain extent exploit available technologies from information science like user profiling, filtering and adaptive hypermedia, and also frequently link to the internet, where already existing data are reused in order to satisfy users' specific needs (Bothma 2011).

The second finding is that the design and construction of specialized online dictionaries should combine the use of lexicographic theories and methods with the active participation of IT experts as well as experts in the field; these should be experts, for instance, in medicine if the dictionary describes medical concepts and language. Such cooperation is of paramount importance, and will lead to the design of a high-quality online dictionary, which, according to Fuertes-Olivera (2016: 229-230)

1. is an information tool that contains (or can contain) many types of lexicographic data — e.g. dictionary articles, systematic introductions, hyperlinks, and so on (see section 3, below);
2. offers the data in a structured way (see section 4, below);
3. is always for specific consultation with the aim of converting its structured data into information quickly and easily;
4. contains specialized relevant data categories for the user: terms, definitions, example sentences, instructions, non-verbal signs, and so on;

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

5. offers reliable data, i.e. the validity and suitability of the data offered have been checked;
6. makes use of internet technologies, especially those favouring individualisation;
7. consists of three related components: a lexicographic database, a search system and a graphical user interface (GUI);
8. is updated continuously.

As the above characteristics show, online dictionaries cover both the language and facts of one or more domains. Regarding facts, i.e. knowledge of a particular domain, Hashimzade et al. (2014: 11-16) explicitly indicate that only experts with proper knowledge can write definitions of economic concepts in dictionaries directed at experts and semi-experts, who are the typical users of specialized dictionaries. As for the language of a particular domain, gone are the days when specialized dictionaries had very few data types, typically lemma and/or equivalent, definition, and sub-domain, e.g. example 1 from the *Diccionario McGraw-Hill de Química bilingüe español–inglés/English–Spanish*:

mezcla azeotrópica *azeotropic mixture* [QUIM]

Disolución de dos o más líquidos, cuya composición no cabía por destilación. También conocida por azeótropo.

Example 1: Dictionary entry in *Diccionario McGraw-Hill de Química*

Instead, modern specialized (online) dictionaries aim at offering a very precise description of specialized lemmas. Fuertes-Olivera and Tarp (2014: 199-200), for instance, indicate that the DWS for the *Accounting Dictionaries* contains up to 21 lexicographic data categories for each accounting lemma (Table 1, below) so as to assist their target users: (1) translators and language staff; (2) accounting experts and semi-experts; and (3) students and laypersons interested in Danish, English and Spanish matters related to accounting.

Table 1: Data types and categories in the accounting lexicographic database

Data Type	Rationale
Lemma	Self-evident: all dictionaries describe lemmas.
Homonymy index (superscript in the dictionary)	Included when homographs have different inflectional paradigms or are countable or uncountable, respectively. It assists all user types in all situations: it restricts the meaning of lemmas.
Polysemy index (Arab numbers in the dictionary	Included when homographs have different definitions. It assists all user types in all situations. It restricts the meanings of the same lemma.

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

Language code to lemma	Included for indicating language and language variant of the lemma. It assists all user types in all situations.
Grammatical data addressed to lemma	Offers inflections, countability, active and passive forms of verbs. It assists all user types in two communicative situations: production and translation. It offers linguistic profiles of lemmas.
Equivalent	One Danish, English or Spanish equivalent in several dictionary combinations: Danish–English (Danish Lemma; Danish definition; English equivalent); English–Danish (English lemma; English definition; Danish equivalent); English–Spanish (English lemma; English definition; Spanish definition; Spanish equivalent); Spanish–English (Spanish lemma; Spanish definition; Spanish equivalent). It assists all user types in communicative situations.
Language code to equivalent	Included for indicating language and language variant of the equivalent. It assists all user types in communicative situations.
Grammatical data addressed to equivalent	Offers inflections, countability, active and passive forms of verbs. It assists all user types in two communicative situations: production and translation. It offers linguistic profiles of equivalents.
Definition of lemma	Explains the meaning(s) of lemmas. Always one definition per sense. Definitions are crafted in the language of the lemma. There is one exception: in the English–Spanish dictionary, definitions are in English and Spanish. It assists all user types in all situations.
Collocations	Short and long phrases but not full sentences. They are crafted in the language of the lemma. They *mainly* assist users in two communicative situations: production and translation. In addition, the Spanish dictionary also includes collocations that assist users in cognitive situations (this is a useful method for dealing with culture-bound subject fields).
Translation of collocations	In some dictionary combinations, collocations are translated: English collocations are translated into Danish and Spanish, and Danish and Spanish collocations are translated into English. They assist all user types in two communicative situations: production and translation.
Language code to translation of collocations	Included for indicating language and language variant of the lemma and equivalent. It assists all user types in two communicative situations: production and translation.
Examples	Full sentences showing lemmas in use. They assist all user types in cognitive situations and two communicative situations: production and translation.
Translation of examples	They are full sentences showing equivalents in use. They are in some dictionary combinations: English examples are translated into Danish and Spanish, and Danish and Spanish examples are translated into English. They assist all user types in cognitive situations and two communicative situations: production and translation.

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

Synonyms and antonyms addressed to lemmas or equivalents	Assist all user types in two communicative situations: production and translation.
Language code to synonyms and antonyms	Included for indicating language and language variant of the synonym and antonym. It assists all user types in two communicative situations: production and translation.
Source	Hyperlinks to external texts. It mainly assists experts in cognitive situations. It can also assist translators and semi-experts in two communicative situations: production and translation.
Lexicographic notes	Usage, and/or contrastive notes that are addressed to lemma or equivalent, especially for indicating factual differences between lemma and/or equivalent, particular usages, cultural details, etc. They mainly assist all user types in cognitive situations.
Grammar notes attached to lemma and/or equivalent	Indicate language profiles of some lemmas and/or equivalents. They mainly assist users in two communicative situations: production and translation.
Proscriptive notes	Used for recommending lemmas. They mainly assist all user types in communicative situations.
Cross-references	Hyperlinks to internal texts. They assist all user types in cognitive situations and two communicative situations: translation and production.

Source: Adapted from Fuertes-Olivera and Tarp 2014: (199-200)

Table 1 also speaks of a new trend in online lexicography, namely, that the time is ripe for creating lexicographic data (section 3) that pay scant attention, if any, to dictionary typologies. Bowker (2018) seems to concur with this idea by explaining the process of convergence of lexicography and terminology (i.e. specialized lexicography in this paper), which she views as inevitable and illustrates in Table 2:

Table 2: Evolution and convergence of characteristics associated with lexicography and terminology

	Lexicography	**Terminology**
Practitioner	mainly lexicographers, but with greater involvement from the general public (via crowdsourcing)	mainly terminologists, but with greater involvement from the general public and subject matter experts (via open and closed crowdsourcing)
Object of study	mainly words, but also some terms	mainly terms, but also some general language words or expressions
Domain	mainly general language, but also some specialised language	mainly the language of a specialised domain, but also some general language

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

Point of view	mainly descriptive	mainly normative/prescriptive in the public and academic sectors, but incorporating more descriptive elements in commercial settings
Approach	mainly semasiological	increasingly semasiological, but retaining some onomasiological elements where useful
Organisation	mainly alphabetical, but sometimes incorporating thematic elements	mainly thematic, but allowing alphabetic searching
Main information provided	words, meanings, examples, usage information (e.g. collocations, frequency, phraseology), a range of linguistic information (e.g. part of speech, pronunciation)	preferred term, variants, context and usage information (e.g. collocations, frequency, phraseology), meaning, conceptual relations
Intended users	lay people, professional and academic audiences	public sector (for language planning), domain experts, scientific/technical writers, translators (for bi- or multilingual resources), commercial enterprises

Source: Bowker (2018: 148)

3. Lexicographic data in the era of the internet

We define lexicographic data as any data that have been prepared or accepted by lexicographers and stored in a DWS with the aim of helping humans and/or machines convert them into information *in a straightforward manner*. This definition merits several comments. Firstly, any data in a DWS implies that the system must allow the inclusion of such data in any format, especially:

— Words such as **azeotrope**
— Running text, e.g. in definitions and systematic introductions
— Symbols, e.g. that of the dollar: **$**
— Sounds, e.g. that of *lion* (Figure 1):

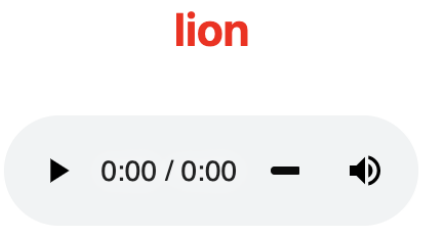


Figure 1: Pronunciation of *lion* in *howjsay*

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

— Films, e.g. how to make pancakes in *YouTube*
— Pictures, e.g. that of a lion (Figure 2)

Figure 2: Picture of a *lion*

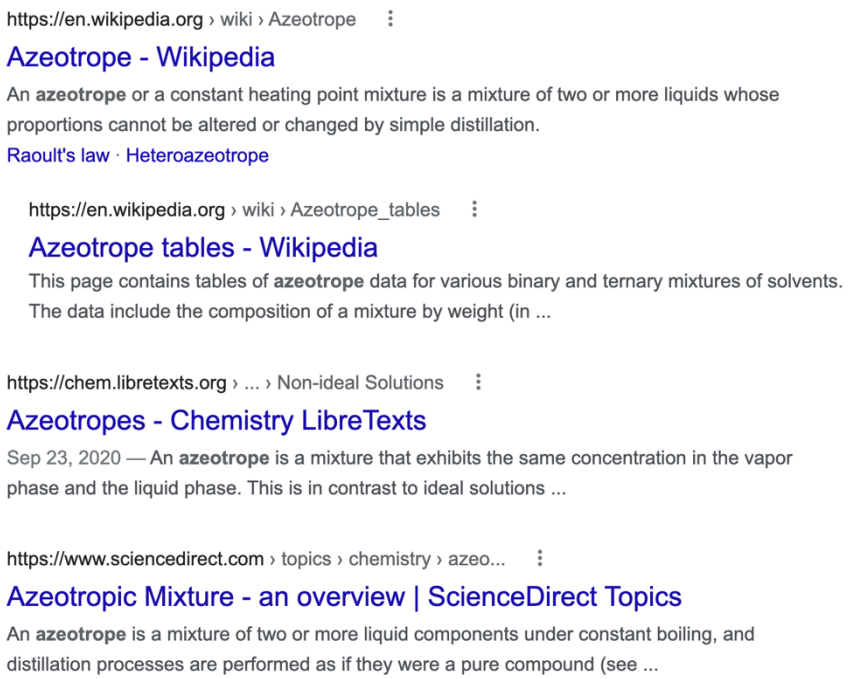
Secondly, the data should be prepared or accepted by lexicographers. This comprises selection, analysis and acceptance of the data to be included in the DWS. Fuertes-Olivera and Tarp (2014), among others, have shown that the process of selection based on restricted corpus data is not recommended in specialized lexicography. Instead, we believe that for selecting, analysing and accepting data we can rely on the internet as a corpus, and make use of big data analytics for extracting, say, the initial lemma list of the dictionary (Fuertes-Olivera et al. 2018), and of Google minitexts for preparing definitions, examples, grammar, etc. of each lemma (Fuertes-Olivera 2012; Fuertes-Olivera et al. 2018; Tarp and Fuertes-Olivera 2016). For instance, IT staff at Ordbogen.com, a Danish language technology company, tracked around one million daily searches in English and Spanish. They found that approximately 80% of these can be matched, i.e. the same search is identified in the logfiles of different dictionaries

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

and may, therefore, be interpreted in order to identify the most popular articles in the dictionaries under scrutiny. After two months of work with the logfiles of the searches, which amount to more than 60 million, IT staff at Ordbogen A/S were able to produce two lists of 20,000 English words and 16,000 Spanish words. They comprise the words most commonly searched in the period under analysis and were used by the editor of the *Diccionarios Valladolid-UVa* for compiling the initial lemma lists of this project (Fuertes-Olivera et al. 2018). Something similar can be done for selecting the initial lemma list of specialized dictionaries.

Regarding Google minitexts, they offer at a glance data that can be easily understood and used, especially if the retrieved data are analysed by an expert in the field. For instance, googling *azeotrope* retrieves around 786,000 hits. The first six hits (Figure 3) offer enough data for one to learn

(a) that *azeotrope* is a countable noun used in chemistry;
(b) its synonym is *constant boiling point mixture*;
(c) *azeótropo* and *mezcla aceotrópica* are its Spanish equivalent, one can be included as equivalent and the other as synonym of the equivalent;
(d) and that it is "a mixture of two or more liquids whose proportions cannot be altered or changed by simple distillation" because it "exhibits the same concentration in the vapor phase and the liquid phase".



http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

Figure 3: Initial hits when googling **azeotrope**

Thirdly, humans and machines, e.g. *Write Assistant* (Fuertes-Olivera and Tarp 2020), should convert lexicographic data into information in a single cognitive process or click. This is a really crucial point in our definition of lexicographic data. In these circumstances, most data in, say, existing Spanish dictionaries are not lexicographic, as they cannot be understood due to several flaws in their treatment and presentation, especially in terms of the use of abbreviations, recursive definitions and lack of adequate information. For reasons of space, we will illustrate this defining condition with two examples of current practices that should be abandoned:

— Recursive definitions should be discarded because they are useless. For instance, Spanish dictionaries define "action nouns", i.e. deverbal nouns that refer to an action or an event with the formula "acción y efecto de" (action and effect of) plus the verb they refer to (example 2):

dilución

acción y efecto de diluir[1]

Example 2: Definition of *dilución* in Spanish dictionaries, e.g. in DLE

Such a definition says nothing and, therefore, is not an example of lexicographic data. As indicated previously, lexicographers can google *dilución* (English: dilution), study the hits found and act accordingly. For instance, we have found that *dilución* has five meanings (example 3): one of them refers to the action of making a solid more dilute by dropping it into a liquid; another meaning refers to the solid that has been diluted (definitions 1 and 2 of example 3). Both are used in general language. In addition, it has a figurative meaning, also used in general language, which refers to the process

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

of weakening some abstract process, e.g. the action of making something weaker in form, content, value and so on (definition 4 of example 3), and two further meanings; a literal one used in chemistry, referring "to the process of decreasing the concentration of a solute in a solution, usually simply by mixing with more solvent like adding water to the solution" (Wikipedia: Dilution (equation) (definition 3 of example 3), and a figurative one found in economics, referring to reducing the value of a shareholding due to offering more shares without increasing assets (definition 5 of example 3). These five meanings, as well as many more data about the same, are all stored in the DWS of the *Diccionarios Valladolid-UVa* (Fuertes-Olivera 2019). These are examples of lexicographic data, as all of them illustrate the meanings of the word in the different contexts and domains found in real usage, and help potential users (humans and/or machine) to immediately disambiguate meanings and usages:

dilución

1. en sentido literal, disolución de un cuerpo sólido en un líquido (literally, making a solid more dilute by dropping it into a liquid)
2. en sentido literal, sustancia que resulta del proceso de disolución de algo por medio de un líquido (literally, solid that comes out of the process of dilution)
3. en sentido literal, reducción de concentración de una sustancia química en una disolución; es un procedimiento para preparar una disolución menos concentrada a partir de otra más concentrada; se usa en química (literally, process of decreasing the concentration of a solute in a solution; it is used in chemistry)
4. en sentido figurado, proceso realizado de forma consciente con el que se intenta un debilitamiento de algo (figuratively, conscious process used for weakening something)
5. en sentido figurado, disminución del valor teórico de las acciones de una empresa debido a la emisión de nuevas acciones sin prima de emisión, es decir, a la par o a un valor inferior al valor de mercado, o a la conversión de bonos u obligaciones en acciones; se usa en economía (figuratively, reducing the value of a shareholding due to offering more shares without increasing assets; it is used in economics)

Example 3: Definitions of *dilución* in the DWS of the *Diccionarios Valladolid-UVa*

— Dictionaries should always contextualize each meaning and usage. For example, Spanish dictionaries do not typically differentiate between literal and figurative meanings. Sometimes they offer a meaning and indicate that the lemma in question has figurative meanings but do not explain them. This is also of no use, so it should be discontinued. The difference between "literal" and "figurative" really matters. The vocabulary stock of specialized languages tends to be made up of three main processes of word formation: (a) derivation, e.g. "zero derivation" in English, such as *audit* (a noun,

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

a verb and an adjective; see examples (5, 6 and 7, below); (b) compounding, especially the creation of extended units of meanings and prefabricated chunks of words, such as *account day*, *define contribution scheme*, *defined contribution pension plan*, and so on; (c) and figurative extensions of general meanings, e.g. definition 5 of *dilución* in example 3. As shown in example 3, above, we believe it necessary to differentiate between literal and figurative meanings when defining words that have both types of meanings. This has led us to write in the DWS of the *Diccionarios Valladolid-UVa* the expression "en sentido figurado" (English: figuratively) at the start of the definition and "se usa en economía" (English: it is used in economics) at the end of the definition, as shown in example 3. These expressions are relevant, as will be shown in the next section.

4. Structured lexicographic data collections

Our definition of lexicographic data merits two main comments. Firstly, dictionaries must always use simple and easy-to-understand language and signs. This means that all language and signs that need more than one step to be understood should be eliminated. Secondly, the lexicographic treatment of lemmas should be as complete as possible. Both ideas will lead to the use of DWSs that should be both complex and dynamic, i.e. they can contain as much data as possible but will permit punctual consultation of the data really needed in a particular usage situation and by a specific user type.

We believe that this can be achieved by creating structured lexicographic data collections, i.e. repositories of data that have four defining characteristics. Firstly, they are created by human analysts, who must study the sociolinguistic context of words and offer detailed descriptions of all their relevant characteristics. These analysts can make use of all the (computer) tools and traditions they consider appropriate for offering deep and logical analyses of the words studied. For instance, the English noun *black swan* originated in day-to-day conversation for referring to *Cygnus stratus*, a term used in zoology, which describes a "species of swan which breeds mainly in the southeast and southwest regions of Australia" (Wikipedia: Black swan). The noun describes the physical and salient characteristics of a swan that was unknown until Europeans landed in Australia in the 17th century. In 2007, Nassim Taleb elaborated on a "black swan theory" in the domains of economics and political science. He referred metaphorically to a *black swan* as an unpredictable or unforeseen event, e.g. the September 11 attacks. This meaning has also re-surfaced again in general language to refer to something extremely rare, which can have positive or negative consequences. Hence, googling *cisne negro* (English: *black swan*) will inform human analysts of the animal *Cygnus stratus,* and two related metaphorical meanings, i.e. an extremely rare event, which is used in general language and may have positive or negative outcomes, and the specialized meaning in economics and political science, where it refers to an event that comes as a sur-

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

prise, has a major effect, and "is often inappropriately rationalized after the fact with the benefit of hindsight" (Wikipedia: Black swan theory). It is interesting to highlight that Spanish dictionaries ignore the metaphorical meanings of *cisne negro*, although this word is found in daily use in Spanish newspapers, magazines and TV programs. For instance, a search of "cisne negro" in Google retrieves 1,660,000 hits, most of which refer to above-mentioned metaphorical meanings and a film released by Hollywood in 2010.

Secondly, a description of the words must be placed in slots, i.e. allotted places in dictionary writing systems, each of which must be reserved for a particular data type, e.g. one slot for indicating that the noun is countable or uncountable, one for the different forms of the conjugation of the verb, one for its meaning, and so on. We believe that we need around 30 slots for each language: (1) lemma; (2) word class; (3) index number for homonymy; (4) index number for polysemy; (5) inflections; (6) grammar; (7) proscription notes, i.e. notes informing on sociolinguistic aspects of the word, e.g. recommended spellings; (8) meaning; (9) synonyms; (10) antonyms; (11) related words; (12) grammar notes; (13) usage notes; (14) diaphasic variant; (15) diastratic variant; (16) diatopic variant; (17) diachronic variant; notes on (18) synonyms, (19) antonyms and/or (20) related words; (21) notes on any variants; (22) sentence examples; (23) chunks of texts showing specific usages of the word, e.g. collocations; (24) figures; links to external sources, e.g. (25) to corpus data, (26) encyclopedic articles, e.g. Wikipedia; (27) symbols, equations, and so on; also, there should be three empty slots for possible use during the process of compilation.

To the best of our knowledge, existing specialized dictionaries tend not to include most of the above data types. For instance, we have investigated the accounting dictionaries retrieved when googling "accounting dictionary" and found a collection of reference works, all of which only offer the definition(s) of words, such as *audit* (example 4):

> AUDIT is the inspection of the accounting records and procedures of a business, government unit, or other reporting entity by a trained accountant for the purpose of verifying the accuracy and completeness of the records. It could be conducted by a member of the organization (internal audit) or by an outsider (independent audit). A CPA audit determines the overall validity of financial statements. A tax audit (IRS in the U.S.) determines whether the appropriate tax was paid. An internal audit generally determines whether the company's procedures are followed and whether embezzlement or other illegal activity occurred.

Example 4: Entry for *audit* in *Ventureline*

In our view, structured lexicographic data collections should contain better descriptions, such as, for example, our description of *audit* in the DWS of the *Diccionarios Valladolid-UVa* (examples 5, 6 and 7):

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

audit
noun: an audit, the audit, audits

definitions:

1. An audit is an official, methodological examination or review by an expert

Synonym: examination

Collocations:
- annual audits
- audit and quality assurance

2. An audit is an examination of the financial report and statements of an enterprise by an independent auditor after which the auditor gives an opinion which is expressed in an audit report.

Synonym: auditing

Collocation:
- Plan an audit
- Perform an audit
- The scope of the statutory audit
- Etc.

Example:
- An audit also includes assessing the accounting policies used and significant estimates made by the Board of Directors, as well as evaluating the overall Annual Report presentation.
- We conducted our audit in accordance with International Standards on Auditing.

Example 5: Dictionary entry for *audit* as a noun in the DWS of the *Diccionarios Valladolid-UVa*

audit
verb: audits, audited, has audited, auditing, is audited, are audited, was audited, were audited, been audited, being audited

Definitions

1. To audit is to independently examine and subsequently express an opinion on the financial statements of an organisation.

Collocations:
- have the annual report audited
- audit the accounts independently

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

Example:

- We have audited the accompanying balance sheet of the ABC Company as of 31 December 2016.

2. To audit is to check something officially.

Collocations:

- audit the accounts

Example 6: Dictionary entry for *audit* as a verb in the DWS of the *Diccionarios Valladolid-UVa*

audit
adjective

definition:
The placement of 'audit' in front of a noun specifies the meaning of the noun to pertain to the field of auditing: audit services, for example, are services performed with a view to auditing financial statements.

Collocation:

- special audit consideration
- an audit requirement
- draw conclusions from audit observations

Example 7: Dictionary entry for *audit* as an adjective in the DWS of the *Diccionarios Valladolid-UVa*

Thirdly, there are three types of slots: "open slot"; "restricted slot" and "expanding slot". The open slot, e.g. the definition slot, is prepared for storing running text without limited space. The restricted slot is reserved for roll-down menus that display sets of previously defined categories of data, e.g. "informal Spanish". For instance, in a Spanish dictionary, there may be 23 geographical labels for referring to an "americanismo" (English: Americanism), a geographical variant typically used for referring to a meaning that is used outside Spain, e.g. in Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Philippines, Puerto Rico, United States, Uruguay, Venezuela and West Sahara. This second type of slot allows the lexicographer to include "and data", e.g. a word can be used in Argentina, Bolivia, and Mexico, and "either or data", i.e. if the word is "informal" it is not formal. Finally, the "expanding slot" is reserved for data types that have different forms, e.g. Spanish verbs can have up to 56 different forms of the same verb, which are stored in sub-slots, each reserved for a specific form. In other words, the slot for verb conjugations can have as many as 56 sub-slots, although not all of them are needed for all verbs. It is the lexicographer's task to analyse each

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

verb and decide which of the different forms are needed in the DWS. In the *Diccionarios Valladolid-UVa*, we have limited the number of sub-slots to eight, as these will allow the retrieval of the verb, no matter which form is used in the search engine. This is important because in Spanish dictionaries search strings such as "comíamos" (English: we ate) retrieves nothing (Figure 4):

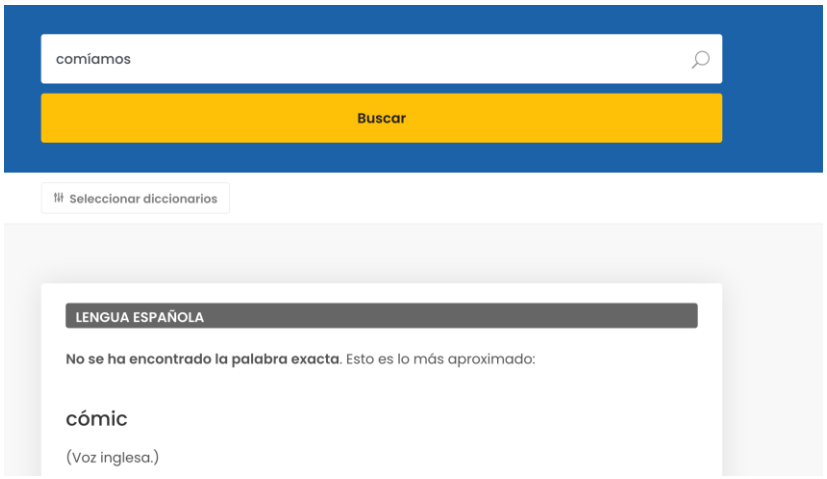


Figure 4: Searching "comíamos" in *Diccionarios.com*

The three types of slots should allow different types of connections among them, especially for enabling different types of retrieval. Table 3 shows the technical document used for preparing a grammar and spelling dictionary with the data types stored in the DWS of the *Diccionarios Valladolid-UVa*:

Table 3: Technical document for preparing a grammar and spelling dictionary of Spanish

search in the fields + search sequence	**Field**	**sequence in the dictionary+text**	**shown as a list if more than 10 results**	**explication text**
1. lemma field (including inflection forms) 2. *lemma* 3. Fuzzy search	1. Lemma	1		

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

	2. Style marker to lemma			
	3. Homonym number	2		
	4. Polyseme number	8 (but only if polysemy by a homonym OR if a grammar remark by a polyseme)		
	5. Meaning	9 (but only if homonymy OR if a grammar remark by a polyseme)		
	6. Lexical remark			
	7. Lexical remark for text production			
	8. Word class and expression class	3		
	9. Grammar, inflexion class 1	4		
	10. Grammar, inflexion class 99	5		too, but not recommended <***>
	11. Valency	6		
	12. Grammar/spelling remark	7		
	13. First reference	10		This variant is not recommended, use instead ***
	14. Second reference(s)			
	15. Collocation(s)			
	16. Example(s)			
	17. Word formation(s)			
	18. Synonym(s)			
	19. Style marker to synonym			
	20. Antonym(s)			
	21. Style marker to antonym			
	22. Synonym remark			
	23. Internet link			
	24. Dictionary grammar			FIELD NOT USED
	25. Memo field			

Source: Henning Bergenholtz (personal communication) and IT staff at Ordbogen A/S.

Table 3 is in line with Gouws's concept of "mother dictionary" (2014). It offers flexibility and allows lexicographers to create a central data collection, which

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

can be displayed differently with the aim of catering to different user needs in different usage situations. Table 3, then, tells IT experts how they should connect the different slots (and sub-slots) in order to retrieve only the specific data type needed in the usage situation envisaged, which is a user searching for grammar and/or spelling information. The label "search in the field + search sequence" informs them that the search string should be prepared for three types of searches: the lemma as it is, including its inflected forms, part of the lemma (*lemma*), and fuzzy search. This will help users retrieve all possible search strings. The label "field" enumerates the possible slots needed for this type of specialized dictionary. For example, it will need a slot for storing a note on grammar and/or spelling (number 12 in Table 3). The labels "sequence in the dictionary + text" and "explication text" indicate which lexicographic data must be retrieved, i.e. which slots are to be retrieved, in which order and possible wording of any rule. For example, searching *a ver* in a grammar or spelling dictionary of Spanish, will retrieve the following (example 8):

- Lemma: *a ver*
- Expression (i.e. the word class). It is used as an adverb.
- Spelling variant: *veamos*
- Meaning:
  1. expresión que se usa para indicar que se siente expectación, ganas o curiosidad por que ocurra determinada situación que se menciona justo después de la expresión, o por observar de qué forma sucede (it works as a formula for showing interest in someone or something)
  2. expresión que se emplea para pedir a alguien que enseñe o muestre algo que se desea observar o conocer (it asks someone to show something he or she has)
  3. expresión que se emplea para invitar a otras personas a adoptar una actitud de espera con el fin de conocer o comprobar qué sucede con determinada situación sobre la que hay incertidumbre (it is used for inviting someone to adopt a particular stance in an unknown situation)

Example 8: Lexicographic data retrieved when searching *a ver* in the DWS of a grammar and or spelling dictionary of Spanish

Finally, some of the slots must be equipped with "add-on" buttons, whose function should also be explained in technical documents such as that of Table 3. Both methods will allow the lexicographic data stored in a DWS to be accessed in multiple ways. For instance, the add-on button in the homonym index slot of *audit* (examples 5, 6 and 7 above) indicates that we have three different words, each of which is linked separately to the meaning slot, which also includes an add-on button for adding as many meanings as necessary. Each add-on button of the meaning slot controls the synonyms, collocations and examples of each meaning. This leads to the joint retrieval of each meaning with its own lexico-

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

graphic data, as shown in examples (5), (6) and (7).

5. Conclusion

According to Fuertes-Oliver and Tarp (2020), lexicography is currently being shaped by three related tendencies: (a) the increasing use of disruptive technologies; (b) the necessity of finding new business models that can finance new lexicographic projects; and (c) the existence of growing competition from other information sources, e.g. Google. These three trends are especially influencing specialized dictionaries, here defined as tools that cover areas outside general cultural knowledge and its corresponding Language for General Purpose (LGP).

This paper has illustrated how the creation of online lexical resources for specialized languages can deal with the above-mentioned tendencies, by making a very precise definition of lexicographic data which must be stored in structured collections. Lexicographic data is any data in any format that has been prepared and/or revised by lexicographers and can be converted into information rapidly and straightaway, e.g. without further look-ups.

Structured lexicographic data collections are repositories which are created by humans who have studied the sociolinguistic contexts of each word, offering detailed descriptions of meanings and usage and placing them in different types of slots, each of which can be linked in different ways so as to allow potential users to retrieve the data they need in each particular usage situation.

If properly implemented, these ideas indicate that the future of online lexical resources for specialized languages depends on implementing data collections that can be restricted during the process of retrieval but not during the process of compilation. In short, lexicographers should store in the DWS as much data as possible and work with IT experts to create DWSs that allow multiple combinations and multiple retrievals. This option can handle the three tendencies, as it uses novel technologies, e.g. ways of profiling or individualizing searches, favours a new business model based on the creation of lexicographic data that can feed many different tools, and offers better and more precise information than that provided by Google and other information tools.

Acknowledgments

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

References

Bothma, Theo J.D. 2011. Filtering and Adapting Data and Information in the Online Environment in Response to User Needs. Fuertes-Olivera, Pedro A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 71-102. London/New York: Continuum.

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

Bowker, L. 2018. Lexicography and Terminology. Pedro A. Fuertes-Olivera (Ed.). 2018: 138-151.

Cabré, M.T. 1998. *La terminologie: théorie, méthode et application. Traduit du Catalan, adapté et mis à jour par Monique C. Cormier et John Humbley.* Ottawa: Les Presses de l'Université d'Ottawa.

Diccionario McGraw-Hill de Química. Parker, S.P., J. Weil, B. Richman, E.J. Fox, J. Faulk and F. Jr. Kotowski (Eds.). 1991. *Diccionario McGraw-Hill de Química bilingüe español–inglés/English–Spanish.* Mexico: McGraw-Hill.

Diccionarios.com. Diccionarios con la garantía Larousse and Vox. https://www.diccionarios.com/. Last access: July 28, 2022.

DLE = *Diccionario de la Lengua Española***.** Real Academia Española. https://dle.rae.es/. Last access: July 28, 2022.

Domínguez Vázquez, María José and Rufus H. Gouws. 2023. The Definition, Presentation and Automatic Generation of Contextual Data in Lexicography. *International Journal of Lexicography* 36(3): 233-259.

Fuertes-Olivera, Pedro A. 2012. Lexicography and the Internet as a (Re-)source. *Lexicographica* 28: 49-70.

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

Fuertes-Olivera, Pedro A. (Ed.). 2018. *The Routledge Handbook of Lexicography*. London/New York: Routledge.

Fuertes-Olivera, Pedro A. 2019. Designing and Making Commercially Driven Integrated Dictionary Portals: The *Diccionarios Valladolid-UVa*. *Lexicography* 6(1): 21-41.

Fuertes-Olivera, Pedro A. and M. Niño Amo. 2018. The Accounting Dictionaries. Pedro A. Fuertes-Olivera (Ed.). 2018: 455-472.

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

Fuertes-Olivera, Pedro A. and S. Tarp. 2020. A Window to the Future: Proposal for a Lexicography-assisted Writing Assistant. *Lexicographica* 36: 257-286.

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.

Gouws, Rufus H. 2011. Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries. Fuertes-Olivera, Pedro A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography:* 17-29. London/New York: Continuum.

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

Hashimzade, N., G.A. Myles and G.D. Myles. 2014. Can Authority Be Sustained while Balancing Accessibility and Formality. *Hermes. Journal of Language and Communication in Business 27*(52): 11-24.

Howjsay. https://howjsay.com/. Last access: July 28, 2022.

Siegele, L. 2014**.** A Cambrian Moment. *The Economist*, January 18, 2014: 1-14 (Special Report).

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.

Ventureline = *Ventureline. Accounting Terms.* https://www.ventureline.com/accounting-glossary/. Last access: July 28, 2022.

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-2-1832 (Article)

Wikipedia. Black swan theory: https://en.wikipedia.org/wiki/Black_swan_theory. Last access: July 28, 2022.

Wikipedia. Dilution (equation): https://en.wikipedia.org/wiki/Dilution_(equation). Last access: July 28, 2022.