ENCODING AND PUBLISHING IDIOM DICTIONARIES USING XML TECHNOLOGIES

Ana Rita Vieira[1], Idalete Dias[1], Alberto Simões[2]
[1]Center of Humanistic Studies/ILCH, Universidade do Minho, Braga, Portugal
pg38980@alunos.uminho.pt; idalete@ilch.uminho.pt
[2]2Ai, School of Technology, IPCA, Barcelos, Portugal.
asimoes@ipca.pt

Abstract

This project aims to provide an interesting and efficient way to publish a bilingual lexicographic online resource of idiomatic expressions. In this prototype, we are encoding and making available Schemann's *Synonym Dictionary of German Idioms* and his bilingual *German-Portuguese Idiomatic Dictionary*. These two dictionaries follow two completely different structures, as the *Synonym Dictionary* uses an onomasiological framework, in contrast to the usual semasiological approach applied in the *Bilingual Dictionary*.

These dictionaries were encoded with the Text Encoding Initiative (TEI) schema and its search is supported by the eXist-DB database, following other works on Online Dictionary publishing. Encoding a dictionary of idioms is not trivial, and it gets more complicated when including different information sources. Thus, and while TEI has a comprehensive set of tags to encode dictionaries, it needs adaptations to be able to encode our data properly.

Another challenge for this project is to understand what is the best way to allow the user to search the dictionaries, finding the desired information, focusing on how to allow the proper use of an online onomasiological dictionary. The current prototype allows the users to search for idiomatic expressions by words, by concepts or to browse an ontological structure. This is supported by the cross-reference and linkage of the dictionaries, bringing together the onomasiological and semasiological approaches. Focused on the user's needs and according to the most recent online dictionary studies, the search tool is prepared to help the users in lexical reception and production, as well as in translation tasks.

Keywords: Idiomatic Expressions, Online Dictionary, Text Encoding Initiative, XML, XQuery

1 Introduction

Online bilingual lexicographical resources specialized in idioms are still quite scarce to this day. If managing multi-word expressions is a great challenge, dealing with idioms is an even greater challenge. Both lack the principle of compositionality, as the meaning of the whole is not obtainable from the meaning of its constituent parts. But while it is rather easy to find common multi-word expressions in dictionaries, lexicons, and computational corpora, there is a significant lack of resources containing idioms.

This project aims to fill in this gap by creating electronically encoded idiom dictionaries based on dictionaries available in print. In this sense, the *German-Portuguese Idiomatic Dictionary*, a semasiological dictionary, and the *Synonym Dictionary of German Idioms*, an onomasiological dictionary, both compiled by Hans Schemann, provided us with the required information. As studies on the modelling of online lexicographic data (Klosa, 2013; Müller-Spitzer, 2014a) have shown, the process of computerizing and

publishing dictionaries on the Internet is not a trivial task, since the relationship between the dictionary data, specific user needs, and possible access routes and search paths to satisfy user demands must be considered (Tarp, 2013). One of our objectives is to use all the information available in the above-mentioned print dictionaries, including structural features and content, to design an integrated model capable of enabling users to carry out focused searches. To achieve this, a detailed XML schema was developed, on the one hand, to represent the complex system of lexicographic information contained in the dictionaries and, on the other, to support the computational tool that will facilitate the dictionary use.

In this paper, we describe the modelling and encoding principles applied to the dictionaries with a focus on: (i) integrating different lexicographic structures and types of information in one electronic dictionary; (ii) providing the users with useful results that match queries in specific communicative and cognitive situations. Therefore, as a first step, we were faced with the challenge of designing a proper granular annotation schema using a dictionary encoding standard to represent both dictionaries' macro and microstructures at the same time ensuring the interconnectedness between the dictionaries' entries and the modelling of all possible querying outputs.

Following the encoding process, the next challenge was to create a search interface that can, on the one hand, mirror and preserve the dictionaries' semasiological and onomasiological approaches; and, on the other hand, provide easier and quicker methods to respond to the needs of the target users. Every decision taken throughout this project was taken with the user in mind, with well-defined target users, user needs, and user lexicographical situations, as advocated by the Function Theory of Lexicography (Tarp, 2012, 2013, 2014) combined with the most recent studies on online dictionary use.

The current prototype is best suited for Portuguese advanced German L2-Learners and German advanced Portuguese L2-Learners, as well as translators, guiding them throughout the interface according to their communicative situations of text reception, text production, and translation.

The remainder of this document is structured as follows: Section 2 discusses related work, focusing on different projects that are publishing dictionaries online, including idiom dictionaries. Section 3 describes the characteristics of the two print dictionaries that were annotated for this project. The annotation schema is explained in Section 4, and Section 5 focuses on the prototype developed to support the user experience. Finally, we conclude in Section 6 with some insights on the current project status and the expected future developments.

2 Related Work

In the last decade, we have witnessed a clear decline in the creation and selling of printed dictionaries. With the advent of digital devices and continuous access to the Internet, users tend to use online tools to gather information about words and expressions. Recent studies on user behaviour have shown that searches on a language problem are performed directly in a web search engine and not in a dictionary website (Sascha et al., 2018). The result is that publishing houses are not interested in further developing their dictionaries, and have been focusing on making them available on the Internet. We can find examples of online dictionaries for most languages, from major publishers: for Portuguese (*Porto Editora*[1]), English (*Oxford*[2] and *Cambridge*[3] Dictionaries), French (*Larousse*[4]), or German (*Duden*[5]), just to mention a few. The same is true for Academy dictionaries, like *Real Academia Española*[6], *Académie Française*[7] or *Academia das Ciências de Lisboa* (Salgado et al, 2019).

1 Integrated in the Infopedia Web Portal, https://www.infopedia.pt/.
2 Available at https://www.oed.com/, through subscription
3 Available at https://dictionary.cambridge.org/.
4 Different dictionaries available at https://www.larousse.fr/.
5 Available at https://www.duden.de/.
6 Available at https://dle.rae.es/.
7 Available at https://www.dictionnaire-academie.fr/.

Most of this work still follows a paper-based approach (Tarp, 2012): each entry of the dictionary is stored in a database, allowing the user to search for the headword and, in some specific situations, by words present in the definitions. The relationship between entries is done in the same way as was presented on paper, with specific "*see also*" sections. More recently, as there is a large diversity of dictionaries online for every language, some attention is being given to the interaction with the user, not just by giving them cleaner interfaces, but also presenting more information than just the word entry.

When it comes to onomasiological dictionaries, most of the available projects on the Internet are thesauri. Some examples are the *Old English Thesaurus*[8] or *VisuWord*[9] for the English Language, the *Caldas Aulete* dictionary for Portuguese[10], the *OpenThesaurus* for German[11] or the multilingual *ConceptNet*[12]. These network-like dictionaries can be compared to the diverse *WordNet* (Miller, 1995) projects, available for most languages.

There are three other lexical projects that we would like to emphasize for supporting both semasiological and onomasiological queries: The *ANW Dictionary* (*Algemeen Nederlands Woordenboek*) [General Dutch Dictionary], *CombiDigiLex* and *Tesouro do léxico patrimonial galego e português* [Galician and Portuguese word bank]:

- The *ANW Dictionary*[13] is "not a clone of an existing printed dictionary [and] it truly represents a new generation of electronic dictionaries in the sector of academic and scientific lexicography" (Moerdijk, 2008). It is a corpus-based dictionary of written Dutch and it pays special attention to '*semagrams'*: "conceptual structure elements which characterise the properties and relations of the semantic class of a word meaning", playing an important role in onomasiological queries (idem, 2008).
- *CombiDigiLex*[14] is an ongoing multilingual project, focused on the analysis of the possible lexical combination of 'communication, movement, emotion, perception and transfer verbs' in German, Spanish and Portuguese, framed within a conceptual macrostructure and built to assist in L2-text production;
- *Tesouro do léxico patrimonial galego e português*[15] is a lexical open data portal for the Galician, European Portuguese and Brazilian Portuguese lexicon. Querying this database provides information to be of use for dialect comparative studies, presenting the search results by a choice of lexical forms, by location and by semantic field. The search interface is built around a semantic classification of words and idioms, composed of twelve major categories with further divisions. To search by concept, the user has to go to the advanced search tool and select a concept (eg. 7.1 - Humans [physical, psychological and behavioural aspects]), showing a list of results that can be filtered by location and grammatical category, for instance selecting 'adverbial expression' (eg. *em leitão* [naked], an European Portuguese adjectival and adverbial expression, which is also an idiom).

Regarding Idiom dictionaries, there are relatively few projects on the web. Most of the available projects are not based on academic dictionaries normally associated with a publishing house. Here we will refer to a few:

8 Available at https://oldenglishthesaurus.arts.gla.ac.uk/category/.
9 Available at https://visuwords.com/.
10 Available at https://www.aulete.com.br/analogico/.
11 Available at https://www.openthesaurus.de/.
12 Available at https://conceptnet.io/.
13 Available a https://anw.ivdnt.org/.
14 See more at https://combidigilex.wixsite.com/website.
15 Available at http://ilg.usc.es/tesouro/en.

- *The Idioms*[16] is focused on the English language. The idiomatic expressions in this resource are organized by topics that lead the user to a list of idioms that are related to the chosen topic. As an example, clicking on the topic 'problem' will present the user with a list of entries, each containing the idiomatic expression (*time puts everything in its place*; *pour oil on troubled waters*; *elephant in the room*; *a hard nut to crack*, etc.), the definition and an example sentence. The user can decide if more information should be presented on a specific idiom by clicking on the button 'Read on': the user will have access to paraphrases, more example sentences, information on the origin of the expression and synonyms. A second access route is provided by the 'Complete List' option that lists all the idioms in the database, although it is not quite clear which ordering strategy was followed. The interface allows for a third access route via a search field. All expressions and example sentences that contain the word entered in the field will appear.
- *Dictionnaire d'Expressions Idiomatiques*[17] is a bilingual Portuguese - French resource of idiomatic expressions. One may search the dictionary via a given list of the expressions or concepts in both languages, resource which serves as a great reference to our project of an online dictionary featuring both semasiological and onomasiological approaches and also synonym relations.
- *Expressio*[18] includes monolingual, bilingual and multilingual dictionaries of idiomatic expressions. When using the multilingual version, results are shown in a table, presenting the equivalent expression in the target language along with its literal translation. For each entry, it includes examples and suggests other idioms with similar meaning. Users can comment on the entries, and suggest changes in the dictionary. Taking the bilingual Portuguese - French as example, for each expression, the user has access to the variant of Portuguese (European or Brazilian) to which the idiom belongs, the French equivalent expression, and a literal translation of the Portuguese source idiom in French, a feature that could prove useful for French learners of Portuguese as a foreign language. For example, the search results for the Portuguese idiom "*não se dar por vencido*" [*to keep one's chin up*] will include "*contre mauvaise fortune bon coeur*" (French equivalent) and "*ne pas se donner comme vaincu*" (literal translation) [do not give oneself as defeated].

3 Dictionaries/Resources

Before initiating the electronic encoding of the dictionaries, we needed to understand how the idioms are presented and organized in both dictionaries and what are the different approaches to using them, particularly in a situation in which the user has to search in both dictionaries to find what he/she wants. The user is faced with a challenging task due to the differences in the dictionaries' structures, as the *Synonym Dictionary* uses an onomasiological approach, in contrast to the usual semasiological approach applied in the *Bilingual Dictionary*. In the former, the idioms are grouped into synonym sets (synsets) which, in turn, are organized into concepts, according to Schemann's knowledge categorization, creating a hierarchical structure that allows the user to explore synonymous expressions that convey the same concept. This lexicographic perspective is particularly useful for writing tasks, that is encoding purposes, to assist users who are looking for expressions that designate a specific concept in their native or a foreign language (Sierra, 2000).

The *Bilingual Dictionary*, on the other hand, follows the more usual alphabetical macrostructure. However, this ordering principle may also present some difficulties to the user because idioms are multi-

16 Available at https://www.theidioms.com/.
17 Available at https://www.cnrtl.fr/dictionnaires/expressions_idiomatiques/.
18 Available at https://www.expressio.fr, https://www.expressio.fr/expressions-idiomatiques-en-portugais.

word expressions. The user will be faced with the challenge of figuring out which keyword was used by the lexicographer to catalogue the idiom. In addition, whenever the same expression is catalogued under different entries, there will be cross-references instructing the user to search for a related entry where the full entry content data can be consulted. This feature may contribute to frustrating user experiences. Note that we are dealing with printed editions, and therefore the lexicographer is bound to space constraints demanding ingenious ways to plan and compile the different textual structures: from the data distribution structure, the access structure, the macrostructure, microstructure, mesostructure, not to mention the addressing structure and search zone structure (Wiegand, 1990; Müller-Spitzer, 2014a). Despite the detailed description of the ordering strategies provided by Schemann in the outer text of the *Bilingual Dictionary*, there is general agreement that, due to several reasons, including the lack of a dictionary culture which can be achieved by dictionary pedagogy, users seldom consult(ed) the outer texts for guidance on how to use a print dictionary (Gouws, 2010). For electronic and online dictionaries we are not dependent on the number of pages allowing the use of technology to automate these indirections and offer a better user experience.

3.1 *Synonym Dictionary of German Idioms*

The macrostructure of the *Synonym Dictionary of German Idioms* was obtained by applying a bottom-up categorization approach that consisted, firstly, in grouping the 18959 German idioms into synsets, followed by the categorization of the synsets with a common denominator into subconcepts that are linguistically realized by what Schemann designates as archilexemes. It is important to emphasize that it is not always possible to find one lexical unit that satisfactorily delimits the semantic content of a group of synsets that belong together. For this reason, Schemann (2012) made use of more than one lexical unit and/or specific idioms within the synsets to act as delimiters. Once at the level of the archilexemes, Schemann grouped these subconcepts into higher-order generic concepts and these, in turn, into nine macro concepts which represent a given organization of the world. Schemann's conceptual system has proved to be a reference for the conceptualization of onomasiological dictionaries (Dias, 2010).

The first section of the dictionary is the browsing section where the nine macroconcepts are presented as follows:

A: *Zeit, Raum, Bewegung, Sinnesdaten* [Time, Space, Movement, Sensory data]

B: *Leben - Tod* [Life - Death]

C: *Physiognomie des Menschen* [Human physiognomy]

D: *Stellung zur Welt* [Attitude to the world]

E: *Haltung zu den Mitmenschen* [Attitude towards fellow human beings]

F: *Einfluß, Macht, Verfügung, Besitz* [Influence, Power, Disposition, Possession]

G: *Kritische Lage, Gefahr, Auseinandersetzung* [Critical situation, Danger, Conflict]

H: *Präferenzen* [Preferences]

I: *Quantitäten, Qualitäten, Relationen* [Quantities, Qualities, Relations]

These concepts are further subdivided into more specific ones, composing a structure made out of three levels, that are identified by a sequence of characters. For instance, the path "*B - Ba - Ba1*" refers to the top-level concept "*B: Leben - Tod*" [Life - Death], the second-level concept "*Ba: Geburt - Tod*" [Birth - Death], and the third-level concept "*Ba1: Geburt*" [Birth]. After selecting the desired concept, the user has to search for it in the second and main part of the dictionary, the *Systematischer Teil* [Conceptual Part], which is where we find the idiom synsets: a collection of idioms grouped under a specific concept, such as

"*Ba1: Geburt*" [Birth] (36 idioms grouped under 21 synsets). The Conceptual Part of the dictionary assists users with text production in the native language and the foreign language. Two possible communication-oriented functions of the dictionary can be described as follows: a native speaker of German who would like to explore possible synonymous idioms that express the same concept by analyzing and comparing the underlying images of the expressions that belong to the same synset and then decide which better fits a given context; an advanced learner of German as a Foreign Language may use the dictionary in a similar situation. In this case, the user might need to resort to complementary lexicographic resources, such as a bilingual dictionary.

The *Alphabetischer Teil* [alphabetical part] contains all the expressions present in the conceptual section of the dictionary, ordered alphabetically by the keyword that was chosen by the lexicographer. Each expression is linked directly to the conceptual part via a synset code, that remits the user to the specific synset in which it is found. This part is more directed to those who have a (key)word in mind and would like to consult the exact form of the idiom. For example, "*(ein Kind) zur Welt bringen*" [*to bring a child/... into the world*] - Ba 1.4. This section was important for the electronic encoding of the dictionary, as it allowed the alignment between the *Synonym Dictionary* and the *Bilingual Dictionary*, as will be described in Section 4.

The last part of the *Synonym Dictionary* is the *Such-und Stichwortregister* [keyword search index], which lists the archilexemes on the subconcept level in alphabetical order. This index at the end of the dictionary forms part of the onomasiological approach as it provides the user with another way to search for a concept via the archilexemes or keywords that are used to categorize the synsets.

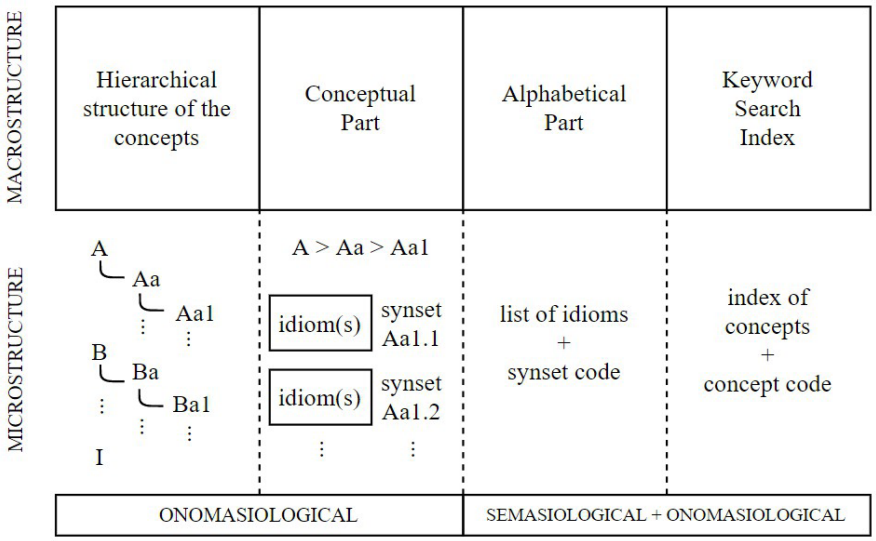

Figure 1 - The *Synonym Dictionary*'s macro and microstructures.

To sum up, the users can search this dictionary starting from any one of the sections presented in Figure 1. The sections can be chosen according to the users' knowledge of the language, the usage situation and corresponding cognitive and communication-oriented functions. The functions are presented in more detail below:

(i) Cognitive functions:

- The user is interested in consulting Schemann's categorization of the world and the methodology used to construct such a system. The user might also be interested in comparing Schemann's conceptual system with other systems that have been developed, such as the *Diccionario Ideológico de la Lengua Española* by Casares, or even take Schemann's system as a reference for a new proposal.

(ii) Communicative function mostly related to text production by a native speaker of German or an advanced learner of German as a Foreign Language:

- The user is familiar with a specific idiom, such as "*(ein Kind) zur Welt bringen*" [*to give birth to (a child)*] and would like to find synonymous expressions. He/she would start by searching in the Alphabetical Part, first by looking for the keyword 'Welt' and, then, selecting the expression from the list of other expressions with the same keyword. The user is redirected to the Conceptual Part through a synset code Ba 1.4, corresponding to the fourth synset belonging to the third level concept *Geburt* [Birth]. In this synset, the user will find the idiom he/she started from and two other synonyms: *(einem Kind) das Leben schenken; (Kinder) in die Welt setzen* [*to bring a child/... into the world, to give birth to a child/a girl/...*].
- The user would like to convey a specific concept but is not familiar with idioms that are used to express that same concept. In this situation, the Keyword Search Index will assist the user by redirecting him/her to the respective synset in the Conceptual Part. To be able to use the Keyword Search Index, the user will have to think of a clue word related to the concept. Since the keywords in the Index are the archilexemes defined by the lexicographer, the user's clue words may not always match the indexed keywords. The concept of 'birth' is delimited or indexed by the archilexeme 'Geburt' [Birth] in the dictionary. This means that the user will find the information he/she is looking for if the clue word is *Geburt*.
- The user looking for an idiomatic expression that conveys a specific concept may start by browsing the hierarchical concept system and select one of the nine macroconcepts (first-level concepts), followed by the respective second-level and third-level concepts. Once at the level of the synsets, the user will need to go through the synsets grouped under a specific concept to be able to identify which one of the synsets contains the expressions with the nuances he/she is looking for. In this case, the search proceeds from the hierarchical structure of the concept system to the Conceptual Part of the dictionary with the synsets. Figure 2 shows three out of a total of twenty-one synsets that are grouped under the third-order concept *Geburt*.

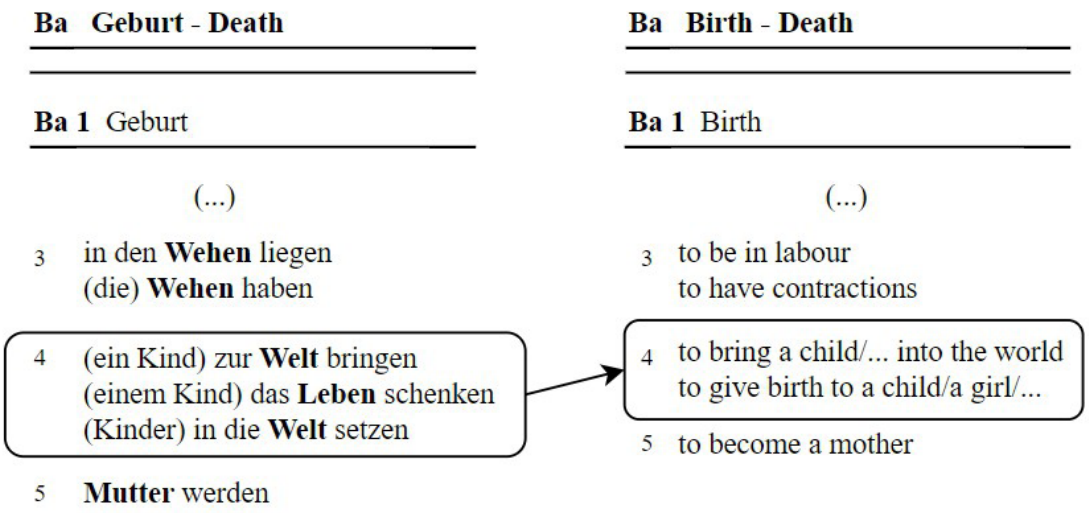

Figure 2 - Synset Ba1.4 in the Conceptual Part and its equivalent synset in English[19].

19 Note that at the present moment we are only using the *Synonym Dictionary of German Idioms*. Nevertheless, we present the same synset in English extracted from the *German-English Idiom Dictionary* from the same author for easier reading.

An analysis of the semantic relation between the synsets in Figure 2 reveals that the synsets are closely related to each other. This exercise of exploring and comparing the subtle semantic differences between the expressions in the contiguous synsets is demanding and can only be achieved by users with an advanced proficiency level. The more familiar one becomes with this dictionary, the more useful it will become.

3.2 *The German-Portuguese Idiomatic Dictionary*

The *German-Portuguese Idiomatic Dictionary*, including about 32.000 German idioms, is one of a series of five bilingual idiom dictionaries compiled by Schemann that stems from the main dictionary *Deutsche Idiomatik* (1993). Although its construction is semasiological (compare Figure 1 with Figure 3), it presents a certain level of complexity. Whereas the *Synonym Dictionary* provides the user with at least three access routes to the main onomasiological part of the dictionary with the synsets, the *German-Portuguese Dictionary* provides the user with only one access route to the entries via the keyword chosen by Schemann to serve as ordering principle. These keywords appear as headwords under which all expressions with the same keyword are grouped. For example, the headword *Welt* [world] subsumes a large number of expressions with the same keyword: *nicht die Welt sein* [*it isn't all that much/long/...*], *das Theater/Bücher/... ist/sind meine/deine/...Welt* [*the theatre/books/... is/are my/her/... world*], *um alles in der Welt* [*at all costs; come what may; for God's sake; for goodness' sake*], including *ein Kind/ein Mädchen/... zur Welt bringen* [*to give birth to a child/a girl/...; to bring a child/... into the world*].

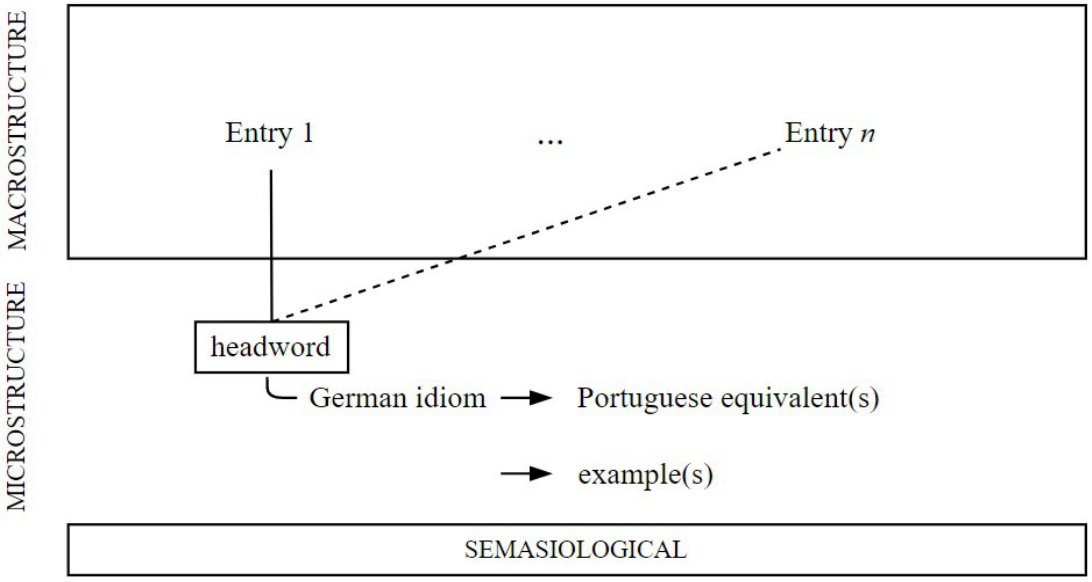

Figure 3 - The *German-Portuguese Idiomatic Dictionary*'s macro and microstructure.

By default, an entry in the *Bilingual Dictionary* has a specific German idiom followed by its translation equivalent(s) in Portuguese and at least one example sentence of the German idiom in context (see example below in Figure 4). To be able to comply with space constraints and to avoid idiom repetition, cross-references appear within the dictionary's microstructure. Therefore, the user is frequently redirected to other entries, which is a common scenario within bilingual dictionaries.

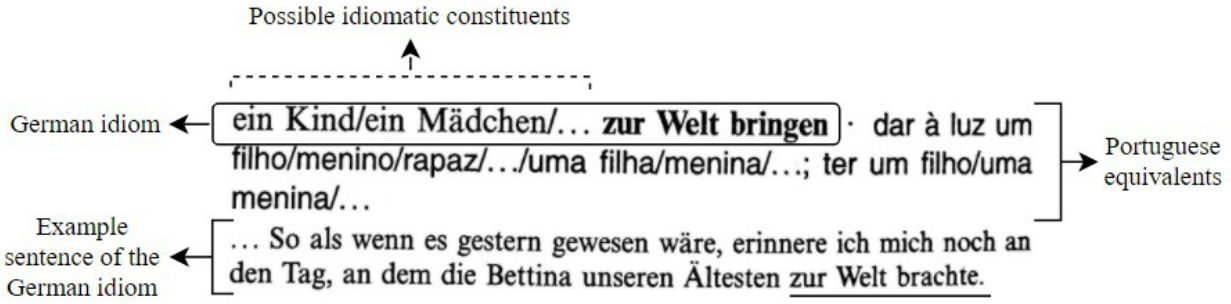

Figure 4 - The *German-Portuguese Idiomatic Dictionary's* entry for *ein Kind/ein Mädchen/... zur Welt bringen*.

However, in this dictionary cross-references are used to establish a relation between idioms but also between example sentences. Usage examples are crucial in any dictionary but are of major importance in such a specialized dictionary because they provide a contextualized use of the idioms, which assist in lexical reception and production, as well as in translation. The cross-references in this resource present one main drawback, namely that the act of remitting from one expression (source) to a quasi-synonymous expression (target) leaves the entry of the source expression devoid of an example sentence (Figure 5). The user only has access to the example sentence(s) of the target expression.

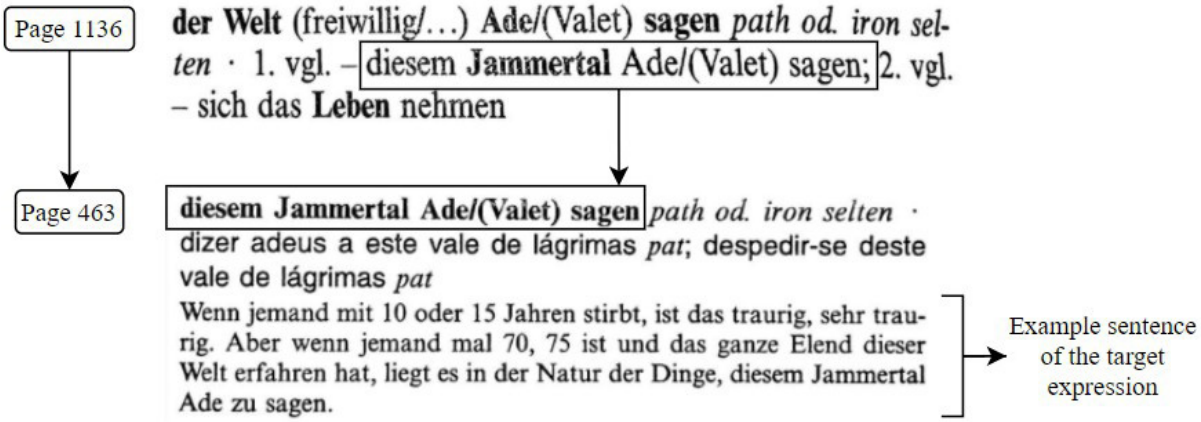

Figure 5 - The entry for *der Welt (freiwillig/...) Ade/(Valet) sagen* with cross-reference to synonymous expressions

Figure 5 shows the methodology used by Schemann for cross-references. The source entry *der Welt (freiwillig/...) Ade/(Valet) sagen* redirects the user to two quasi-synonymous expressions: 1. *diesem Jammertal Ade/(Valet) sagen [to leave/to say farewell to/to say adieu to/to bid adieu to/...this vale of tears*]; 2. *sich das Leben nehmen* [*to take one's life; to commit suicide*]. There is no example sentence for the source expression, only for the target expressions.

In view of the above, merging the *Synonymy Dictionary* and the *German-Portuguese Dictionary* into one electronic resource taking into account possible ways of interlinking the different contents and modelling common access structures to the lexicographic data will make it possible to cater for a wider range of user types and use situations. In addition, it has been possible to pick up on expressions that belong to a specific synset but were not originally included in the *Synonym Dictionary*.

4 Dictionary Encoding

This section briefly describes the preliminary electronic processing tasks undertaken to merge the two dictionaries, followed by the semi-automatic encoding of the lexicographic data using a subset of the

Text Encoding Initiative (TEI) schema for dictionaries. It also identifies the encoding challenges when annotating these dictionaries.

4.1 Preliminary Electronic Processing

This project emerges after previous work on Schemann's dictionaries[20], which produced the first digital version of the dictionaries in XML files with minimal annotation. The first processing step involved converting the data marked up using TUSTEP (Tuebingen System of Text Processing Tools) into XML with minimal descriptive tags. The second step was dedicated to linking the data between the dictionaries: a Perl script was created to match the German idioms in the *Synonym Dictionary* with the German idioms in the *German-Portuguese Dictionary.* With every match, the synset code in the synonym dictionary was added to the respective bilingual dictionary entry. In order to be able to automatically pick up expressions that were not perfect matches due to differences in paradigmatic lexical units, specific patterns were identified to assist in the matching process. Since the bilingual dictionary consists of almost twice the number of idioms in the synonym dictionary, many German idioms do not have a synset code. One very important result of this process is that the Portuguese equivalents have also automatically been assigned a synset code.

Figure 6 shows the first entry of the *German-Portuguese Idiom Dictionary* in XML with minimal annotation, linked to a synset code from the *Synonym Dictionary*.

```
<eintrag>
    <st>A</st>
    <de><b>das A anschlagen/angeben</b> <i>Musik</i></de>
    <po>dar o lá</po>
    <bs>Mein Gott, ist die Geige verstimmt. Es scheint, du hast nicht das A angeschlagen,
    sondern das H!</bs>
    <kat> Dc10.22 </kat>
</eintrag>
```

Figure 6 - First entry of the *German-Portuguese Idiom Dictionary* with minimal XML annotation.

A minimal entry, `<eintrag>`, included a headword, `<st>` (*Stichwort*), the German idiom annotated with the `<de>` (*deutsch*) element, highlighted using the bold element `<b>`. The Portuguese equivalents were encoded with `<po>` (*portugiesisch*) tag and the example sentence using the `<bs>` (*Beispiel*) element. There are other elements like notes, encoded using `<note>`, and usages (using the italic tag `<i>`) and other non-descriptive markup (e.g. #s for letter-spacing). The element `<kat> (Kategorie)` provided the synset code which enabled linking the dictionaries using eXtensible Stylesheet Transformations (XSLT)[21].

20 The processing tasks were performed by Idalete Dias, Center of Humanistic Studies, University of Minho, and José João Almeida, Department of Informatics, University of Minho.

21 XSLT is a World Wide Web Consortium Recommendation, that defines a language to manipulate and perform structural transformations on XML documents: https://www.w3.org/TR/xslt-30/

4.2 Reencoding into TEI

The Text Encoding Initiative (TEI, 2021) is the result of the work of a consortium of individuals in the creation of an XML schema for Digital Humanities. It supports the encoding of diverse types of resources, ranging from simple prose or poetry up to dictionaries, mathematics, or even linguistic corpora.

While there are diverse formats to encode dictionaries, Chapter 9 of the TEI-P5 schema is one of the most used and was chosen for this project. The conversion of the original encoding into TEI was performed with a set of handwritten rules, compiled in a GNU Make makefile. This allows the use of the rules over the different documents, and to easily edit the encoding process to test new rules.

The selected TEI elements, presented in Figure 7, are suitable for modelling the structure of the dictionary entries and for describing the semantic nature of the components. As long as it is accurate and relevant, the more granular the annotation we apply to the data, the more information the user will be able to retrieve from a given search query.

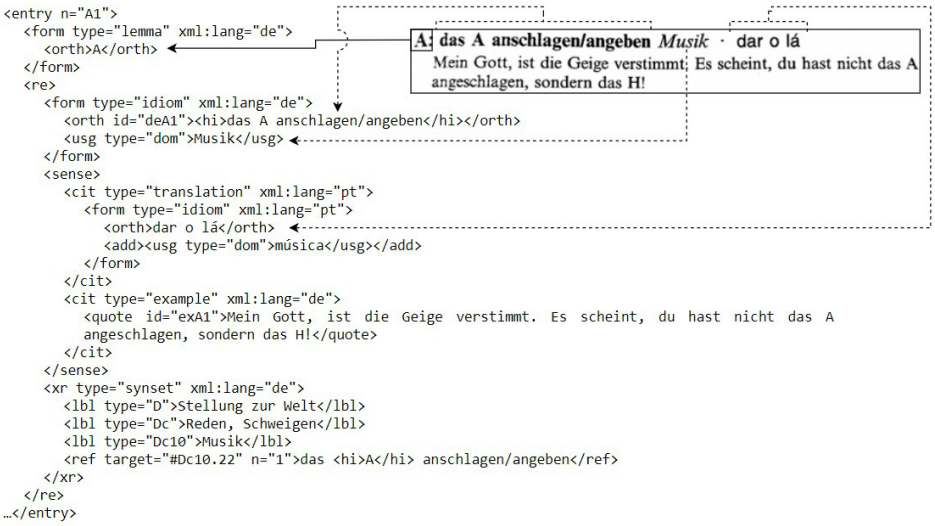

Figure 7 - The first entry of the dictionary in its print version, compared to its annotation using XML/TEI after linking the dictionaries.

Given the specificity of the document, the TEI schema was extended with some custom annotations to suit the dictionaries' content. For instance, as seen in the example above, the Portuguese idiom "*dar o lá*" [*to hit A*] could also have its usage displayed, which happens to be the same as for its German equivalent. Although the following TEI elements are not allowed as proposed below by TEI's dictionary schema, it is our understanding that these rules best represent the structure and semantics of the dictionary entries:

- <add> must contain the citation element <cit>, in cases where we want to improve the dictionary's content by adding more examples;
- <add> must contain <xr>, when it is clear that a cross-reference is missing;
- <add> must contain <usg>, if an accurate and relevant usage is identified to provide more information to an idiom or an example;
- <cit> must contain <del> and <del> must contain <form>, to delete some clear idiom repetition within the Portuguese equivalents;
- <quote> must contain <usg>, where examples also provide usage information.

These extended specifications allowed us to preserve the dictionaries' original content and distinguish it from new information that was added in this phase of the project. Regarding the last addition, it is of utmost importance to show usage information because an idiom may have several meanings and the nuances between them are given through contextualized examples of their usage situation (e.g. formal, ironic).

4.3 Identified challenges

Two particular challenges emerge from the cross-reference notation in the *Bilingual Dictionary*, one being, as already mentioned, that sometimes an idiom lacks an example sentence because the user is redirected to another entry; and secondly, the user may find more or fewer idioms that share synonymy relationships while going through the cross-references, meaning that, depending on the starting point, the user may never come across some idioms because this linking system is not bidirectional. Furthermore, German idioms that only appear in the *Synonym Dictionary* do not have example sentences because there were idioms that did not match with the ones in the *Bilingual Dictionary* due to slight differences (e.g. punctuation or other suggested co-occurrence lexical units) or German idioms that have no Portuguese equivalents. These issues stem from the print version and we are working to improve this with the aim of providing all idioms with example sentences and Portuguese equivalents.

Further challenges appeared after the TEI encoding of the dictionaries, namely, idioms that point to a synset without belonging to it, when they should; and, sometimes, two or more idioms appearing together separated by a slash, that may not be clear to the user. Therefore, we have identified further annotation improvements within the idioms to provide better assistance in idiom learning.

In regard to the search interface, querying by concept should be improved by implementing an autocomplete function to the search box. If the user chooses to type in the search box, rather than explore the semantic field structure to check for the available concepts, the query may return no results, which is unproductive and frustrating for the user. Suggesting keywords in the search box will not only assist in writing without spelling errors but also instantly show which words are identified as concepts according to the *Synonym Dictionary*.

5 Web Application

In this section, we briefly describe the technologies that are supporting our prototype and share our ideas on how to develop an intuitive and versatile interface for querying an Idiom Dictionary.

5.1 Supporting Technologies

Considering the choice to encode the dictionary using XML technology (namely using the TEI schema), the prototype was developed with technology focused on the World Wide Web Consortium (W3C) standards, namely supporting XPath and XQuery for the querying of the XML dictionary files, and HyperText Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript (JS) for building the interface.

With this goal in mind, a set of different document-oriented databases were analyzed, and eXist-DB was chosen. eXist-DB is an open-source document-oriented database that treats XML documents as first-class citizens in its environment. This means that the encoded files can be imported directly into the database, without any kind of extra-processing, and that all the database environment is prepared to deal with such documents.

eXist-DB works not just as a database, but as a full development framework. It allows the development of web applications with XQuery and supports direct XML transformations through eXtensible Style Sheets Transformations (XSLT).

Thus, the XML documents encoded in TEI (one per letter of the alphabet) were imported into the database, and the full interface, discussed in the next section, was developed using XQuery and web technologies (HTML, CSS, JS).

5.2 Prototype Development Guidelines

The creation of an online dictionary requires well-thought considerations related to its "content, presentation, users and usage" (Klosa, 2013), hand in hand with Tarp's lexicographical function theory (2014), which states the following:

- *Dictionaries are utility tools*
- *designed for consultation*
- *and produced with the genuine purpose of meeting punctual information needs,*
- *which specific types of potential user*
- *may have in specific types of extra-lexicographic situation,*
- *by providing access to carefully prepared data*
- *from which the users can retrieve information*
- *which can subsequently be used for different purposes.*

Therefore, the search interface must focus on the targeted users - German and/or Portuguese advanced learners and translators - and their information needs.

Given the dictionaries' content, we have identified the user's needs in communicative situations and cognitive situations. 'Communicative situations' refer to "text reception and production in the mother tongue or a foreign language, translation from and into the mother tongue, and text revision" (Tarp, 2014), while 'cognitive situations' occur when there is a need for knowledge acquisition. The situations covered by the search tool range from Portuguese text production, German-Portuguese translation, Portuguese-German translation, German text reception and text production and Portuguese text reception to users with German or Portuguese as their mother tongue. This makes a total of 12 identified lexicographical situations for which we developed the 'search methods' for idiom comprehension, assistance in writing and translation, as well as the advanced search for more focused results.

In regard to text reception needs (Leroyer, 2018), even though we do not have a definition of the idioms, we can still assist in idiom comprehension (also a cognitive situation) by providing the idiom's synset, whenever it is possible, and/or one or more contextualized examples.

Though the search tool provides results for each of the use situations listed above, it's mostly directed towards tasks related to text production, which can also involve translation and idiom comprehension (Fuertes-Olivera & Bergenholtz, 2018). While full-text search is always provided and the user may search for one or more words that are part of an idiom or identified as a concept, he/she can also explore the hierarchical structure of the concepts, performing an onomasiological search that mirrors the *Synonym Dictionary*'s approach, as intended. This is particularly useful if one is looking for idioms related to a specific concept for writing assistance or, for instance, looking for synonyms while performing text revision.

The search interface also assists in translation from German to Portuguese and its reverse, providing all the information available related to an idiom in both languages.

If the user can express the need for information about a language problem, then the interface can show the results according to the selected search methods, which are driven by the user's lexicographical situation.

Last but not least, the interface design was created considering the recent research in interface and usability design for online dictionaries, namely carried out by the Institut für Deutsche Sprache (IDS), Wolfer et al. (2018a, 2018b), Fuertes-Olivera (2016), Bergenholtz et al (2015), Lew & de Schryver (2014), Müller-Spitzer et al (2011, 2014b), and by analysing the online dictionaries mentioned in Section 2, among others.

5.3 Developed Prototype

The current prototype for the search interface (Figure 8) has the following layout design:

- a left side menu to search for idioms by concept, preserving Schemann's original conceptual system, which can be open (compass button) and closed at any time;
- a top menu with the tabs *Ajuda* [Help], *Sobre* [About], *Contacto* [Contact] and a globe button to select the interface language (Figure 8 shows the Portuguese interface[22]);
- a search box, which is always displayed to perform a full-text search at any moment. Note that this search is performed in the entry idioms and in the examples. It is also possible to restrict the search looking for results in *Português* [Portuguese], *Alemão* [German] or searching by *Palavras exactas* [exact words]. It is also possible to restrict the search on specific parts of the dictionary entries: *Em tudo* [all] or only in idioms or concepts and a button '*+ Tipos de Pesquisa'* [+ Search Methods] to focus on the results to assist in writing, idiom comprehension, translation and also advanced search with boolean operators;
- the search results area, showing the dictionaries' entries according to the performed query, which can provide information related to the German idiom, its usage, its Portuguese equivalent(s), the associated concept and synset, as well as suggestions to check for more idioms within the same concept and its higher-order concepts: for example, as shown in Figure 8, there are hyperlinks to search for more idioms in the concepts Dc10 - *Musik* [Music] < Dc - *Reden, Schweigen* [Speaking, Silence] < D - *Stellung zur Welt* [Attitude to the world].

Figure 8 - Search result for '*dar o lá*' [*to hit A*], showing 1 result.

6 Conclusions and Future Work

There are more offers in the market of lexical resources prepared for the decoding task, than for encoding. Studies reveal that "the use of a dictionary for assistance with writing is very high (...) [and] many lexicographers recognise users need dictionaries to look for a word that has escaped their memory although they remember the concept" (Sierra, 2000). For this kind of task, the *Synonym Dictionary*'s onomasiological structure is very useful, namely given its focus on idioms. However, searching an analogical ideological or synonym dictionary can be difficult for some users who might not know how to search in these dictionaries. Thus, a good solution is to make them available in digital format, with their content correctly annotated using XML/TEI. TEI-encoding adds a layer of meaning with more metadata and better data structure, preserving the original content but also leaving space for improvement. If the dictionaries' content is well-

22 The current prototype is only available with its interface in Portuguese.

annotated with appropriate TEI elements and is well-formed in a solid structure, then the search results will be of more relevance to the user as long as these are presented clearly to assist in idiom learning and/or translation.

The interface design of online dictionaries should break the 'codex layout' that the users know, displaying information in more attractive and interactive ways, to grab the users' interest and attention and give them what they need. This includes the scenarios where not even the users know exactly what they need or are looking for and the search interface should offer suggestions to try to meet what they have in mind. Also, it is important to guide the user throughout the interface whenever it is possible, by offering several search methods, so he/she is not so dependent on his/her searching skills. These search methods are built around the users' needs and will become particularly useful if one identifies his/her needs of information, for instance, assistance in writing a text, understanding an idiom by its synonyms and/or contextualized examples and translating idioms.

We hope that our project provides more direct access to the dictionaries' content and to do so, we will continue to improve it by: i) working on the cross-reference system, ii) adding further annotation within the idioms, iii) adding more keywords related to the top-level and mid-level concepts, so the user will have better chances in finding what he/she needs, and iv) adding examples where there are none. Regarding the examples for Portuguese idioms, these will also be included from *Schemann's Portuguese-German Idiomatic Dictionary.*

Furthermore, a thorough test with the targeted users must be carried out, which may lead to new design specifications.

We believe that our model can later include other languages by encoding four more bilingual Idiomatic Dictionaries of the same author, namely in English, Spanish, French and Italian.

Acknowledgements

This project was funded by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the project UIDB/05549/2020.

References

Bergenholtz, H., Bothma, T., & Gouws, R. (2015). Phases and steps in the access to data in information tools. *Lexikos, 25*(1). https://doi.org/10.5788/25-1-1289

Dias, I. (2010). *Sinonímia - campo semântico - contexto - texto: uma análise sinonímia com particular relevância para as expressões idiomáticas: estudo sistemático e contrastivo*. PhD Thesis. Braga: Universidade do Minho. http://hdl.handle.net/1822/11263

Fuertes-Olivera, P. & Tarp, S. (2014). *Theory and Practice of specialised online dictionaries*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110349023

Fuertes-Olivera, P. (2016). A Cambrian explosion in lexicography: Some reflections for designing and constructing specialised online dictionaries, *International Journal of Lexicography*, Volume 29(2), 226-247. https://doi.org/10.1093/ijl/ecv037

Fuertes-Olivera, P., & Bergenholtz, H. (2018). Dictionaries for text production. In P. Fuertes Olivera (Ed.), *The Routledge Handbook of Lexicography* (pp. 267-283).

Gouws, R. (2010). Outer texts in bilingual dictionaries. *Lexikos, 14* (pp. 67-88). https://doi.org/10.5788/14-0-683

Klosa, A. (2013). 26. The lexicographical process (with special focus on online dictionaries). In R. Gouws, U.

Heid, W. Schweickard & H. Wiegand (Eds.), *Supplementary Volume Dictionaries. An international Encyclopedia of Lexicography* (pp. 517-524). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110238136.517

Leroyer, P. (2018). Dictionaries for text reception. In P. A. Fuertes-Olivera (Ed.), *The Routledge handbook of lexicography* (pp. 250-266).

Lew, R. & de Schryver, G. (2014), Dictionary users in the digital revolution, *International Journal of Lexicography*, Volume 27(4). (pp. 341-359). https://doi.org/10.1093/ijl/ecu011

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, No. 11: 39-41, https://doi.org/10.1145/219717.219748

Moerdijk, F., Tiberius, C., & Niestadt, J.(2008). Accessing the ANW dictionary. In M. Zock, & C-R. Huang (Eds.), 22nd International Conference on Computational Linguistics. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (pp. 18-24). Brighton, UK: One Digital.

Müller-Spitzer, C., Koplenig, A., & Töpel, A. (2011). What makes a good online dictionary? - Empirical insights from an interdisciplinary research project. *Proceedings of eLex 2011*. (pp. 203-208).

Müller-Spitzer, C. (2014a). 11. Textual structures in electronic dictionaries compared with printed dictionaries: A short general survey. In R. Gouws, U. Heid, W. Schweickard & H. Wiegand (Ed.), *Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography* (pp. 367-381). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110238136.367

Müller-Spitzer, C. (Ed.). (2014b). *Using online dictionaries*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110341287

Salgado, A., Costa, R., Tasovac, T. & Simões, A. (2019). TEI Lex-0 in action: Improving the encoding of the dictionary of the Academia das Ciências de Lisboa. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 417-433). Sintra, Portugal.

Simões, A., Salgado, A., Costa, R. & Almeida, J. J. (2019). LeXmart: A smart tool for lexicographers. In I.

Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 453-466). Sintra, Portugal.

Sierra, G. (2000). The onomasiological dictionary: a gap in lexicography. *Proceedings of the 9th Euralex International Congress* (pp. 223-235). Stuttgart.

Tarp, S. (2012). Online dictionaries: today and tomorrow. *Lexicographica, 28*(2012), 253-268. https://doi.org/10.1515/lexi.2012-0013

Tarp, S. (2013). *Dictionaries. An International Encyclopedia of Lexicography: Supplementary volume: Recent Developments with Special Focus on Computational Lexicography.* Gouws, R. H., Heid, U., Schweickard, W. & Wiegand, H. E. (eds.). New York: De Gruyter, Vol. 5.4. pp. 460-468.

Tarp, S. (2014). Dictionaries in the internet era: Innovation or business as usual? (Enrique Alcaraz Memorial Lecture 2014). *Alicante Journal of English Studies / Revista Alicantina de Estudios Ingleses*, (27), 233-261. https://doi.org/10.14198/raei.2014.27.13

TEI Consortium, eds. (2021). Chapter 9: Dictionaries. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.4.2, last modified on 9th April. TEI Consortium. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html (Accessed May 24, 2021).

Wiegand, H. E. (1990). Printed Dictionaries and their Parts as Text. An Overview of More Recent Research as an Introduction. *In Lexicographica, 6,* 1-126.

Wolfer, S., Nied, M., Dias, I., Müller-Spitzer, C. & Domínguez, M. J. (2018a). Combining quantitative and qualitative methods in a study on dictionary use. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek, Proceedings of the XVIII EURALEX International Congress – Lexicography in Global Contexts. Ljubljana: Ljubljana University Press, pp. 101-112.

Wolfer, S., Kosem, I., Lew, R., Müller-Spitzer, C., & Silveira, M. R. (2018b). Web-based exploration of results from a large European survey on dictionary use and culture: ESDexplorer. *Lexikos*, 28, 440-447. https://dx.doi.org/10.5788/28-1-1473

Dictionary References

Schemann, H. (1993). *Deutsche Idiomatik. Die deutschen Redewendungen im Kontext*. Stuttgart/Dresden: Ernst Klett.

Schemann, H., Schemann-Dias, M. L., Amorim-Braun, L., Martins, T., Duque-Gitt, M.J. & Costa, H. (2012). *Idiomatik Deutsch-Portugiesisch* (2nd ed.). Hamburg: Buske.

Schemann, H. (2012). *Synonymwörterbuch der deutschen Redensarten* (2nd ed.). Berlin: De Gruyter