Citation for published version (APA):
Silvestre, J. P., & Villalva, A. (Eds.) (2015). *Planning non-existent dictionaries*. (Dicionarística Portuguesa; Vol. IV). Lisboa - Aveiro: Centro de Linguística da Universidade de Lisboa - Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro.

DICIONARÍSTICA PORTUGUESA - IV

Planning non-existent dictionaries

João Paulo Silvestre & Alina Villalva (eds.)

CENTRO DE LINGUÍSTICA DA UNIVERSIDADE DE LISBOA
UNIVERSIDADE DE AVEIRO

2015

Título | Planning non existent dictionaries

Autores editores | João Paulo Silvestre, Alina Villalva

Coleção | Dicionarística Portuguesa

Edição | Centro de Linguística da Universidade de Lisboa - Universidade de Aveiro

Impressão | Clássica - Artes Gráficas

Depósito Legal | 391177/15

ISBN | 978-989-98666-1-4

Tiragem | 250 exemplares

Apoio | FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

Contents

I. Looking for heritage dictionaries

II. Dictionaries for special purposes

III. The e-lexicography challenge

Presentation

There is an increasing number of dictionary types and lexical search-tools designed to respond to an ever-growing array of user needs, there are those that are used for language-learning purposes, those that serve specialised knowledge requirements, and those that aim to be suited to search meaning in the lexicon, to name but a few.

The quest for innovation, however, is not over and this is what this book shall shed light on. In the autumn of 2013, a conference entitled *Planning non-existent dictionaries* was held at the University of Lisbon. Scholars and lexicographers were invited to present and submit for discussion their research and practices, focusing on aspects that are traditionally perceived as shortcomings by dictionary makers and dictionary users. The topics for debate were intended to be provocative: the identification of dictionary types that have never been developed for certain languages or for a given lexical domain, as well as typological and linguistic problems that may compromise the development of lexicographic projects. We hoped that the discussion would lead to the presentation of problem-solving strategies, especially those related to corpora documentation, information technology and data presentation. We received an incredible response and have had the opportunity to acknowledge several projects that are different in size, novelty and degree of accomplishment.

This conference left both organizers and participants with a memorable experience and the wish to have access to the information on these ongoing projects on a longer lasting basis. As a result, we decided to edit this book, which contains a collection of papers divided in three sections.

The first section is devoted to heritage dictionaries, referring to lexicographic projects that aim to register all the documented words in a language, particularly those that can be described as early linguistic evidence. This kind of project requires specialist training and the allocation of substantial manpower. Consequently, these projects tend to gain a national dimension, which helps to explain why some of them are at a more advanced stage than others.

ROBERTA CELLA's paper focuses on lexicographic beginnings in Italy, presenting early Italian glossaries and wordlists as tools for a broader, non-linear and sometimes chaotic linguistic standardisation. They are neither dictionaries in the modern sense, nor should they be considered as imperfect steps leading to Crusca's dictionary (1612). Cella shows just how deeply Italian lexicography from the first half of the sixteenth century depends on the process of standardisation of the Italian language.

Moving from an analysis of sources to an example of their use in the compilation of a dictionary, CLOTILDE MURAKAWA and MARIA FILOMENA GONÇALVES present the *Historical Dictionary of Brazilian Portuguese* project. This reference work, based on a database with approximately 10 million entries, aims to document the lexical variety Brazilian Portuguese is built upon, between the 16th and 18th centuries. They discuss some corpus selection issues, such as combining manuscripts and printed texts, produced in very diverse conditions over more than three centuries, which raised philological and linguistic problems.

IVANA FILIPOVIĆ PETROVIĆ reports on the completion of a descriptive dictionary of the Croatian Literary Language. This dictionary, which Julije Benešić began compiling in 1949, was left unfinished after the lexicographer's death in 1957. The excerpted material used in the dictionary was stored in the Linguistic Research Institute of the Croatian Academy of Sciences and Arts, and in 2008 an extensive project began to complete the dictionary. In this paper, she presents the main typological characteristics of Benešić's dictionary, which make it a unique accomplishment of Croatian lexicography. Ivana reveals, through a number of examples, Benešić's lexicographic methods, as well as lexicographic solutions from the new volumes and guidelines for a more systematic approach in the future.

To conclude this section, MARIA-PILAR PEREA describes recent developments in the study of Catalan lexicography. The *Lexdialgram* project aims to recover and digitize nineteenth-century dictionaries, providing important data that can help us to learn more about the status of the Catalan language and its linguistic variation in a period when dialect studies as such were largely ignored.

The second section is devoted to dictionaries for special purposes and it gathers papers that describe innovative lexicographic projects. One of the challenges faced by modern dictionaries has to do with the need to account for the multiple meanings that words display in context when combined with each other. ELENA DE MIGUEL, currently in charge of the research project *Multilingual Electronic Dictionary of Motion Verbs*, presents a dynamic dictionary of minimal verbal definitions, from which the so-called periphrastic, figurative or metaphorical uses can be derived once combined with the information provided by other words in context. It is frequently assumed that there is a literal, canonical meaning, and that all the other interpretive possibilities are figurative or metaphorical. However, in her paper, Elena de Miguel claims that the generation and interpretation of different word meanings (both literal and non-literal) is governed by the same general and regular principles, assuming the existence of *agreement of sub-lexical features* and the existence of minimal word definitions. Since different meanings can be deduced from the minimal definition, enumerating them in the lexical entries of nouns, verbs and adjectives become unnecessary and the task of the lexicographer becomes much easier.

ALINA VILLALVA and JOÃO PAULO SILVESTRE question the sources and methods to compile a morphological historical root dictionary for Portuguese. Although Portuguese dictionaries generally include information about etymology and morphological structure, they do so quite inconsistently, for they are generally the product of the accumulation of material that can be found in previous dictionaries. This *modus operandi* leaves no room for a systematic analysis of the set of words and word families that form each dictionary's entry list. *Roots* is a research project that aims to produce a prototype for a specialized dictionary that intends to be a useful tool for linguists, lexicographers, translators and language teachers. More specifically, *Roots* is intending to provide a thorough description of word families, considering their presence in the Portuguese lexicon, from a semantic and a morphological point of view, both in diachrony and in their contemporary usage.

Some domains of the lexicon are mostly excluded from general lexicography. JEAN-LOUIS VAXELAIRE identifies the drawbacks of traditional lexicography regarding proper names. Though there are specialist proper names dictionaries, their lexicographic treatment is not appropriate, focusing on etymology or descriptions of famous bearers of the name, rather than a description of meaning. As they are elements of culture, there is need for a more language-oriented work for students and translators (i.e. providing a phonetic transcription, gender, etc.).

The aim of IVO FABIJANIĆ's paper is to provide some theoretical and methodological models in preparing both a macro- and micro-structural framework for the compilation of a future bilingual, specialized and explanatory dictionary of abbreviations in linguistics, one that has never been developed for the Croatian language. The dictionary would cover both core areas of linguistics (phonology, morphology, syntax, semantics and pragmatics) and its interdisciplinary areas (e.g. sociolinguistics, psycholinguistics, neurolinguistics, ethnolinguistics, computational linguistics, etc.).

CRISTINA FARGETTI discusses the lexicographic description of kinship terms in Juruna, an indigenous language from Brazil. In linguistic studies, sometimes the complexity of kinship systems is not made clear, raising many questions about the use of the terms presented. For instance, the prohibition of incest is commonplace, yet relative, given what is considered incest varies from culture to culture. This prohibition is the basis of kinship systems, which differ in each nation. She reflects on how to recollect and analyse data regarding the cultural knowledge that a linguist should have. To conclude, she presents sample entries for a Juruna dictionary that is in the process of being completed

Pedagogical dictionaries are also special purpose dictionaries. In their work, MARIA EGIDO VICENTE, MANUEL FERNÁNDEZ MÉNDEZ and MÁRIO FRANCO BARROS describe *Diconale*, a project aiming to develop a conceptual online bilingual dictionary of contemporary Spanish and German. This bidirectional bilingual dictionary is primarily designed as a reference tool to be used from B1 level onwards of the Common European Framework of Reference for Languages in the area of ELE (Spanish as a Foreign Language) and DaF (Deutsch als Fremdsprache). Access to the dictionary is both onomasiological and semasiological. It is based on a model of lexical, modular and multilateral description that allows for the presentation of the information of verbal and deverbal lexemes for each language separately, as well as from a contrastive perspective. Therefore, the central point of interest is the study of both syntagmatic and paradigmatic relations. In a second paper, MEIKE MELISS & PALOMA SÂNCHEZ HERNÂNDEZ present some theoretical and methodological aspects of the *Diconale* project, in which they expose some difficulties that involve the onomasiological-conceptual approach.

The last section in this volume provides an overview of contemporary e-lexicography projects. ÁLVARO IRIARTE examines the process of reverse search in electronic dictionaries. The "ontological search" is an underused function in electronic dictionaries, allowing users to track information that is not explicit in the dictionary. Using a set of rules, it is possible to calculate plausible lexical-conceptual relations. As an example, he proposes to calculate lexical-conceptual relations using the transitivity relation in hypernym-hyponym relations.

Finally, Mauro Villar presents a survey of the lexicographic corpora that a team of lexicographers have developed under his supervision for the *Dicionário Houaiss da Língua Portuguesa*. This dictionary was originally compiled based on manual selection of illustrative quotations, since it was initially developed when there were no other methods available in Portugal or Brazil. However, as soon as was possible, lexicographers engaged with corpora resources, that are far more straightforward and practical to work with, although for the time being they are not so rich on historical information about the language. Villar refers to online media as a way of improving previous versions of the Houaiss dictionary, documenting new uses and senses, reading rare works that had previously been out of reach, adding to what was already gathered and organised.

Lexicographers seek to add to the dictionaries on offer in order to include new languages or new vocabulary fields. In addition, lexicographic critics challenge the feasibility of dictionaries that fail to implement validated theoretical models. In the case of language dictionaries, they often lack lexical documentation and description, and, more often than not, lexicographers try to meet users' expectations, rather than provide for their needs. Just like the conference that preceded it, this book was intended as an opportunity to discuss pending issues and a chance for less divulged dictionary projects to reach a larger audience and thus a chance for debate. We hope to have accomplished these goals, even if only on a smaller scale.

João Paulo Silvestre & Alina Villalva

I

Looking for heritage dictionaries

Before the Crusca's dictionary: Italian glossaries and wordlists in first half of sixteenth century

Roberta Cella

1. Introduction

For modern scholars, a dictionary is a scientific representation of a lexical-semantic system, both diachronic (how the system was and how it has changed) and synchronic (how is the system and how it could be, or specialized or bilingual, but anyway a representation of a lexical-semantic system, that is obviously related to morphology and syntax, but is regarded as autonomous from any of them. Modern scholars have a descriptive (not a prescriptive) aim for creating dictionaries: they discuss what kind of corpus could offer a better representation of reality, they take care of the selection of headwords, the coherence of definitions, the distinction and ranking of senses, the choice of examples – in order to represent in the dictionary-microcosmus a real mirror of a macrocosmus. Even if they are aware that their representation is a partial and autonomous conceptualization of the lexico-semantic domain of language, they tend to forget that, at the very beginning of modern lexicography, linguistic perspectives, conditions and aims were different. In particular, they forget that lexico-semantic autonomy was not a starting point but a slow achievement[1].

In this paper, I will present a survey of Italian glossaries and wordlists from the first half of the 16th century. In its beginnings, lexicography was strictly connected to grammar (regarded as standard morphology), encyclopaedia (a world classification by words), literary genres theory and philology (edition of "corrected" texts). Shortly: wordlists and glossaries were tools for a broader, not-linear, and sometimes chaotic linguistic standardisation.

[1] See: «il riconoscimento della specificità e dell'autonomia del lessico non è un punto di partenza ma una delle conquiste che si guadagnano nel corso del suo *iter* cinquecentesco» ('the acknowledgement of lexical specificity and autonomy is not a starting point but an achievement that occured by sixteenth century lexicographers' (personal translation), Manni 1991: 69).

2. Glossaries and history of Italian language

Italian glossaries and wordlists of the first half of sixteenth century show how apparently aseptic linguistic tools can react to the historical-cultural context in which they originated (on early Italian lexicography vd. Olivieri 1942, Poggi Salani 1986, Tancke 1984, Trovato 1988, Manni 1991, Lobodanov 1999, Marazzini 2009: 55-126, Paccagnella 2013). First of all, we have to reject a misunderstanding that until recently conditioned our approach to old dictionaries: they can not be regarded in a teleological way as imperfect forerunners to the *Vocabolario degli accademici della Crusca* (1612), they must rather be analysed in their own context. For this purpose we will consider self-standing works (Liburnio 1526, Alunno 1539, Alunno 1543, Acarisio 1543, Alunno 1548) and glossaries (sometimes called *dichiarazioni,* i.e. 'explanations') added to editions of Giovanni Boccaccio's *Decameron* (Minerbi 1535, Brucioli 1538, Dolce 1541, Sansovino 1546, Ruscelli 1552). My overview will leave apart Fabricio Luna's *Vocabulario* (1536) and some other minor works, such as Giovanbattista Verini's *Dictionario* (1532). Luna's *Vocabulario* is a marginal work, compiled in Naples, far from the mainstream language standardisation (and, in despite of the title "dictionary of 5.000 Tuscan words", rather prone to offer a mixed language). Verini 1532 is a basic glossary, compiled for beginner students (cf. Poggi Salani 1986: 54-56, 66-69, Marazzini 2009: 64-66, 68-72).

All the works considered here, both self-standing glossaries and editions of the *Decameron* with glossaries, were published in Venice, which at the time was the heart of the linguistic standardisation powered up by Pietro Bembo's *Prose della volgar lingua* (1525). This author projected a communal language that today we call Italian, on the basis of thirteenth century Florentine variety. That is the reason why contemporaries called it Tuscan or, as Bembo, volgare i.e. 'not Latin'. Bembo's rule follows Petrarch for poetry and Boccaccio for prose, and leaves aside Dante, too various to be imitable.

Venice was the heart of the linguistic standardisation especially for its flourishing publishing activity. Editors and publishers had been aiming for standardised editions (*corrette* 'normalised, regulated') that could be sold in broader markets. They spread a standard language, indeed based on a literary archaic Florentine model recommended by Bembo (cf. Trovato 1991). Only Acarisio's *Vocabolario* is not connected to Venice: it was published at Cento, near Ferrara, by a non-professional and quite isolated author.

Furthermore, all these works deal with canonical authors, regarded by then as linguistic models: Petrarch and Boccaccio (according to Bembo) and Dante, who was neglected by Bembo, but highly regarded by sixteenth century lexicographers.

The *osservazione* ('examination', i.e. textual analysis) of the fourteenth century masterpieces (Dante's *Comedy*, Petrarch's *Rerum vulgarium fragmenta* and Boccaccio's *Decameron*) produced glossaries and wordlists, with the double aim of explaining the ancient Tuscan words to non-Tuscan speakers and giving an imitable model of literary language and style, according to the principles of Renaissance.

The very feature shared by all these glossaries is their *non-autonomy,* settting them apart from Crusca's *Vocabolario*. In effect, from a conceptual and even material point of view they depend on other elements: on grammar (Liburnio's *Le tre fontane* [1526], Acarisio's *Vocabolario* [1543]), on the spreading of linguistic models for poetry and prose (Liburnio's *Le tre fontane* [1526] again, Alunno's *Osservazioni* [1539] and *Ricchezze* [1543]), on philology – regarded both as textual comprehension and fixing of a "correct" text (glossaries included in editions of the *Decameron*). Even Alunno's *Fabrica del mondo* (1548), the first Italian methodical dictionary, did not consider lexicon by itself. The lexicon is related to encyclopaedia, one one hand, in ten thematic categories (God, sky, world, elements, soul, body, human being, quality, quantity, hell), and on the other hand, the eleventh section exhibits a connection of the lexicon to grammar. This section is devoted to *particelle* (i.e. grammatical elements, such as conjunctions, adverbs, prepositions, articles, interjections).

In sum: these works are not dictionaries as we see them nowadays, with an autonomous value and *raison d'être* given only for the sake of a well-organised representation of lexicon and semantics. They are tools planned for a very special purpose – declaring and spreading the standard language. They played a similar role to publishers and editors, and successfully contributed to standardisation of Italian language. This purpose explains their external features, which seem so far from modern dictionaries that they are often accused of primitiveness.

3. Glossaries and grammar

In Niccolo Liburnio's (1474-1557) *Le tre fontane* (1526) - i.e. 'three springs' of Italian language: Dante, Petrarch and Boccaccio - grammar establishes the work's very structure[2]. *Le tre fontane* - three books, each one devoted to an author – offers simple wordlists, divided into parts of speech: verbs, adverbs, interjections, pronouns, prepositions, conjunctions, relative pronouns and adjectives, irregular nouns («nomi eterocliti») and other nouns, idioms («modi figurati», in which Liburnio lists both true idioms and onomasiological periphrasis, for ex. «Et noi movemmo i piedi inver la terra» 'and we moved our feet to the place' Dante, *Inferno* IX 104, is given under headword *andar* 'to go'). Headwords are not lemmatised (they remain in the original inflectional form) nor explained, but rather followed by an one-line example with reference. As in grammars, many sections of *Le tre fontane* begin with a short definition of the part of speech they are devoted to. See for instance section "conjunction" in book I, devoted to Dante: «Congiuntione è parte di oratione non declinabile, il cui ufficio è la forza e l'ordine dell'altre parti della compositione insieme copulare» ('conjunction is an invariable part of speech. It

[2] «*Le tre fontane* [...] rappresenta nel modo più esplicito una situazione di dipendenza del lessico nei confronti della grammatica. Il lessico è incapace di trovare in se stesso un elemento ordinatore e arriva a darsi una sistemazione solo inserendosi negli schemi della categorizzazione grammaticale» ('*Le tre fontane* [...] explicitely represents
the redependence of the lexicon from the grammar. Lexicon cannot find in itself an order and can be arranged only if inserted into grammatical categories', Manni 1991: 70). On Liburnio cf. Poggi Salani 1986: 52-54; Marazzini 2009: 58-64).

connects and puts in order the other parts of speech', Liburnio 1526: 30v). We can conclude that lexical items are structured as in a grammar.

In Alberto Acarisio's *Vocabolario, grammatica e ortografia della lingua volgare* (1543) lexical items and grammar are better identified, but likewise complementary (cf. Poggi Salani 1986: 69-71, Marazzini 2009: 66-68, and especially Trovato 1988). In 1536 Acarisio published a successful *Grammatica volgare*, in which he spread and simplified the Bembo's *Prose della volgar lingua* (which was an extremely complex work, structured as a classical dialogue, unsuitable as a reference book). The *Vocabolario* included the *Grammatica volgare* (as a section) and a brief treatise on orthography (Acarisio 1543: 1r-25v, 26r-27v). Even in the glossary (Acarisio 1543: 28r-315v) there are systematic annotations on morphology. On the contrary, definitions consist of the simple Latin translation. See for instance the entry *appartengo* 'I belong' (with headword in first person present tense, as in Latin dictionaries, not in infinitive)[3]:

> *Appartengo* è latino *pertineo*, segue la regola di *tengo*: noi diciamo *i parenti appartenersi*, cioè tenersi d'una parte medesima. BOC. nel *prin(cipio)* "& quelle, che più gli appartenevano, piangevano" ('*Appartengo* is equivalent to Latin *pertineo*, and has the same inflective rule as *tengo*: we say *relatives belong to each other*, i.e. they are of the same group. Boccaccio, in the preface: "women from his group cried"', Acarisio 1543: 44v).

Whereas the meaning is limited to the simple comparison with Latin *pertìneo*, the focus is on morphology, as the annotation «segue la regola di *tengo*» 'it has the same inflection as *tengo*' reveals, on collocations - as we say today - (*parenti* 'relatives' *si appartengono* 'are of the same part, belong to each other'), and on Boccaccio's authority (with an example elicited by *Decameron*).

For Acarisio, the main focus is always on grammar, especially on morphology; in the entry *empio* 'I fill up', after the simple meaning equivalence with Latin *ìmpleo*, Acarisio offers the whole verbal paradigm («*tu empi*, *quegli empie*, *empieva*, nel te(m)po passato *empiei*, et ne la terza p[e]rsona *empiè*, nel futuro *empierò*, nel soggiuntivo *empia*, *empierei*, nel'infinitivo *empiere*, *esser empiuto*, nel gerundio *empiendo*»), with a new morphological annotation at the end, after examples by Petrarch and Boccaccio: «i suoi composti lo seguono, come *adempiere*, *compiere*, *riempiere*» ('its compound words, such as *adempiere*, *compiere*, *riempiere*, follow its inflection', Acarisio 1543: 122r). Note that the basic form is infinitive in cross-reference, whereas it is first person present tense in headword.

We can recognise the same attention to phono-morphology and orthography — that were main issues in standardisation of Italian — through all Acarisio's work:

[3] Quotations from sixteenth century works are ligthly modified: I integrate modern interpunction, italics and inverted commas for quotations from Dante, Petrach and Boccace; abbreviations are resolved in brackets. In square brackets I insert explanations or integration to facilitate the reading.

> *Albero* & *arbore* si dice questo di genere di femina & di maschio, quello di maschio solamente (‘*Albero* & *arbore* [‘tree’] people use the last one as feminine and masculine noun, the fist one only as masculine’, Acarisio 1543: 37v).

> *Lascio* da *laxo* latino: alcuni non vogliono che per doppio *ss* si possa scrivere ma per *sc*, il che mi piace per essere così la pronontia thoscana, ma chi ne’ versi per due *ss* lo scrivesse non errerebbe, per ciò che il Petr(arca) in rima l’ha detto (‘*Lascio* [‘I leave’] is from Latin *laxo*; some don’t approve writing it with double *ss* instead of *sc*, and I agree because of the Tuscan pronunciation, but writing with *ss* is not a fault in poetry because even Petrarch used it as rhyme’, Acarisio 1543: 173r).

4. Glossaries and linguistic models for prose and poetry

One of the ground principles in Bembo’s *Prose della volgar lingua*, so deep-rooted as to influence Italian literary language up to the nineteenth century, is the distinction between language of poetry and language of prose. This distinction was not new, since it had been implicitly practiced by poets in all ages; the novelty comes from its theoretical explicit expression, up to the codification of allotropes such as *due* / *duo* ‘two’, *niuno* / *nessuno* ‘nobody’, *dee* / *debbe* e *deve* ‘he, she should’, *sparso* / *sparto* ‘scattered’, *aperse* / *aprì* ‘he, she opened’, *vederò* / *vedrò* ‘I will see’, *udirò* / *udrò* ‘I will heard’, *fo* / *faccio* ‘I make, I do’ : the first elements of each pair was considered to be typical of prose, the second part was typical of poetry.

The previously mentioned eleventh section of Alunno’s *La Fabrica del mondo* (1548) is devoted to «particelle» ‘particles’, i.e. function words or grammatical words, listed in alphabetical order. In the entry *nondimeno* ‘nevertheless’, after the Latin equivalent and a list of synonyms, Alunno specifies that «è voce più delle prose che del verso» (‘it is a word more suitable for prose than for poetry’, Alunno 1548: 254r). Recommendations like this are found in early sixteenth century glossaries, as evidence of loyalty to Bembo’s rule of linguistic peculiarity of prose and poetry, and of educational purposes for “comporre bene” (‘correct writing’).

Niccolò Libunio, in his preface (*La cagione della presente opera* ‘Reason for this work’, Liburnio 1526: 4r-9r), explicitly declared:

> ardisco dire cosa essere di favore e ornamento mirabile, se non nel domestico favellare nel scrivere di lettere almeno e in altre componiture dell’uno e dell’altro stile, haver in pronto cotai forme della volgar eloquenza bellissime (‘I affirm that it is useful and fair to have such beautiful words available, if not for informal speech at least for literary writing and for composing in both styles [i.e. in prose and in poetry]’, Liburnio 1526: 5v-6r).

However, the loyalty to Bembo’s rule was tempered by modern usage, especially in Acarisio (Poggi Salani 1986: 69). Forcing selective principles of Renaissance classicism, he included Dante into canonical models (as Liburnio 1526 did, cf. Paccagnella 2013: 55) and rejected unused words, even if used by Dante, Petrarch and Boccaccio. In his preface *A’ lettori* (‘to the readers’) Acarisio wrote:

> per che ho notato alcuni vocaboli da' nostri scrittori usati che hoggidì sono da schifare, vi priego ben considerarli, e tutti quelli che a questo tempo non sono in uso lasciare ('because I noticed that some words used by our authors should be avoided today, I ask you to carefully evaluate them and to reject all those that in these days are not used', Acarisio 1543: non-numbered page).

Doing so, Acarisio established, in facts more than in theory, a principle that some years later would be developed by Benedetto Varchi and Lionardo Salviati: example of ancient Tuscan writers must be submitted to contemporary usage.

5. Glossaries and philology

Glossaries and wordlists added to editions of Giovanni Boccaccio's *Decameron* represent a very interesting lexicographical experience. First of all, they demonstrate the importance of printing correct texts for the spread of standard language and, then, they present unusual typologies of lexicographical tools that will have negligible subsequence, at least in the official, 'national' lexicography.

Lucilio Minerbi composed in 1535 the first glossary printed together with an edition of Boccaccio's *Decameron*. It is a kind of alphabetical word index with desultory additions of a synonym (introduced by *ciò è* 'i.e.') and references to page and line in which the item occurs (Marazzini 2009: 56-57). The synonyms explain items perceived as ancient (see for instance «*membranza* ciò è *ricordanza*», «*nimistà* ciò è *nimicitia*») or for some reason difficult («*malvagio* ciò è *cattivo*»), but the general criteria of annotation remain unclear — in his preface, Minerbi keeps things vague: «i vocaboli seco(n)do l'alphabetico ordine [...] dove fie bisogno maggiore co(n) somma dilige(n)za dechiarati» ('alphabetically ordered words [...] are carefully explained whenever it is necessary', Minerbi 1535: *Alli valorosi giovani*). It is remarkable that some added synonyms are of Northern origin (especially Venetian), which is revealed by phonetics and morphology («*macino* ciò è *il massinar*», «*mescolare* ciò è *misciare insieme*» «*mica* ciò è *miga*» «*munaio* ciò è *mulinaro*») (Poggi Salani 1986: 59; Richardson 1993: 98). Since Minerbi was from Rome, Northern words were taken from Liburnio 1526, as also other evidences suggest (Paccagnella 2013: 49-50). Sporadically, concise grammatical information is given: «*messer lo frate*, & non *il*» (on the use of the ancient form of the article after consonant), «*nocq(ue)* p(assa)t(o) di *nocere* [..] *noce* la terza persona del presente» (on the declension of irregular verb *nuocere* / *nòcere* 'to damage'). In his preface *Alli valorosi giovani e all'amorose donne* ('to valiant youth and amorous ladies') Minerbi declared the aims of his work, explaining the meaning of unknown words and supporting 'youth and ladies' in their own composition in prose in imitation of Boccaccio:

> sove(n)te leggendole [= le «favole» di Boccaccio stampate a cura di Minerbi] [i giovani] potra(n)no co(n) agevolezza co(m)porre et correttamente scrivere, & la forza del vocabolo da loro altrimenti no(n) intisa apertamente conoscere ('often reading them [= Boccaccio's tales in Minerbi's edition], [youths] could easily compose and correctly write, and clearly understand the meaning of unknown words') (Minerbi 1535: *Alli valorosi giovani*).

Minerbi's examples were followed three years later by Antonio Brucioli (1538), albeit with a different structure: he gives lexical annotations — sometimes closer to a commentary than to a glossary — at the end of each tale, with a general index at the end. In 1541 Ludovico Dolce proposes the same structure, but with shorter notes (Richardson 1993: 98-100).

Much richer, in terms of lexical tools, than any *Decameron* previously published was Francesco Sansovino's edition (Sansovino 1546): it contains a glossary (*Dichiaratione di tutti i vocaboli, detti, proverbi e luoghi difficili* 'explanation of all the words, idioms, proverbs and difficult passages'), information about historical characters cited in Boccaccio's tales with reference to literary sources from which they are elicited (*I luoghi e gli auttori quali il Boccaccio ha tolto i nomi* 'passages and literary authors from which Boccaccio has elicited characters'), names of ancient Florentine families, and finally an alphabetical wordlist of collocations used by Boccaccio (*Epitheti usati da m. Giovanni Boccaccio posti per ordine di alfabeto*). The *Dichiaratione* is a relatively large glossary (eighteen quarto-pages) containing entries of a certain size and complexity; headwords are both words and idioms or brief passages, organised in inaccurate alphabetical order. The main focus is on meaning, etymology and usage, but there are entries of an almost encyclopaedic kind, such as:

> *Cappuccio* [...], da *capo*, o 'cappello' o 'berretta', & cappuccio è propriamente quella cosa di panno fatta a guisa di manica che portavano i Fiorentini tempo fa, da l'un de' lati pendeva una stola & in cotal modo si veggano i veri ritratti di Dante & qualch'un'altra antica figura ('Hood [...], derived from *capo*, [means] 'headgear' and 'cap', and *cappuccio* is specifically a sleeve-like cloth thing that Florentines were used to wear some time ago, with a lateral tail; you can see such a hood in the old portraits of Dante and in some other ancient depiction', Sansovino 1546: s.v. *cappuccio*).

Sansovino's aptitude for linguistic and extra linguistic explanations can be seen in entries devoted to idioms, as s.v. *pan per focaccia*, an idiomatic equivalent of *tit-for-tat*:

> *Pan per focaccia* [...] 'il medesimo per il medesimo', ma in forma diversa: la qualità della focaccia & del pane è una medesima, ma la forma dell'uno è tonda e dell'altro tonda & stiacciata ('*Pan per focaccia* [...] 'like for like', but in a different way: bread and focaccia bread are of the same sort, but the first is round and the second round and flat', Sansovino 1546: s.v. *pan per focaccia*).

Anyway, Sansovino mixed different kinds of entries: some are mostly "linguistic", explaining etymology and meaning, others resemble commentaries to single passages (see for instance s.v. *bolognini in grossi*, *cavalcando la bestia*, *laude che cantavano i secolari*, etc.), others grammatical advising about ancient and standard usage (see s.v. *cotesto*, *isdrucire*, *in colui danno*, etc.), others "philological" discussions on textual variants (see s.v. *alloppiato*, *cappi*, *posta*, etc.).

Section *Epitheti usati* is the first dictionary of collocations I know of: it alphabetically lists nouns followed by adjectives or prepositional phrases, participial phrases, infinitive phrases as headword specifier. For instance, headword *acqua* ‘water’ is followed in random order by *calda* ‘hot’, *salsa* ‘salted’ (like *saltwater*), *fredda* ‘cold’, *freddissima* ‘very cold’, *fresca* ‘fresh’, *benedetta* ‘holy’, *santa* ‘holy’, *mortifera* ‘poisoned’, *avelenata* ‘poisoned’, *rosata* ‘rose’ (as in *rose water*), *di fior d’aranci* ‘orange flower’, *da bere* ‘drinking’, *chiara* ‘clear’, *lavorata* ‘blended’, *alloppiata* ‘poisoned’, *rosa* ‘rose’ (as in *rose water*), *di fior di gelsomini* ‘jasmine flower’, *nanfa* ‘orange flower’, *odorifera* ‘scented’, *marina* ‘sea’ (as in *sea water*), *viva* ‘running’. It was a basic aid to composing literary works in prose and poetry: it offered a set of mosaic tiles ready to use, imitating the style of canonical authors. This approach was very far apart from the basis of a modern dictionary of collocations, but it implies an idea of real language as interconnected lexico-semantic relations consolidated by usage.

Few years later, in 1552, Girolamo Ruscelli reprinted Sansovino’s *Epitheti* in his edition of the *Decameron*, to which he added an alphabetical index of all characters with the corresponding story abstract (*Tavola di tutti e nomi propri* ‘index of all proper names’) and a glossary (*Vocabolario generale di tutte le voci usate dal Boccaccio bisognose di dichiaratione, d’avvertimento o di regola*). In his short introduction, Ruscelli highlights that the *Vocabolario* explains «tutta la grammatica, o regole o commentari della lingua» (‘all grammar, or rules or annotations concerning the language’), except for «i soli spiegamenti o coniugationi de’ verbi» (‘verbal declensions or conjugations’). Although it could be an overstated assertion, in his entries Ruscelli devoted real attention to grammatical rules, from pronunciation to orthography, from morphology to syntax: results are not always satisfactory from a strictly prescriptive point of view (concurrent forms of a word are often considered as equivalents), but his effort to combine textual comprehension and grammatical rules is clearly visible. See for instance how entries *ahi* (an exclamation like *ay*, *ouch*) and *autore* explain the distinctive value of grapheme *h* and of simple or double *t*:

> *Ahi*, voce di gridare o dolersi. Et si scrive co(n) la *h* in mezo, p(er)ché se(n)za potrebbe pigliarsi per *a i* cioè *alli* che il verso usa [...]; & se si ponesse la *h* avanti farebbe il verbo *havere* [...] (‘*Ahi*, for crying or aching. It needs an *h* in the middle, because without *h* it would be *ai* i.e. *alli* [‘to the’] when used in poetry [...]; with an *h* at the beginning it would be a verbal form of *avere* [‘to have’]’, Ruscelli 1552: s.v. *ahi*).

> *Autore* con una *t* sola scrivono la più parte, & anco con due, ma il primo è più ragionevole, perché ancor nel Latino quando non significa ‘accrescitore’ si scrive con una sola. Significa *autore* a noi il medesimo che ‘compositore di qualche opera’ e anche ‘facitore’ o ‘principale nel far la cosa’ (‘Most of the people write *autore* with a simple *t*, but also with double *t*: the first is more reasonable, because even in Latin they write it with a single *t* except for the meaning of ‘accrescitore’[4]. For us, *autore* means ‘writer of some work’ and also ‘maker’ or ‘principal maker’’, Ruscelli 1552: s.v. *autore*).

[4] Two different words originate from Latin *auctor*: *autore* ‘author’ and, by learned derivation, *auttore* ‘who increases sth’ (the second one does not exist any more in Italian).

6. Conclusions

This short overview on glossaries and wordlists of the first half of the sixteenth century highlights that lexicographical practice's autonomy was not a starting point but a slow achievement, consolidated in 1612 with the edition of the *Vocabolario degli Accademici della Crusca*. Early modern Italian lexical tools are deeply implicated in linguistic standardisation and diffusion of the common linguistic type[5] [linguistic type , quer dizer "variedade linguística"? Parece-me pouco claro]. They satisfied "explanatory" requests (what a certain Tuscan word means) and "prescriptive" requests (how to use correctly a certain word) asked especially by non-Tuscan speakers. These are, with few adjustments, the same requests of a modern ingenuous user. In this perspective semantic, grammar and editing of canonical texts (and even "good style", i.e. distinction of language for prose and poetry) are necessarily linked together.

References

Acarisio, A. (1543). *Vocabolario, grammatica et orthografia de la lingua volgare*. Cento (Ferrara): in casa de l'autore [repr.: Acarisio, A. (1988). *Vocabolario, grammatica e ortografia della lingua volgare*. P. Trovato (ed.). Ferrara: Forni].

Alunno, F. (1539). *Il Petrarca con le osservationi di messer Francesco Alunno*. Venezia: Francesco Marcolini.

Alunno, F. (1543). *Le ricchezze della lingua volgare*. Venezia: Eredi di Aldo Manuzio [Alunno, F. (1551). *Le ricchezze della lingua volgare sopra il Boccaccio* [Second edition]. Venezia: Eredi di Aldo Manuzio].

Alunno, F. (1548). *La Fabrica del mondo*. Venezia: Niccolò de Bascarini.

Brucioli, A. (1538). "Tavola di tutti i vocaboli, detti e modi di dire incogniti e difficili". In Boccaccio, G. *Decamerone*. Venezia: Giovanni Giolito da Trino.

Dolce, L. (1541). *Il Decamerone di Messer Giovanni Boccaccio nuovamente stampato e ricorretto per Messer Lodovico Dolce, con la dichiarazione di tutti i vocaboli, detti, proverbi, figure et modi di dire incogniti e difficili che sono in esso libro*. Venezia: Curzio Navò e fratelli [or Bindoni e Pasini].

Liburnio, N. (1526). *Le tre fontane*. Venezia: Gregorio de Gregori.

Lobodanov, A. (1999). "Cenni sulla storia del pensiero lessicografico nei primi vocabolari del volgare". *Studi di lessicografia italiana* 16. 253-265 [repr. in Lobodanov, A.P. (2011). *L'arte della parola. Saggi di linguistica slavoromanza*. Pisa: Tipografia Editrice Pisana. 155-172].

[5] «La prima lessicografia è strettamente legata alle teorizzazioni e alle discussioni che, in ambito veneziano, [...] si svolgevano sui connotati di fondo del volgare [...]. Di questa discussione teorica la lessicografia è parte integrante, anzi contribuisce a fissare l'ipotesi dominante [...]. Storia della lessicografia, storia della questione della lingua, storia delle prime grammatiche sono strettamente integrate e il "vocabolario" concorre alla creazione di una norma condivisa» ('Early lexicography is strictly linked to theories and discussions on basic features of volgare held in the Venetian circle [...]. Lexicography is an essential part of this theorical discussion, contributing to fix the main perspective [...]. History of lexicography, history of linguistic debate, history of early grammars are closely related and the "dictionary" contributes to create a communal rule', Paccagnella 2013: 50-51).

Luna, F. (1536). *Vocabulario di cinquemila vocabuli toschi non men oscuri che utili e necessarij del Furioso, Bocaccio, Petrarcha e Dante novamente dechiarati e raccolti da Fabricio Luna per alfabeta ad utilità di chi legge, scrive e favella*. Napoli: Giovanni Sultzbach.

Manni, P. (1991). "Note sull'idea di lessico nei primi vocabolari italiani". In Giannelli, L. et al. (eds.). *Tra Rinascimento e strutture attuali. Saggi di linguistica italiana*. Atti del I Convegno della Società Internazionale di Linguistica e Filologia Italiana (Siena, 28-31 marzo 1989). I. 69-79. Torino: Rosenberg & Sellier.

Marazzini, C. (2009). *L'ordine delle parole. Storia di vocabolari italiani*. Bologna: il Mulino.

Minerbi, L. (1535). "Vocabulario". In *Il Decamerone di m. Giovanni Boccaccio col Vocabulario di m. Lucilio Minerbi nuovamente stampato et con somma diligentia ridotto*. Venezia: Bernardino di Vidali [non-numbered pages].

Olivieri, O. (1942). "I primi vocabolari italiani fino alla prima edizione della Crusca". *Studi di filologia italiana*. 6. 64-192.

Paccagnella, I. (2013). "L'editoria veneziana e la lessicografia prima della Crusca". In Tomasin, L. (ed.). *Il Vocabolario degli Accademici della Crusca (1612) e la storia della lessicografia italiana.* Atti del X Convegno ASLI (Padova, 29-30 novembre 2012 — Venezia, 1 dicembre 2012). Firenze: Cesati Editore. 47-64.

Poggi Salani, T. (1986). "Venticinque anni di lessicografia italiana delle origini (leggere, scrivere e «politamente parlare»): note sull'idea di lingua". In Ramat, P.; Niederehe, H.-J.; Koerner, K. (eds.). *The History of Linguistics in Italy*, Amsterdam/Philadelphia: Benjamins Publishing Company. 51-83 [formerly in: Historiographia Linguistica 9. 265-297].

Richardson, B. (1993). Print culture in Renaissance Italy: the editor and the vernacular text, 1470-1600. Cambridge: Cambridge University Press.

Ruscelli, G. (1552). "Vocabolario generale di tutte le voci usate dal Boccaccio bisognose di dichiaratione, d'avvertimento o di regola". In *Il Decamerone di m. Giovan Boccaccio, nuovamente alla sua intera perfettione, non meno nella scrittura che nelle parole ridotto per Girolamo Ruscelli. Con le dichiarationi, annotationi et avvertimenti del medesimo sopra tutti i luoghi difficili, regole, modi et ornamenti della lingua volgare [...]*. Venezia: Vincenzo Valgrisi. [non-numbered pages].

Sansovino, F. (1546). "Dichiaratione di tutti i vocaboli, detti, proverbi e luoghi difficili [...] aggiuntevi alcune annotationi de' luoghi [...] e nel fine una parte delle voci con i loro più propri epiteti, con i nomi delle casate più antiche in Firenze". In Boccaccio, G. *Il Decamerone [...] con gli epitheti dell'autore, espositione de' proverbi et luoghi difficili.* Venezia: Gabriel Giolito de Ferrari. 2 vol. [non-numbered pages].

Tancke, G. (1984). *Die italienischen Wörterbücher von den Anfängen bis zum Erscheinen des "Vocabolario degli Accademici della Crusca" (1612). Bestandsaufnahme und Analyse*. Tübingen: Niemeyer.

Trovato, P. (1988). "Introduzione". In Acarisio, A. (1988). *Vocabolario, grammatica e ortografia della lingua volgare*. Trovato, P. (ed.). Ferrara: Forni. VII-XLI [repr. of Acarisio 1543].

Trovato, P. (1991). *Con ogni diligenza corretto. La stampa e le revisioni editoriali dei testi letterari italiani (1470-1570).* Bologna: il Mulino [repr. Ferrara: UnifePress, 2009].

Verini, G. B. (1532). *Dictionario [...] in lo quale si contiene tutti li nomi masculini e feminini di tutte quante le cose del mondo vive et morte in lingua toscha*. Milano: Gottardo da Ponte.

The corpus of the *Dicionário Histórico do Português do Brasil* (*DHPB*)

Clotilde Murakawa & Maria Filomena Gonçalves

Introduction

The aim of this paper is to show the importance of the textual corpus of the *Dicionário Histórico do Português do Brasil* (*Dictionary of the History of Brazilian Portuguese*), ranging from the sixteenth to the eighteenth centuries, hereinafter referred to as *DHPB*[6]. The chronological milestones of the "textual corpus databank" range from the year 1500 to 1808, the date in which the Portuguese court arrived in Brazil. Assembled specifically for the *DHPB*, the *databank* of the abovementioned project presents some problems in what relates to texts (manuscripts and printed) produced in very diverse conditions over more than three centuries, which raised a number of philological and linguistic problems. In fact, the corpus of the project gathered documentation about all Brazilian regions and the most varied local realities.

I. The DHPB project and its features

The *DHPB* is a reference work which documents the lexical collection that would have originated the Portuguese language in its Brazilian variant, based on a database with approximately 10 million occurrences, from which the *DHPB* word list was extracted. The corpus that composed the database was produced from the sixteenth through the eighteenth centuries, with the year 1500 - the time of the

[6] The project *DHPB* received financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). It was designed by Prof. Maria Tereza Camargo Biderman and began in December 2005. Following the death of Prof. Biderman, in May 2008, the project was carried on by a team of professionals under the supervision of Clotilde Murakawa at the Faculdade de Ciências e Letras, UNESP - Univ Estadual Paulista, Araraquara, Laboratório de Lexicografia. The project was completed in November 2012 with 10,470 entries, gathered in 19 volumes with a total of 11,051 pages.

letter by Pero Vaz de Caminha on the discovery of Brazil - as the starting date for the collection of documents, and the year 1808 - the arrival of the Portuguese royal family in Brazil - as the final collection date. The word list is composed of nouns, adjectives, and verbs. The definition of the entry word or lemma is based on the contexts extracted from the database via the search engine of the *PhiloLogic*™ software program, which was especially adapted for the *DHPB*.

1. Written corpus selection and IT processing

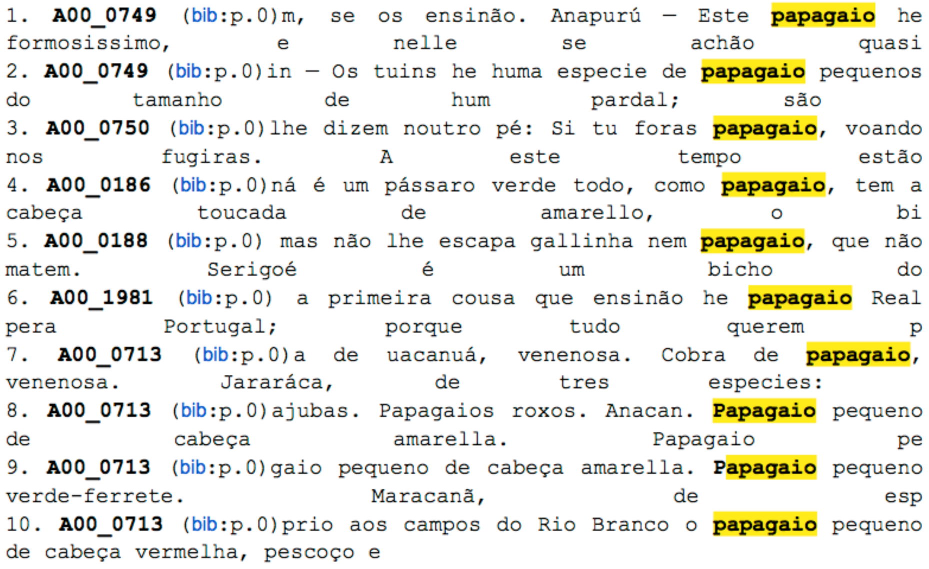
The selection of documents produced during the intended period (16th-18th) gathered a representative corpus with various text genres and nature, such as: literary works of missionaries (mostly Jesuits settled in Brazil), logbooks, letters of land grants, descriptive itineraries of the Brazilian flora and fauna, geographical descriptions, business letters exchanged between traders from the colony and Portugal, private letters, notarial documents, law suits, inventories, wills, licenses, legal documents, military regulations, works on medicine, pharmacy, agriculture, and mining, the literary production of the Baroque and Arcadian periods.

In order to input the previously mentioned corpus in the database, all texts had to be submitted to an IT process that gave them a special format, compatible with the *PhiloLogic* software program adapted for this purpose. The following sequence was obeyed:

a) scanning of texts and editing of images;
b) OCR and correction with the ABBYY Fine Reader software;
c) inclusion of cataloging data in the texts already corrected;
d) XML markup;
e) input of the text in the PhiloLogic software program.

Based on the whole IT processing, a database with a total of 9,541,721 occurrences was built from 32,358 scanned pages.

The software allows to extract lexical units in the format of the following query example, for *papagaio*:



```
1. A00_0749 (bib:p.0)m, se os ensinão. Anapurú — Este papagaio he
formosissimo, e nelle se achão quasi
2. A00_0749 (bib:p.0)in — Os tuins he huma especie de papagaio pequenos
do tamanho de hum pardal; são
3. A00_0750 (bib:p.0)lhe dizem noutro pé: Si tu foras papagaio, voando
nos fugiras. A este tempo estão
4. A00_0186 (bib:p.0)ná é um pássaro verde todo, como papagaio, tem a
cabeça toucada de amarello, o bi
5. A00_0188 (bib:p.0) mas não lhe escapa gallinha nem papagaio, que não
matem. Serigoé é um bicho do
6. A00_1981 (bib:p.0) a primeira cousa que ensinão he papagaio Real
pera Portugal; porque tudo querem p
7. A00_0713 (bib:p.0)a de uacanuá, venenosa. Cobra de papagaio,
venenosa. Jararáca, de tres especies:
8. A00_0713 (bib:p.0)ajubas. Papagaios roxos. Anacan. Papagaio pequeno
de cabeça amarella. Papagaio pe
9. A00_0713 (bib:p.0)gaio pequeno de cabeça amarella. Papagaio pequeno
verde-ferrete. Maracanã, de esp
10. A00_0713 (bib:p.0)prio aos campos do Rio Branco o papagaio pequeno
de cabeça vermelha, pescoço e
```

There are 87 more occurrences for the plural **papagaios**, one for **papagajo** (from 1591), six for **papagayo**, and seven for **papagayos.** The oldest document recording **papagayo** is the 1500 Letter by Pero Vaz de Caminha.

3. Microstructure

The types of definition that guided the construction of the entries is based on the theories of definition proposed by Imbs (1960), Dubois & Dubois (1971), Bosque (1982), Rey-Debove (1984), Porto-Dapena (2002), Garriga Escribano (2003), and Castillo Carballo (2003), adding the authors of *La Lexicografia – De la Lingüística Teórica a la Lexicografia Práctica* (1982), especially Haensch. The typology proposed by I. Bosque (1982) in his classic article *Sobre la teoria de la definición lexicográfica* (On the theory of lexicographical definition) presented the best theoretical support to the writing of definition in the *DHPB*.

The microstructure was designed setting apart compulsory information and optional information. Compulsory pieces of information in the *DHPB*, i.e., those to be contained in all entries, include: 1) the entry word or lemma, that was spelled according to the *Vocabulário Ortográfico da Língua Portuguesa* (*VOLP*) from 2009. The updated spelling facilitates consultation at the *DHPB*. In the vast majority, the database recorded the spelling of the *VOLP*, but there were some cases where it was not included. In such cases, the spelling of the *VOLP* prevailed and the spellings recorded in the database were considered variants; 2) the grammatical class: noun, adjective or verb; 3) all the meanings or values that the lemma presents in the many contexts of the database, that were necessarily accompanied by the context with complete bibliographic reference; 4) date record that selects the oldest text in the corpus where the entry was documented.

Optional pieces of information, which were dependent on being registered or not in the database, include: 1) graphic, morphological, or phonetic variants were placed immediately after the lemma and were always accompanied by the context; 2) figurative sense; 3) the entry word can integrate a noun, adjective, verb, prepositional, adverbial, or conjunctive phrase; 4) the entry word has joined another word to form a nominal (noun or adjective) or verbal phrase; in this case, the entry recorded it under the label of syntagmatic expression; 5) encyclopedic information - whenever some historical information was interesting to clarify the lexicographical definition, it was recorded as a note; 6) use of cross references, whenever the reader is invited to go to another entry where he can get more information. The six pieces of information just mentioned occur strictly in this order within the entry.

Although basic principles for the organization of the entries have been established along the search in the database, some other options and solutions had to be considered, taking into account the relevance of information since this is a historical dictionary where historical data can greatly assist the construction of the lexical definition.

The remaining sections illustrate problem-solving choices, as they occur in the *DHPB*. Notice that the contexts that illustrate the meanings of the *DHPB* are accompanied by bibliographic information as shown in the database; number in brackets indicate the file number and text page where the lexical unit of the entry can be found.

3.1. Orthographic entry according to the *VOLP*

When the database did not record the entry according to the *VOLP*, the spelling of this vocabulary was kept and the form was registered in the database as a variant, e.g.:

ginseng *n.m.*
variants: jinsen, jensen.
Perennial herb with rhizomes and thick, aromatic, medicinal root.
Padu é um cipó do Amazonas ainda pouco vulgar, e conhecido, mas na verdade digno de muita estimação, e pode correr parelhas com o famigerado jinsen da China [...]
PE. JOÃO DANIEL (1976) [1757], [...] Tratado terceiro - Da riqueza do amazonas na preciosidade da sua madeira - cap. 6º - De algumas ervas mais notáveis do Amazonas [A00_1866 p. 374].
1st register [1757]
Jensen. É ũa erva de muita estimação no Império da China tanto, que se compra a peso de dinheiro, ou para melhor dizer um peso de Jensen vale muitos pesos de prata. Tem grandes préstimos especialmente para fortalecer, e avivar os espíritos, para suprir as faltas de comer, sede, e somno, ou tomando o seu chá, ou mastigando a sua raiz. PE. JOÃO DANIEL (1976) [1757], [...] Tratado terceiro - Da riqueza do amazonas na preciosidade da sua madeira - cap. 6º - De algumas ervas mais notáveis do Amazonas [A00_1866 p. 373].

3.2. Variants

We chose to register all the variant forms found in the database. With the contextualization of the entry, we state the literary work and the time it was used. Some examples of variants:

semear *v.*
variants: cemear, semiar.
sacristão *n.m.*
variants: sachristão, sacristam, sacristaõ, sanchristaõ, sancristão.
malagueta *n.f.*
variants: malaguêta, malangueta.
chumbo *n.m.*
variants: xumbo, chunbo.

3.3 Homonymy

For homonymous words - homophones and homographs - the three criteria established by lexicographical practice - etymological, grammatical or functional, and semantic - consequently lexical units could be inserted separately. Of the three criteria previously mentioned, the most commonly used in dictionaries is the etymological, which states that lexical units of different origin should be placed in separate entries; however, the grammatical or functional criterion was very productive, as verified in the database.

melado[1] *adj.*

variant: mellado.

Smeared with honey.

Para enrolar o Tabaco, dobraõ a corda já curada, & melada, de comprimento de tres palmos, ſobre hũa eſtaca, naõ muito groſſa, & leve, que nas extremidades tem quatro taboinhas em cruz: [...]. André João Antonil (1711) [1711], Segunda parte - Na lavra do tabaco [A00_2579 p. 113].

1st register [1618]

A ordem é esta: depois do açúcar limpo e melado nas caldeiras, se passa a umas tachas também de cobre, aonde à fôrça de fogo o fazem pôr no ponto necessário para haver de coalhar [...]. Ambrósio Fernandes Brandão (1966) [1618], Diálogo terceiro - Em que se trata das mercancias do açúcar, pau, algodão, madeira [A00_1583 p. 85].

melado[2] *n.m.*

Thick, dark syrup made from sugarcane; it is improved after cooking and placed in large pans.

[...] porq. o tenpo me não da lugar p.ª regallos, as seias paço com hum prato de milho cozido em agoa com huma colher de mellado, cujo prato lhe dão qua de quangiqua [...]. Francisco da Cruz (1973) [1726], Cartas remetidas para Lisboa-Minas Gerais [A00_0437 p. 288].

3.4. Phraseologies

Under the label of syntagmatic expressions, the entries present nominal, adjectival and verbal expressions, as well as conjunctive, prepositional and adverbial phrases. Many of these nominal expressions are still in use. The current use of many of them often leads to the thought that they are recent, but the corpus proves differently.[7]

3.4.1. Nominal syntagmatic expressions

The syntagmatic expressions in the *DHPB* can be found under the entry of the first nominal element. See the following examples:

Cabeça de julgado (main seat; seat of the judged); *cabo de esquadra* (military rank below sergeant); *cabo de guerra* (commander, chief who became renowned in war campaign); *capa consistorial* (cloak used by cardinals when they meet in the consistory); *capa de asperges* or *capa pluvial* (long cloak with no pleats used by priests in certain ceremonies); *capelão-mor* (priest who was in charge of the religious service of the royal chapel); *crime capital* (capital crime; punished with the death penalty); *pecado capital* (deadly sin); *camisa de onze varas* (very embarrassing situation); *carta de data* (document that records the donation of goods or properties to others); *carta de sangrar* (document authorizing the application of suction cups and leeches on ill people); *pé de boi* (a cautious man); *roupa de franceses* (objects

[7] Failing to accurately translate the phraseologisms shown in items 5.4.1, 5.4.2, and 5.4.3., the meaning in English is registered in parenthesis.

that have no owner); *roupa branca* (a set of linen or cotton shirts, dresses and tablecloths); *roupa de cama* (bed linen); *pé de moleque* (candy made with a mixture of caramel sauce and crumbed peanuts); *moleque de assentar* (a thick wooden stick which served as a stickle to equal the sugar inside boxes in the old sugar mills); *moleque de quebrar* (a utensil similar to a shovel which was used to crumble bread in the old sugar mills); *coroa de areia* (clustering of sand above the water level); *baraço e pregão* (rope and notice of guilt that prisoners had to wear outside the prision); *inimigo jurado* (avowed enemy who has implacable hatred and has sworn revenge); *gota coral* (epilepsy); *corrente e moente* (it is said of the sugar mill that is ready to operate); *criança de peito* (the child who is still breastfed); *mel de tanque* (molasses collected from reservoirs in sugar mills); *talhado a pique* (of great perpendicular height); *taipa de mão* (wall made of cross-woven slats, and then covered or plastered with mud); *taipa de pilão* (wall made of crushed clay and gravel); *légua quadrada* (the largest unit of area of the Imperial measuring system, equivalent to approximately 24 km^2); *armada sutil* (light and small vessel).

Nominal expressions also document the diversity of the Brazilian flora and fauna, as well as the flora that was transplanted to Brazil, brought by the colonizers; the meaning of the terms expressed ahead is given only indicating the genus next to the lexicographical definition, as in:

Martim-pescador (bird), *couve-flor* (cauliflower), *couve-murciana* (type of cabbage), *couve-rábão* (type of cabbage), *couve-tronchuda* (type of cabbage), *couve verde* (type of cabbage), *tatu-caba* (wasp species), *sabiá-branco* (bird), *sabiá-coca* (bird), *sabiá-da-praia* (bird), *sabiá-poca* (bird), *sabiá-verdadeiro* (bird), *sabiá-vermelho* (bird), *araticum-apê* (tree), *araticum-bravo* (tree), *araticum-panã* (tree), *erva-andorinha* (herb), *erva-babosa* (medicinal plant), *erva beltrona* (medicinal plant), *erva-botão* (herb used in dyeing), *erva cavalinha* (aquatic weed), *erva cheirosa* (greenery used as seasoning), *erva cidreira* (aromatic medicinal herb), *erva-de-bicho* (herb), *erva-de-cobra* (medicinal plant), *erva-teiú* (medicinal herb), *erva-de-rato* (tree), *erva jequeri* (poisonous herb), *erva-lombrigueira* (medicinal herb), *erva ginseng* (medicinal herb), *erva-tostão* (medicinal herb), *jacu-pema* or *jacu-pemba* (guan species), *jararaca do rabo-branco* (poisonous snake)

3.4.2. Verbal syntagmatic expressions

The same way as the nominal syntagmatic expressions, verbal expressions compose the entries of the *DHPB*, and they are always equivalent to a verb in the lexicographical definition; therefore, there is always a verb heading the definition, as in the following examples:

Assinar o rogo (to sign on behalf of someone who is illiterate); *pregar no deserto* (to speak in vain; not being heard by the people to whom the preaching is being addressed); *virar a casaca* (to change position or opinion); *pagar o pato* (to bear the consequences of something that one is not responsible for); *fechar os ouvidos* (to refuse to listen to reason or excuse); *ficar no tinteiro* (to fail to accomplish something by oversight or omission); *passar revista* (to inspect); *meter alguém na dança* (to involve someone in business through fraud); *aturar a bucha* (to endure something uncomfortable); *meter a foice em seara alheia* (to intrude on something); *meter-se como piolho em costura* (to appear everywhere untimely, without being called or invited); *deitar água na fervura* (to lose heart); *estar na prancha da língua* (to be ready to be told); *cozinhar gato por lebre* or *vender gato por lebre* (to deceive); *falar aos cotovelos* (to talk too much); *fazer das tripas coração* (to take courage from weakness; fill up with courage); *dar couto* (to give refuge, asylum; protect someone); *meter na cabeça* (to convince, persuade); *não fazer bom cabelo* (to dissatisfy); *doer o cabelo* (to be afraid of some evil); *estar muito ao cabo* (to be close to death); *dar jus a* (to grant the right); *criar corpo* (to grow, fatten, gain consistency); *furtar o corpo* (to dodge to avoid a stroke or blow); *ir às correntes* (to discard the poor quality sugar); *tomar a ocasião pela calva* (to miss); *ficar/ estar em talas* (to get in trouble); *ter em talas* (to get under mastery); *ficar por portas* (to be in misery); *dar em rosto* or *deitar em rosto* (to censure, say confronting things face to face); *levar (algo) na unha* (to imprison); *dar à vela* or *fazer(-se) à (a) vela* (to start navigating).

3.4.3. Other phrases

Phrases are equivalent to a grammar tool in contexts, that is, they are prepositions, conjunctions or adverbs. Their main lexical unit is the inserted word; thus, the phrase 'bedridden/in bed' (bedridden, lying in bed because of illness, *de cama*), occurs under the lemma 'bed' (*cama*). Other examples: *pelos cabelos* (at the expense of work and sacrifice); *com unhas e dentes* (fiercely, in all possible ways); *de corrida* (quickly, without delay); *a queima roupa* (very closely); *a olho* (without weight or measure; aimlessly); *com olho sobre o ombro* (with contempt); *a olhos vistos* (clearly, in a way everyone can see).

There are lexical units, however, that are only registered in phrases, as in: *de permeio* (permeating, in between) and *de cor* (by heart). In the last two, there was the need to open an entry where only the phrase appears, since the word *cor*, meaning memory, and the word *permeio*, meaning in between, appear only in these phrases.

3.5. Dating

At the end of the entry, the date is compulsory information, as well as the oldest context where the lexical unit is documented. If the precise date is not available, the century is registered. In older texts, the entry may be the same form as in the VOLP, or in variant or inflected form in the case of nouns and adjectives.

With respect to verbs, concerning the dating, we decided to register the infinitive form, owing to the impossibility to go through all the conjugated verb forms in the database. When appropriate, the older variant infinitive form was chosen. See the following examples:

> suspeitar *v.*
> variants: sospeitar, çuspeitar.
> 1st dated [1562]
>
> Mas tambem nos intristecia sua tardança por chegarem outros navios que partirão detrás e dizião que era a nao mais veleira que elles, o que nos fazia sospeitar serem tomados de Franceses ou averem aribado às Antilhas. P. LEONARDO DO VALE (1956) [1562], Carta do P. Leonardo do Vale por comissão do P. Luís da Grã aos Padres e Irmãos De S. Roque, Baía 26 de Junho 1562 [A00_0057 p. 486].

taxar *v.*
variants: taixar, tachar.
1st dated [1669]
[...] e o que comprar de Angola e Santo Thomé possamos tachar conforme a abundancia ou carestia que delle houver, e que hum e outro se venda sem mostura [...]. desconhecido (1951) [1669], Registo da carta que, se escreveo ao procurador Joze Moreira de Azevedo, 1669 [A00_2166 p. 74].

II. The curpus databank and the tracking of 'brasileirismos'

There are many lines of thought provided by the *DHPB* corpus. Our attention will focus therefore on philological and linguistic aspects that best demonstrated the value of the documentary collection of the *DHPB* both in respect to the dating of the Portuguese lexicon in the land of Vera Cruz as regards the story of "brasileirismos" (that is, Brazilian variety specific words)[8].

It should be noted that, although the project was formally terminated in what regards to the sponsoring entity (CNPq, Brazil), the *DHPB* will still be subject to a general revision, in order to integrate the data from the *Textual Databank 2* documents, which will surely result in an increase of the nomenclature of the dictionary and in updating the entries. The *Databank 1* consists of 23,858 pages and accounts for 7,492,472 occurrences; in turn, the *Databank 2* contains 2,049,249 occurrences and 8,009 scanned pages. Both *corpora* include documents belonging to the most diverse textual and speech genres, produced between 1500, date of the *Carta de Pêro Vaz de Caminha*, the letter reporting the discovery of Brazil to King D. Manuel, and 1808, the year of the arrival of D. João VI (1767-1808) to the Brazilian territory.

There are several reasons to consider the *DHPB* a "historical dictionary". The first is the usage of a chronological criterion underlying the selection of a three centuries period of Portuguese lexicon; the second is the inclusion of lexical units used in the selected period based on their frequency in the corpus; the third concerns the contextualization of the meaning units by means of quotes extracted from the *Corpus Databank*; the last one is the dating of the oldest of all occurrences of each word in the corpus. In addition, the *Dictionary* registers some additional information, such as graphical variations and combinations of syntagmatic units taken from the *Textual Databank*.

The *DHPB* is different from "diachronic dictionaries" in general, and from "panchronic" dictionaries in particular, because it does not consider the entire lexicon, nor does it consider all changes in lexical diachrony. According to the typology of lexicographical works of Günther Haensch (1982: 161), the *DHPB* is not, therefore, a diachronic dictionary (Murakawa 2010; Murakawa 2012), but is not a "general dictionary"[9] either; it is, indeed, a “historical” and “partial dictionary”,

[8] On the concept of “brasileirismo”, we refer to: Boléo (1943), Cunha (1987), Oliveira (1998), Murakawa (2005). On the introduction of the lexicographical mark “brasileirismo”, confer Gonçalves (2006).

[9] The *DHPB* only integrate selected nouns, adjectives and verbs.

which aims to define the units taken from the corpus, depending on their frequency and according to the contexts in which they occur. The chronological span – three centuries of history – enables us to inventorying and characterizing the lexicon of the Portuguese colonial period.

The presence of the Portuguese court in Brazil constituted a turning point in the history of Brazil. This is why the date of independence (1822) was not taken as a *terminus ad quem* for the *Dictionary*. If the project would target the entire diachrony, the corpus would have to cover the pre-1500 synchronies and those after 1808. In turn, the *DHPB* should contemplate the etymology, the first register of the words, as well as all variants of the word over time and changes occurred both at a formal and a semantic levels (Haensch 1982: 160-163).

According to the periodization of Portuguese (Castro 1999; Castro 2006: 73-77), the three centuries covered by the *DHPB* range from so called *middle Portuguese* to *classical Portuguese*, which will last until late 18th century, a period to which Rosa Virginia Mattos e Silva (2006: 21-26) prefers to call *modern Portuguese*. It should be noted that the lexical-semantic dynamics occurred in the Portuguese lexicon in Brazilian soil is a result of the processes of "re-semantization" or "re-signification" of fully Portuguese words, the creation of new combinations and the import of words from African and indigenous languages, and of the influence of the dialectal and sociolinguistic variations taken to Brazil with the first settlers and those who followed them[10]. In addition, the demographic and social dynamics have been of capital importance, for example, during the 18th century "gold rush", when the settlement became more intense.

1. Textual corpus databank

With a *terminus a quo* in 1500 and a *terminus ad quem* in 1808 expectations are that the *Textual Corpus Databank* brings to the table problems of various kinds, namely technical, philological and linguistic. It should be reminded that this corpus included unpublished manuscripts, and other printed texts of the three centuries in study. Manuscripts had to be transcribed to integrate the *Databank*; the others were either presented in *princeps* editions – see the work of Antonil (1711) [11] – or in 19th century editions, elaborated accordingly to philological disparate criteria.

Given the complexity of the documentary heritage and the diversity of philological situations, in order to match the lexical model designed for the *DHPB*, the project team faced difficult but necessary decisions.

[10] We do not intend to summarize the main explanatory hypotheses about the origins of Brazilian Portuguese, but we may address the reader to Mattos e Silva (2005), Lucchesi (2012) and Naro and Scherre (2007).

[11] It is *Cultura e opulência do Brasil por suas drogas e minas*. One of the few copies known belongs to the public library of Évora and it was scanned for the *Textual Corpus Databank*.

1.1. Challenges

Since the corpus was a mass of texts with different stories and editorial fates — autographs, apocrypha, prints, unedited documents, first editions, 19th century diplomatic and semi-diplomatic editions, modernizing editions, etc. —, the first challenge was, of course, of philological order.

It was necessary to decide whether the texts available solely on editions made by historians and philologists of the 19th century would be the target of a reliable genre edition, marked by appropriate criteria to tradition and "lectio" of each text. This huge task would constitute a separate project that would render unenforceable the *DHPB*, at least within the time limit established by the funding agency. As the purposes and, consequently, the requirements of a diachronic dictionary differ from a historical dictionary, it was decided that the *Textual Databank* would welcome the greater number of documents relating to Brazil, regardless of its philological circumstances, since the dictionary was designed to offer the inquirer definitions of each word in the *Textual Corpus Databank*, but not the "general" history of each unit. Thus, only the manuscripts had a conservative (diplomatic) edition. In the case of printed matter, whenever there was a first edition, one would opt for the latter. Otherwise, an available edition would be used, even though it would have been a modern one.

Given the specific nature of this *Dicionário Histórico*, the necessary condition for integrating a document (printed or manuscript) was in dealing with matters pertaining to the various regions of Brazil, in the three centuries of the studied period, in such a way as to be representative of each of the three centuries, but also of the most varied local realities, namely: fauna, flora, landforms, indigenous peoples, missionary activity, colonial administration, inland exploration of the territory, maritime and river navigation, mining, agriculture, trade, fishing, hunting tools, food, clothing, traditions, weights and measures (Gonçalves 2012a), domestic and social life, medicine and popular pharmacopoeia, among many other areas.

Consequently, The *DHPB*'s *Textual Corpus Databank* brings together a vast informative pool about Brazil in the 16th, 17th and 18th centuries, including manuscript sources unknown until now, while recovering and including printed works merely studied in specialized fields, as zoology or botany, for example, works which proved, in fact, to be paramount language sources, whether for the history of the lexicon, whether for the history of language in general.

The second challenge of the project was the *Textual Corpus Databank* model and the solutions offered by Corpus Linguistics, having been chosen the tool *PhiloLogic*, whose functionalities provide search and analysis in full texts. *PhiloLogic* has been developed by a consortium consisting of the *American and French Research on the Treasury of the French Language* (ARTLF) and the *Digital Library Development Center* (DLDC) at the University of Chicago, in cooperation with the ATILF (*Analyse et Traitement Informatique de la Langue Française*), based in Nancy, and the CNRS (*Centre National de la Recherche Scientifique*), and it has been used in several countries for projects related to different languages.

2. Lexical chronology: contributions of the *DHPB*

In this section, based on a small sample extracted from the *Textual Corpus Databank*, we proced to a demonstration of the documentary value of this compilation. Firstly, however, there will be a brief historical contextualization of the selected referential domains.

On economic history of Brazil, we can usually identify three distinct cycles – the ones of sugar, coffee and gold –, which are well represented in the *DHPB*'s *Textual Corpus Databank*. Let us consider the first one.

Originated in Madeira Island, in 1532, the first sugarcane seedlings are taken to Brazil by Martim Afonso de Sousa (c. 1490-1570). Sugar production reaches its pinnacle in the mid-16th century, so it is not so surprising that many Portuguese words emerge in the texts of that era, on the ground where was planted the sugarcane – the "plantation" – and in the sugar production place – *engenho-de-açúcar*[12] (i.e., the sugar mill) – have been "re-semantized", acquiring specific (or even exclusive) meanings pertaining both to that activity and context. The specialization of the lexicon used in the mills led to the establishment of a true "sugar terminology" (Gonçalves, 2012b), that is, a set of "terms", concerning the activity of a *engenho-de-açúcar*, in order to name whether the varieties of sugarcane and everything connected to this activity, which included the utensils used in its handling and milling or, even, the people who worked in the mill.

In Portuguese lexicography, some of these terms were attested, for the first time, in the *Vocabulario Portuguez, e Latino* (1712-1721), by D. Rafael Bluteau (1638-1734), whose entry corresponding to "sugar" was, in good part, extracted from the *História Natural do Brasil* (1648), a work by the Dutch naturalist George Marcgrave (1610-1644). In the *DHPB* corpus, the complex unit "cana-de-açúcar"[13] (sugarcane) is recorded in a document of Fr. Manuel da Nóbrega (1517-1570), which allows us to place this unit in the 16th century. Similarly, the example of *balceira*, which is an adjective that applies to a "shoddy, low in sucrose, and growing on moist land" sugarcane (Houaiss 2001), clearly brings forward the benefits of *corpora* in the preparation of dictionaries because the textual collections allow the dating of these units – *cana-de-assucar* (1557) and *balceira* (1802[14]) – which, in a dictionary of reference such as the Houaiss (2001), are not dated (Noll, 2012).

The sugar "terminology"[15] (Gonçalves, 2012b) is well represented in the *Textual Corpus Databank*. Although with different spellings, the word "sugar" has a high

[12] In the *Textual Corpus Databank*, the complex unit "engenho de cana" has its first occurrence in a letter sent by Luís de Góis to King D. João III, dated of May 12, 1548.

[13] Houaiss (2001) does not present a date for this unit.

[14] Date of a letter of Luiz dos Santos Vilhena (1744-1814). See the excerpt provided by *PhiloLogic*: "[...] porque de huma semente tirão fructo por quatro, cinco e mais safras, segundo a valentia das terras; e por este prejuizo não querem tirar duplicado lucro, receando-se tão bem do pouco que no primeiro anno dão os novedios, a que chamão canna **balceira** [...]"**.** In Houaiss (2001), this adjective receives the mark of a *pernambucano* lexicographic regionalism.

[15] On the concept of "terminology" and theoretical approaches on terminology, see: Cabré (1990); Krieger and Finatto (2004).

number of occurrences: *assucar*, 642 occurrences (between the years of 1583 and 1808); *asucar*, 107 (between 1585 and 1804); *açucar*[16], 266. Although the nouns "açúcar" (sugar) and "cana" (sugar cane) were belonging to the so-called "common" lexicon, because they did not integrate a "specialized language", the same cannot be said of the complex units as "açúcar cândi"[17] (sugar candy) and "açúcar mascavo" (brown sugar), whose use was not as usual excepting on sugar production and trade. Therefore, these two units are "terms". In Houaiss (2001), "cândi" (candy), or "cande", and "mascavo" (brown sugar) are not dated. Now, in the *DHPB*'s *Textual Databank*, "cândi" was recorded in 1583; the adjective "mascavo", which had only three occurrences in the corpus, attested in 1765, had a more frequent use as its equivalent "mascavado" (37 occurrences).

Let us see some examples:

> (1) [...] quando se quer servir delle, se lhe mistura huma sufficiente quantidade de assucar **candi**; e se formão pequenos rebuçados [...] Frei Mariano da Conceição Velloso (1805) [**1798**].

> (2) O asucar fica reduzido a três qualidades; branco; **mascavo**, e **mascavado** q' hé o inferior do fundo das fôrmas [...] Anônimo (muito provavelmente Joseph Barbosa de Sáa) (1999) [**1765**].

As for the names of occupations or positions performed in the mill, we highlight complex units as "mestre-de-açúcar" (sugar-master) and "torrão-de-açúcar" (sugar candy): the former, which Houaiss classifies as a Northeastern Brazilian dialecticism, refers to the "individual who supervises the work relating to the manufacture of sugar" (Houaiss 2001), and the latter, refers to a small portion of compressed rectangular shaped sugar; neither of which are, however, dated in Houaiss. In the *DHPB*, the former unit has attestation in 1583, and the latter, in 1699. However, we owe the first lexicographic record of "torrão-de-açúcar" to Bluteau (1721: 214).

> (3) [...] Tem necessidade cada engenho de feitor, carpinteiro, ferreiro, **mestre de assucar** com outros officiaes que servem de o purificar; os **mestres de assucares** são os senhores de engenhos [...]. Padre Fernão Cardim (1980) [1583].

> (4) [...]. e como era mui limpa e branca e alem disso doce como um **torrão de assucar** [...]. Padre. João Felippe Betendorf (1910) [1699].

We may discern identical creative processes in the "cycle of gold". Gold mining began in the Captaincy of São Vicente, in the 16th century; however, given the profitability of sugar production, mining only became economically appealing as Brazil began to have competition in that activity. Therefore, the so-called "cycle of

[16] According to Houaiss (2001), the word "açúcar" dates from the 14th century.

[17] According to Bluteau (1712: 116), this type of sugar is the one that, after three or four boils, ends out very white and hard as rock.

gold" only peaked in the mid-18th century, as a result of the gold and diamonds mining in the States of Minas Gerais, Goiás and Mato Grosso. The depletion of the deposits occurred around 1760.

Of course, the specific context of mining work determined the use of a lexicon that, following the example of sugar terminology, was known and used particularly among those who dedicated themselves to this activity.

It should be noted, of course, a form of variation recorded in the *Textual Corpus Databank*: the word "ouro" (gold) has the variation "oiro"[18], being that the former, which is the oldest form occurring in Portuguese, emerges in the *Databank* for 3655 times (from 1500 to 1808) and the former, which is a more recent form, occurs only 278 times (from 1590 to 1804). There are many examples of linguistic variation that, such as the alternation between "ou" and "oi", still require further study, what clearly shows the linguistic value of the *DHPB* corpus for research in various areas of the history of language.

Let us attend on one of the oldest occurrences of "ouro" and "oiro":

(5) [...] pero huũ deles pos olho no colar do capitam e começou daçenar cõ amaão pera aterra e depois perao colar como que nos dezia que avia em tera **ouro** e tam bem vio huũ castical de prata e asy meesmo acenaua peraa tera e entã perao castical como que avia tan bem prata. Pero Vaz de Caminha (1964) [1500].

(6) [...] Por estes rios acham alguãs pedras e no rio de Janeiro ẽ hũ se achou **oiro**, afora as minas das terras ha muitos veados de muitas castas, corcos etc. Francisco Soares (1966) [1590].

In what regards the names related to the extraction of precious metals in Minas Gerais, it should also be noted that the unit "mina" (mine) has 424-recorded occurrences in texts ranging from 1562 to 1804. In the context of mining, in addition to the noun "ouro", the noun "minas" (mines) is the most frequent word, noting 4709 occurrences in texts, from 1560 to 1805; the singular noun, "mina", has, in turn, 424 occurrences (from 1562 to 1804) and the verb "minerar" (mining), 133 occurrences. According to *DHPB*, the units "mineralogia" (mineralogy) and "minerador(es)" (miner[s]), 'relating to the mine or the ones working there', are less frequent, considering they account for mere 17 and 3 occurrences, respectively. It should be noted that in the *Databank*, "mineralogia" is first recorded in 1653, and "minerador", only in 1760, which allows us to backdate these units, since both Cunha as Houaiss record much later dates: for "mineralogia", 1789 (Houaiss 2001) or 1813 (Cunha 1994: 522); for "minerador"[19], 1858 (Houaiss 2001).

[18] In the Portuguese language, this word can be pronounced in two different ways: as a diphthong [ow] that became a monophthong in [o], in the southern dialects, and in the standard language. Also, it can be pronounced as a diphthong [oj], mainly produced in southern varieties.

[19] A. G. Cunha (1994: 522) does not include the form "minerador" (miner) among those derived from "*mina*" (mine), yet considers the following units: *mineira, minero, mineração,*

In what concerns to the "cycle of coffee", first of all, it should be noted that the plant was introduced into Brazil in 1727 by Francisco de Melo Palheta (1670-?), who took the first seedlings from French Guiana. However, the interest in the planting of this species in Brazilian lands only increased around 1760, when the coffee plantations in Rio de Janeiro started. As in previous cycles –sugar and gold– the lexicon reflects the new realities and practices. Thus, from the word "café" (coffee) are formed new units to designate not only the plant but also the ground upon which it is grown extensively. From "café" are formed "cafezal" (coffee plantation), "cafezeiro" (coffee plants) and "cafeteira" (coffee pot), for example. Thanks to a document of 1757, by Fr. João Daniel, the *Databank* allows us to backdate the noun "cafezal", which means, a 'great coffee extension or coffee plantation', since Cunha (1994: 136), and Houaiss (2001) assign to it subsequent dates: 1776 and 1844. More significant still is the backdating of "cafezeiro", which means a 'bush or tree whose fruit is the coffee', which, with this meaning, and according to Houaiss (2001) and Cunha (1994: 136), is dated from 1836 and 1844, a chronology that can be revised thanks to *DHPB*, because this unit appears in a document of 1757, by the aforementioned Fr. João Daniel. However, in the *Databank,* "cafezeiro" still does not present the meaning 'coffee plantation owner or producer of coffee', to which Houaiss assigns the date of 1836. As the chronological span of the *DHPB* corpus ends by 1808, it is not surprising that this meaning is not yet represented in it, because the economic activity of planting coffee only had its peak in the mid-19th century, yet this does not mean that at the beginning of this century the word "cafezeiro" would not name already the "coffee plantation owner or producer of coffee".

On the aforementioned dates, we should underline that the proposed dates by Cunha and Houaiss are based on lexicographical works. Indeed, 1836 and 1844 correspond, respectively, to the first and the second edition of the *Novo Diccionario Critico e Eymologico* (1844), by Francisco Solano Constâncio (1777-1846).

The situations of backdating and the possibility of dating, for the first time, many other words, demonstrate well the documentary and linguistic value of *Textual Corpus Databank,* in which it is based the *Dicionário Histórico do Português do Brasil,* reason why this will substantially improve the condition of lexical chronology, one of the most problematic aspects of the history of Portuguese lexicon and, by extension, the lexicography of the Portuguese language. No less expressive are the so-called "brasileirismos", i.e. the words forged in Brazilian soil, from a Portuguese base, in order for naming local realities and practices, which, then, acquired new meanings or became integrated in terminological domains pertaining to the American context and, still, the words originated from the Brazilian indigenous languages –"indigenismos"– and those of African origin –"africanismos"–, which were conveyed by the slaves and incorporated into the Portuguese lexicon.

mineral, mineralogia, mineralógico, mineralurgia, minerar e minério (mining, mineral, mining, mineral, mineralogy, mineralogical, mineralogy, mining and ore).

Houaiss (2001) defines “brasileirismo”[20] as “any fact of language (phonetic, morphologic, syntactic, lexical, and stylistic) characteristic of Brazilian Portuguese”, which designates, under the lexical scope, any “word or phrase (vocabular dialectalism) or meaning (semantical dialectalism) specific to the Portuguese in Brazil”, definition that does not allude to the unit’s Amerindian origin (Dietrich and Noll 2010) nor those of African origin (Bonvini 2002; Bonvini 2008), although the use of such words is a distinctive trait between the Brazilian lexicon in face of the European Portuguese. In the next section, we will examine some examples of “brasileirismos” of the abovementioned types.

2.1. *Brasileirismos* in the *DHPB*

In the literature on the history of Portuguese lexicon, one of the most controversial aspects is the definition of “*brasileirismo*”, that is to say, the lexical unit whose usage is unique or characteristic of Brazil. Of course, phytonyms and zoonyms abound among “brasileirismos”, but the truth is that there are “brasileirismos” in the most diverse semantic fields and referents: landforms and orography, family, domestic life, food, clothing, utensils and devices, practices and techniques, among many other fields of Brazilian nature and culture.

The Viscount of Pedra Branca (1780? -1875), Ambassador of Brazil's Emperor at the Court of France, is the author of the first list of “brasileirismos”, which, according to the typology of the author, resulted from the meaning change of Portuguese words or were used exclusively in Brazil. This list was published in 1826 by Adrien Balbi, in the *Introduction* to his *Atlas ethnographique du globe* (Balbi 1826: 172-175). Examples of both kinds of “brasileirismos” are the units “babado” and “nuelo”.

The semantic change inherent to the transfer of grammatical class or “conversion”[21] is one of the Portuguese units’ transformations in Brazil, as it demonstrates the nominalization of the past participle “babado” (from verb “babar” 'to slobber'), which designates an “embellishment used in women's clothes or in the house”. Indeed, the noun “babado” is registered in the *DHPB*, where it is found in a document of 1791; however, in Houaiss (2001), this “brasileirismo” is dated from 1899. Confer the sample extracted from the *Databank*:

> (7) [...] Dois lençoes de bretanha de Hamburgo com **babados** de panno de linho aberto já usados em seiscentos reis $600 [...]. Joan B.a Lustoza (1936) [**1791**].

Similarly, in the *DHPB*, as we can see in (12), “nuelo(s)” appears in a document of 1770 with a meaning not used in Portugal, namely, “completamente nu; nu em

[20] According to Houaiss (2001), this unit will have been registered for the first time in 1899, no *Nôvo Diccionario* de Cândido de Figueiredo.

[21] About this process, we refer to Espinosa Elorza (2008: 176-178)

pêlo"[22] (completely naked; buck-naked) (Houaiss 2001), a definition that was marked in Houaiss as a "regionalism of Minas Gerais", dated from 1836.

(8) Entre o grande número de **nuelos** vinham 5, ou 6, com camisas muito alvas [...]. Afonso Botelho de S. Paio e Sousa (1962) [**1770**].

No less interesting is the case of "lambida", past participle of "lamber" (to lick), which, in Brazil, as an adjective or a noun, it is equivalent to "lambidela", that is, the 'action or effect of licking'. In our corpus, the word has a quite different meaning from that which the Portuguese-Brazilian lexicography presents. Indeed, in the *Databank* corpus, "lambida" means "empty, deserted", a meaning that is not mentioned by Houaiss (2001). Thanks to the *DHPB*, we may not only added a meaning to the word "lambida", as it was also possible to date it, because the document that attests it is from 1732, as it is shown in the following snippet.

(9) [...] Aonde ha hum monte cuja falda he. **lambida** de todo o genero de caça [...]. Francisco Tavares de Brito (1732) [**1732**].

If the examples adduced above reveal undoubtedly both the informational wealth, and the linguistic worth of the *DHPB*'s *Databank*, so do the words created in Brazil to designate non-existent realities in Portugal, as well as the "brasileirismos" of indigenous or African origin.

2.1.1. Lexical morphology

Many were the Portuguese words that were subject of derivation[23], thus giving rise to words that name Brazilian realities. Such is the case of "bandeirante", that is, the 'individual who, in the colonial Brazil, took part in "bandeira" ('expedition'), a word that is evidently derived from word "bandeira"[24] (flag), and that, in the *DHPB*, has a first attestation in 1722, dating well before the current lexicography, since both Cunha (1994: 96) and Houaiss (2001), sustained on Domingos Vieira (1871: 718) place the noun "bandeirante" in 1871. This example alone enables us onto assessing the huge documentary and linguistic value of the *Textual Corpus Databank*, in which the *DHPB* is based upon.

(10) [...] Ora, não há dúvidas de que de S. Feliz, nome d'um **bandeirante** que lá foi ter em 1728, naturalmente por informação colhida da bandeira do Anhanguéra [...]. José Peixoto da Silva Braga (1982) [**1722**].

[22] Not being a "brasileirismo" *per se*, however, "nuelo", as an adjective, means, "who or what is featherless, who or what does not have fur or feathers (it is said of a newborn bird)" (Houaiss 2001).

[23] Included in the so-called "multiplied lexicon" (Álvarez de Miranda 2008: 145), that is, one of the process of "lexicogenesis" (word formation).

[24] Each expedition to the interior of Brazil was a "bandeira" (flag) because each group was headed by a flag that would identify it.

2.1.2. Words of Amerindian origin

The Brazilian indigenous languages – the tupi, in particular (Dietrich and Noll 2010) – passed on to Portuguese many words belonging to the most varied referential fields, as it was pointed out earlier, even though not all have equal distribution. Thus, the lexical unit “nana” (or ananás[25] [pineapple, in Eng.]), which is a ‘plant belonging to the pineapple family, native to Brazil’, originates from the tupi, and is already attested in the *DHPB* through a document from 1585; however, Houaiss (2001) points out this word as dating from 1899; such a chronological gap renders the importance of *corpora* in lexicographic projects.

> (11) [...] Ha outros caraguatás que dão humas folhas como espadana muito comprida, de duas ou tres braças, e dão humas alcachofras como o **naná**, mas não são de bom gosto. Padre Fernão Cardim (1980) [**1585**].

On the noun “ananás”, as an annotation on its etymology, the Houaiss Dictionary (2001) points out that the word “abacaxi” (also pineapple) is not documented in Portuguese until the 19th century. Now, in the *DHPB* corpus, the word “abacaxi”, whose origin is tupi, also has its first attestation in 1757, as it is shown in the snippet extracted from the *Databank*, contrary to the assertion made in that important lexicographical work.

> (12) [...] e todas as suas folhas são do mesmo feitio, excepto, que as do ananás ordinário são lisas pelas bordas: as do ananás e **abacaxis** tem pelas bordas uma serrinha, que serra muito os dedos [...]. Pe. João Daniel (1976) [**1757**].

Even restricting our attention to the field of food, there are many “brasileirismos” of indigenous origin. Consider the cases of “mocotó”, “pamonha”, and “pupunha”, all of them recorded in the *DHPB* corpus.

The noun “mocotó”[26] comes from the tupi and it names the ‘veal shank, without the hull’. According to Houaiss, the unit was recorded in 1836 with such meaning, also referring to “chambaril” or “mão-de-vaca”, culinary terms that correspond to the dish made with that part of veal; however, the *DHPB* corpus allows placing the designation of this dish in 1735, what constitutes a remarkable backdating.

> (13) [....] ou comerá pés, e maõs de carneyro, ou em falta **mocotós** de boy, ou pés, e maõs que he o meſmo [....]. Luis Gomes Ferreira (1735) [**1735**].

In what regards to “pamonha”, which designates a ‘delicacy prepared with mashed corn, cooked and wrapped in its own straw or corn leaf’, according to our *Databank*, the noun was used in the colony of Brazil in 1749, exactly with that

[25] According to Houaiss (2001), the form “ananás” (pineapple) dates from 1557.

[26] The form is also mentioned in Houaiss as a "regionalism" from Trás-os-Montes, a Northern region of Portugal, to designate a "fat old man". Hence, the origin of the Portuguese word cannot be tupi, so the coincidence should be due to a homonymic collision.

meaning, hence it follows that the delicacy has an ancient tradition, reason why the dating of Houaiss –1877– is completely clueless of historical reality.

> (14) [...] Do bagaço que fica se fazem uns bolos, que metidos em folhas debaixo das brasas do fogão e assados são gostosos, e se chamam **pamonhas**. Caetano da Costa Matoso (1999) [1749].

In turn, the unit "pupunha", a 'variety of American palm tree and its fruit', which Houaiss dates from 1833, appears in our textual corpus in the 18th century, in documents of Alexandre Rodrigues Ferreira, who traveled about the *Amazonas* between 1783 and 1792, and described the region in detail. Based on this attestation, it is possible to correct the dating proposed by Houaiss –1833– which, in fact, was already being used through the last decades of the 1700's.

> (15) PUPUNHAS Similhante ao dendê de Angola; é palma vistosa, e muito mais quando tem fruto, sendo umas amarellas e outras encarnadas; a sua arvore é povoada de espinhos e por todos os talos: umas são oleosas, e outras não. Alexandre Rodrigues Ferreira [n.d.].

If the *DHPB* allows us to dating new meanings, correcting dates and assigning first dates to "brasileirismos" of indigenous origin, we can notice the same about many "brasileirismos" originated from African languages spoken by slaves, and which were assimilated into the Portuguese language in Brazilian ground. We will cover this subject in the following section.

2.1.3. African origin words

The African origin "brasileirismos" (Bonvini 2002; Bonvini 2008) are well represented in the *DHPB*'s *Databank*, which allows us to asserting that they were quickly assimilated. Such words account for the strength of several African languages taken to Brazil and how they managed to influence the Portuguese language that tended to be the hegemonic language since the mid-18th century. In fact, there is good sample of "brasileirismos" that came from the contact of Portuguese with Kimbundu, Kikongo, and Yoruba, all them African languages that contributed to the multiculturality of the lexicon existing then (c.XVI-XVIII), and still today, in Brazil.

Among this African root of the "brasileirismos", we would highlight the names of some of the Afro-Brazilian culinary specialties, namely, "abará", "acarajé"[27], "acaçá"[28], "angu", "bobó"[29], "dendê", "moqueca"[30], "pé-de-moleque[31]", "quibebe"[32] and

[27] The name of a "black-eyed peas cookie", perhaps from the Yoruba (Houaiss 2001).

[28] According to Houaiss, it is a denomination of "jeje" origin for "an Afro bahiano dumpling made from rice or corn flour, cooked in thick gelatin and involved, still hot, in banana leaves" (Houaiss 2001). In the *Dicionário Houaiss* it is dated from 1871.

[29] Recipe made with beans "cooked to the consistency of porridge, with palm oil and seasonings" (Houaiss 2001). The unit seems to have its origin in the language "jeje", from the region of Benin and neighboring regions.

"vatapá"[33]. Some of these nouns –"acarajé", "bobó", "pé-de-moleque" "quibebe", "vatapá"– received in Houaiss the date of 1899, when such words are, in fact, quite older in Portuguese, considering they were already present in a text from 1792:

> (16) [....] Quiabos, e carurús, De que se fazem jambês. Temos **quibebes**, quitutes, **Moquecas**, e quimgombôs, Gerzelim, bôlos d'arrôs, **Abarás**, e manauês. Temos a canjica grossa, Pirão, **bobôs**, **caragés**, Temos os jocotupés, Orapronóbis, tutús. Tambem fazemos em tempo Do milho verde, o corá, Mojangues, e **vatapá**, **Pés de muleque,** e cuscús.

In the case of "dendê" (17), originally from Kimbundu, which is a name that designates a 'variety of palm tree, its fruit, and the oil one extracts from it', and the "abará", which designates a 'serving of black-eyed peas, ground and seasoned"[34], Houaiss indicates the dates of, respectively, 1836 and 1871, but in the *DHPB*, these units have attestation in 1765 and 1792.

> (17) [...] No Brazil se conhecem Coco, Coquilho, Andaiá, Andaiaoasú, Andaiá Merim, Agoasú, Bacori, **Dendê**, Aguê, Jaraiva, Penohi, Airi, Mocaiuba, Seriba [...]. Anônimo (muito provavelmente Joseph Barbosa de Sáa) (1999) [**1765**].

Indicated by Houaiss as a 'baiano' regionalism, the word "acarajé" emerges on the *DHPB* in 1802, as noted in the excerpt below:

> (18) [...] pamonhas, cangicas; isto é, papas de milho, acassás, **acaragés,** abarás, arroz de côco, feijão de côco, angús, pão de ló de arroz [...]. Luiz dos Santos Vilhena (1921) [**1802**].

Lastly, "angu", which is a word of obscure origin, signifies a 'thick dough that one prepares by mixing, under fire, cornmeal ("fubá", or maize flour), cassava ("mandioca") or rice, with water and, sometimes, salt', is dated from 1799 in Houaiss; however, the *DHPB* attests that unit in a 1749 document.

[30] Designation of a "stew of fish, seafood, meat or eggs, made with coconut milk, palm oil and plenty of spice" (Houaiss 2001). The unit is originated from the Kimbundu.

[31] Designation of a "hard-candy sweet made from sugar or brown sugar with roasted peanuts" (Houaiss 2001). The word is originated from Kimbundu.

[32] Name of a "recipe made of mashed pumpkin, to which one may add coconut milk" (Houaiss 2001). From the Kimbundu. Houaiss also indicates a Portuguese meaning: "porridge of anything else".

[33] Houaiss points out this word as regionalism from Bahia. This is a "highly prized delicacy, which is based on softened bread (or wheat flour) to which one adds shredded fish meat, fresh shrimp, dried shrimp and seasonings".

[34] The seasoning consists of "salt, onion, palm oil and dried shrimp" (Houaiss 2001), being the portion of beans cooked in a water bath or steam, and subsequently wrapped in banana green leaf".

(19) [...] Dele se faz o *fubá, assim chamado nas Minas e, em Portugal, farinha; deste, o **angu** para os negros, cozido em um tacho de água até secar; só se diferencia da broa em ser esta cozida no forno e levar sal. Caetano da Costa Matoso (1999) [**1749**].

As it happens with the "brasileirismos" of Amerindian origin, those of African origin above do not incorporate the European Portuguese lexicon, even though some of them, referring to gastronomic specialties broadcast on television or brought to Europe by Brazilian immigrants, are already known to the speakers of the European variety.

Conclusion

A dictionary is, by nature, a work in permanent construction, since neither the nomenclature nor the lexicographic utterance are final, and someone can always put a date on certain meanings or complex units; it will always be possible to backdate other words or assign a date to many others.

The data extracted from the *Textual Corpus Databank* allow us to conclude, without the slightest doubt, that a historical dictionary must be based on a broad, diversified and representative corpus of a given chronological period, since this is the only way to remediate the shortcomings of the lexical diachrony and the chronology of words. Indeed, only the use of texts produced within the various synchronies of the given period, texts that are representative of various textual and speech genres will make possible a historical knowledge of both the Brazilian lexicon during the colonial period and, of course, the Portuguese lexicon.

In the case of the *Dicionário Histórico do Português do Brasil*, the three centuries represented in the corpus reflect the standing of the Portuguese lexicon since the end of the medium Portuguese until the classic or modern Portuguese, including, therefore, the 18th century, a century which, as it is well known, has the utmost importance to understand the drift of the Brazilian Portuguese. Yet, the historical value and the linguistic corpus of the *DHPB* does not only manifests itself in the studies of lexical chronology, for the *Databank* is equally valuable to the study of such specific areas as the lexical morphology, as the examples adduced in 2.2.1. illustrate so well, which reveal a self-reliance in terms of suffixal derivation process, or the study of syntax in the period under consideration.

On the other hand, the sample analyzed in this work also demonstrates that contemporary lexicography should be increasingly based in diversified textual corpora, so that certain units may not be attested by lexicographical works alone, as we have seen in several of the analyzed samples. In fact, querying the *Dicionário Houaiss* (2001) revealed that, despite of this dictionary being today a reference in the lexicological and lexicographic fields, it produces very out-of-phase dating of the history of lexicon in general and, in particular, of the indigenous and African-influenced "brasileirismos". These findings highlight the many things that must be done not only in the next Houaiss editions, but also in the study of the History of the Lexicon and of Historical Lexicology, areas that are now being enhanced as a result of corpora Linguistics.

The *Databank* and the *DHPB* are excellent examples of both the challenges underlying the elaboration of historical *corpora*, and the advances that these allow concerning the historical knowledge of the Portuguese lexicon, under various different perspectives, whether synchronic, diachronic, terminological, morphological or phraseological, between the 16th century and the end of the 18th century.

Despite the benefits corpora Linguistics brought in, the truth is that neither the *DHPB* nor any other current lexicographic project will refute the epigraph that presides over the *Dicionário Histórico do Português do Brasil*:

> The making of a dictionary is an exceedingly laborious matter that requires, in addition to scientific capabilities as spectacular as sharpness of mind, fantasy, coherence and critical judgment, many discrete virtues, related to those of craftsmen, such as patience, diligence, perseverance, precision in details and – last but not least – a great collector's passion.
>
> (Harald Weinreich)

References

Álvarez de Miranda, P. (2008). "Neología y pérdida léxica". In: Miguel, E. de (ed.). *Panorama de la lexicología*. Barcelona: Ariel. 133-158.

Antonil, A. J. (1710). *Cultura e opulência do Brasil por suas drogas, e minas* [...]. Lisboa: No Officina Deslandesiana.

Balbi, A. (1826). "Introduction". In *Atlas ethnographique du globe* [...]. I. Paris: Rey et Gravier. 172-175.

Bluteau, R. (1712-1721). *Vocabulario Portuguez, e Latino*. Tomos I, II, (1712); III, IV, (1713); V, (1716); VI, VII, (1720); VIII, (1721). Tomos I-IV, Coimbra: Collegio das Artes da Companhia de Jesus; Tomos V-VIII. Lisboa: Officina de Pascoal da Sylva.

Boléo, M. de P. (1943). *Brasileirismos (Problemas de Método).* Coimbra: Coimbra Editora.

Bonvini, E. (2002). "Palavras de origem africana no português do Brasil: do empréstimo à integração". In Nunes, J. H.; Petter, M. (eds.). *História do saber lexical e constituição de um léxico brasileiro.* São Paulo: Humanitas/FFLCH. 147-162.

Bonvini, E. (2008): "Línguas africanas e português falado no Brasil". In Fiorin, J. L.; Petter, M. (eds.). *África no Brasil: a formação da língua portuguesa.* São Paulo: Contexto. 15-62.

Bosque, I. (1982). "Sobre la teoria de la definición lexicográfica". *Verba*. 9. 105-123.

Castillo Carballo, Mª. A. (2003). "La macroestrutura del diccionário" In Medina Guerra, A. M. (ed.) *Lexicografía española*. Madrid: Ariel. 79-101.

Castro, I. (1999). "O Português Médio segundo Cintra (nuga bibliográfica)". In Faria, I. H. (ed.). *Lindley Cintra. Homenagem ao Homem, ao Mestre e ao Cidadão.* Lisboa: Cosmos. 367-370.

Castro. I. (2006). *Introdução à história do português.* [2ª ed.] Lisboa: Colibri.

Constâncio, F. S. (^{1}1836[^{2}1844]). *Novo Diccionario Critico e Etymologico da Lingua Portugueza*. [2ª ed.] Paris: Angelo Francisco Carneiro, Editor Proprietario.

Cunha, A. G. (1994). *Dicionário etimológico Nova Fronteira da língua portuguesa.* [2ª ed.] Rio de Janeiro: Nova Fronteira.

Cunha, A. G. (1996) *Dicionário Etimológico Nova Fronteira da Língua Portuguesa*. [2ª ed.] Rio de Janeiro : Editora Nova Fronteira.

Cunha, C. (1987). *Que é brasileirismo?* Rio de Janeiro: Tempo Brasileiro.

Dietrich, W.; Noll, V. (2010). “O papel do tupi na formação do português brasileiro”. In Noll, V.; Dietrich, W. *O português o tupi no Brasil*. São Paulo: Editora Contexto. 81-103.

Dubois, J.; Dubois, C. (1971). *Introduction à la lexicographie: le dictionnaire*. Paris: Librairie Larousse.

Espinosa Elorza, R. M. (2008). “El cambio semántico”. In Miguel. E. de (ed.). *Panorama de la lexicologia*. Barcelona: Ariel. 159-188.

Figueiredo, A. C. (1899). *Nôvo Diccionário da Língua Portuguêsa.* 2 vols. Lisboa: Tavares Cardoso & Irmão.

Garriga Escribano, C. (2003). "Microestrutura del diccionário: las informaciones lexicográficas". In Medina Guerra, A. M. (ed.). *Lexicografía española*. Madrid: Ariel. 103-126.

Gonçalves, M. F. (2006). “A marca lexicográfica «Termo do Brasil» no Vocabulario Portuguez, e Latino de D. Rafael Bluteau”. *Alfa – Revista de Lingüística*. 50(2). 205-228.

Gonçalves, M. F. (2012a). "Aspectos do léxico português e brasileiro no século XVIII: pesos e medidas no «Erário Mineral» (1735), de Luís Gomes Ferreira". *Confluência - Revista do Instituto de Língua Portuguesa*. Rio de Janeiro. nº 43 (2º semestre de 2012). 47-67. http://llp.bibliopolis.info/confluencia/pdf/651.pdf [Access date: 10 October, 2013]

Gonçalves, M. F. (2012b). “La terminología azucarera en Brasil: el testimonio de los lexicógrafos Rafael Bluteau y António de Morais Silva”. In Viña, A.; Corbella, M. D. (eds.). *La ruta azucarera atlántica: historia y documentación*. Funchal: Centro de Estudos de História do Atlântico. 101-132.

Haensch, G. (1982). "Tipologia de las obras lexicográficas". In Haensch, G. (et al.) (eds.) *La Lexicografia* (eds.). *De la Lingüística Teórica a la Lexicografia Práctica*. Madrid: Editorial Gredos. 95-187.

Houaiss, A. (2001). *Dicionário eletrônico Houaiss da língua portuguesa* [CD-ROM. Versão 1.0.]. Rio de Janeiro: Objetiva.

Imbs, P. (1960). "Au seuil de la lexicographie." *Cahiers de Lexicologie*. Vol. 2. 3-17.

Krieger, M. G.; Finatto, M. J. B. (2004). *Introdução à terminologia. Teoria & prática.* São Paulo: Editora Contexto.

Lucchesi, D. (2012). “A diferenciação da língua portuguesa no Brasil e o contato entre línguas”. *Estudos de Lingüística Galega*. 4. 45-65.

Mattos e Silva, R. V. (2004). *Ensaios para uma sócio-história do Português Brasileiro.* São Paulo: Parábola.

Mattos e Silva, R. V. (2006). *O português arcaico: fonologia, morfologia e sintaxe.* São Paulo: Editora Contexto.

Murakawa, C. de A. A. (2010). “Dicionário histórico do português do Brasil: um modelo de dicionário histórico”. *Filologia e Linguística Portuguesa*. 12(2). São Paulo. 329-349.

Murakawa. C. A. A. (2012). “Trois siècles de mots du portugais du Brésil”. *Cahiers de lexicologie*. 101(2). 73-91.

Naro, A.; Scherre, M. M. (2007). *Origens do português brasileiro.* São Paulo: Parábola.

Noll, V. (2012). “Para uma revisão do dicionário Houaiss – Vocabulário e datações”. *Confluência – Revista do Instituto de Língua Portuguesa*. 43, 2º semestre de 2012. 68-77. http://llp.bibliopolis.info/confluencia/pdf/653.pdf [Access date: 08.09. 2013]

Oliveira, A. M. P. P. de. (1998). “Brasileirismos e Regionalismos”. *Alfa – Revista de Lingüística.* São Paulo: UNESP. 42. 109-120.

Porto Dapena, J-A. (2002) . *Manual de técnica lexicográfica*. Madrid, Arco/Libros, S.L.

Rey-Debove, J. (1984) *Léxico e Dicionário*. Trad. Clóvis Barleta de Moraes. Revista Alfa. Vol. 28 (sup.), São Paulo. 45-69.

Silva, A. de M. (1922, ^{2}1813). *Diccionario da Lingua Portugueza recopilado dos vocabularios impressos até agora, e nesta segunda edição novamente emmendado, e muito accrescentado por* .... Lisboa: Na Typographia Lacerdina. Fac-simile da segunda edição

(1813). Freire, L. (dir.). Rio de Janeiro: Officinas da S. A. Litho-Typographia Fluminense. 2 v.

Silva, A. de M. (^{1}1789): *Diccionario da Lingua Portugueza*, 2 vols. Lisboa: Officina de Simão Thaddeo Ferreira.

Silva, A. M. (s/d) *Diccionario da língua portugueza*. [9 ed.]. Lisboa: Editora Empreza Litteraria Fluminense de Santos, Vieira & Commandita. 2v.

Vieira, Fr. D. (1871-1874). *Grande diccionario portuguez ou thesouro da língua portugueza.* Porto: Editores Ernesto Chardron e Bartholomeu Moraes. 5 vols.

Vocabulário Ortográfico da Língua Portuguesa. (2009). [5 ed.] Editora Globo.

Completing the unfinished: a descriptive dictionary of the Croatian Literary Language

Ivana Filipović Petrović

1. Introduction

Although the Croatian language is represented in works of European bilingual and multilingual lexicography dating back to the 16th century, monolingual lexicography in Croatia did not begin to develop until the 20th century. In 1949, Julije Benešić, a renowned Croatian translator, lexicographer, literary critic and writer, was entrusted with the mission of compiling a dictionary of the contemporary Croatian Literary Language. Aspiring to present a portrait of the Croatian Literary Language in the process of its development, Benešić decided to base his dictionary upon quotations excerpted from the works of 113 of the most distinguished Croatian authors who were active writers between the mid-nineteenth and the mid-twentieth century. Benešić worked on collecting quotations and defining the corpus until his death in 1957, leaving an unfinished dictionary with the word *serenade* as the last completed entry (Benešić 1985: XXVII). Some thirty years after the lexicographer's death, the first twelve volumes of the *Dictionary of the Croatian Literary Language from the Revival to Ivan Goran Kovačić*[35] were finally published. The most recent volume, PROTIVAN–RZATI, was published in 1990. Given that the aspiration for the completion of the dictionary was never completely extinguished, and that the excerpted material used in the dictionary was stored at the Linguistic Research Institute of the Croatian Academy of Sciences and Arts, in 2008 extensive work on the dictionary's completion began anew.

In this paper, we will indicate the main typological characteristics of Benešić's dictionary which make it a unique accomplishment of Croatian lexicography and point to the necessity of its completion. For the lexicographers who are working on this task, these typological properties also bring up questions and dilemmas that have not yet been solved by contemporary lexicography. Foremost, the dictionary is

[35] Ivan Goran Kovačić being the last of the 107 Croatian writers whose works are cited in the *Dictionary*.

entirely the creation of its author. Although the history of lexicography shows that many lexicographers rely on their predecessors (Béjoint 2010: 62), Benešić used very few existing sources and consulted them only for the sake of comparison. There are two reasons for this. Firstly, despite some comparable works, Benešić's dictionary is typologically unique in Croatian lexicography. Secondly, Benešić's biography reveals a man who relied on his own linguistic knowledge, experience, and intuition. Benešić aimed to produce a descriptive dictionary that would be open to the corpus it was based upon, a dictionary that «lists those words which our writers have used in the past one hundred years, words with obsolete contents and forms, words that have become extinct, as well as those which are new. There are no rules or instructions on how writers should write correctly; only examples illustrating usage up until today are given» (Benešić 1985: XXV). As a result, he designated word senses and wrote definitions according to the quotations, but also according to his own judgement. Thus, the present-day project is confronted with the problem of reconciling two different aspirations: the goal of finishing the dictionary the way the author would have done so, and the urge to amend his misjudgments (Nikolić-Hoyt 2010: 8).

In this paper we will first say a few words about the *Dictionary's* progress from the idea of its compilation to present. Then we will show, through examples, Benešić's lexicographic methods. In the end, we will illustrate lexicographic solutions from the new volumes as well as indications for a more systematic approach in the future.

2. A glance into the past: the creation of The Dictionary

In order to understand Benešić's *Dictionary* as well as the problems that his follwers have to deal with while trying to complete it, we need to mention that Julije Benešić (1883 – 1957) was a Geography teacher who soon abandoned his profession to become a poet and a novelist, editor, linguist, translator, theatrologist and polonist. Thanks to his involvement in editing the works of renowned Croatian authors, many of whom were also his contemporaries, Benešić had an excellent knowledge of literary works. He came into contact with lexicography while writing his *Croatian-Polish Dictionary* and, earlier, the dictionary of the dialect of his hometown, *The Dictionary of Local Speech of Ilok* (manuscript). At the age of 65 he had in front of him a new extensive lexicographic task to which he ardently devoted himself. It is therefore not surprising that, in defining the guidelines for *The Dictionary's* compilation, Benešić started from his position of a novelist and clung to his personal feeling for language paying not too much attention to lexicographic standards. When he handed in *The Dictionary's* manuscript up to the letter K in 1954, the principles of interpretation seemed to him clearly defined. They reflected the character of both *The Dictionary* and its author: Benešić, the poet, editor and translator brought to the task of compiling the dictionary primarily his passion for the written word and his capacity for a profound enjoyment of literary works. The circumstances in which *The Dictionary* was compiled and which had impact on its character are also mentioned in the foreword to *The Dictionary's* first volume: »its [*the Dictionary's*] departure from certain rules which apply in strictly philological lexical works should not confuse: from the very beginning the dictionary was imagined as a sort of lexical anthology of Croatian literature in the last hundred and fifty years« (Benešić 1985: Foreword).

Envisaged, therefore, as the dictionary of Croatian literary language, i.e. as the collection of quotations from the works of the finest Croatian authors who were active writers during the stormy one hundred years period in the history of Croatian literature and language, *The Dictionary* aimed to show the development of Croatian literary language:

> The purpose of this Dictionary is to present a portrait of the literary language which Croatian writers used from the period of Illyrian movement (1835) till the beginning of World War II, exactly: in the period of one hundred years, and with quotations accompanying every word, in the form in which that word was used in the works of Croatian poets and belletrists (Benešić 1985: XVII).

There existed a certain disagreement between Benešić and Yugoslav Academy of Arts and Sciences, which was the publisher of *The Dictionary*, or more precisely between the reviewers appointed by the Academy to evaluate the manuscript that Benešić handed in. In his description of the principles of the treatment of the entries, Benešić, discussing the interpretation of meaning, emphasised his belief that explanations should be kept to a minimum. In other words, it is not necessary to explain some words since Croatian language is not a foreign language to the reader of *The Dictionary* so there is no point in describing what *leg, nose, church, mass* is, or what *happiness, nobleness, wrath, work, think, write* etc. means (Benešić 1985: IX). He therefore concluded that the words will mainly be explained through synonyms and examples from the texts: «I repeat: the reader of the dictionary is familiar with and speaks Croatian language and the dictionary is not really the same as a comprehensive grammar book» (Benešić 1985: IX).

Furthermore, Benešić explained that he did not cite the title and the page of the literary work from which the confirmatory quotation was excerpted but only the name of the author out of consideration for the space available. Those who might be interested in more details are referred to the notes containing such information which constitute the material collected for *The Dictionary* and which can be found at the Linguistic Research Institute of the Croatian Academy of Sciences and Arts. Although he was asked, in the reviews written for the Academy by linguistic experts, to accompany every word with an explanation, Benešić in his answer further explained the main characteristics and the purpose of the dictionary he was compiling at the same time sticking to his choice.

> The dictionary of any literary language cannot be normative like an instruction how to write correctly.[...] The dictionary of a literary language is informative (and therefore not normative). Such is this dictionary that I am compiling and editing (Benešić 1985: XXVI).

Moreover, the philologists reviewers insisted that the title, edition and page number of the literary work to which the examples refer be indicated alongside the author so that confirmatory quotations may be checked. In the light of contemporary state of lexicography but also in the light of the fact that today's users are considerably better informed, the final decision to leave out the data about the title, the year of publishing and the page number of the literary work caused massive damage[36] to Benešić's *Dictionary* (cf. Nikolić-Hoyt 2010). The polemics about the principles of dictionary compilation were ended by this and Benešić continued to work on *The Dictionary.* He reached the word *serenade* when his work was cut short by death before Christmas in 1957.

Almost thirty years after the author's death, the first volume of Benešić's *Dictionary* was finally printed in 1985. Eleven more volumes followed after that and the last, the twelfth volume (P–R) was published in 1990. After another period of *silence*, at the Linguistic Research Institute of Croatian Academy of Sciences and Arts, the work on completing *The Dictionary* was resumed in 2008. This work is based on accepting and carrying out the guidelines defined in 1948 at the beginning of the project and it sets out two conditions: the dictionary should be completed as if Benešić had completed it himself but without repeating the mistakes related to inconsistent treatment. Apart from the fact that reconciling these two aspirations represents a great challenge, we should also take into account that the amount of material waiting to be interpreted is greater than the material contained in the first twelve volumes of *The Dictionary*, which undoubtedly points to the conclusion that the task of dictionary completion is very extensive and time-consuming.

3. Benešić's lexicography

Before we turn to the problems concerning the completion of the dictionary, it is necessary to sum up its main features, which follow from the short description given in the first chapter and should be preserved in the new volumes as well. Foremost, the dictionary is entirely the creation of its author. Although the history of lexicography shows that many lexicographers rely on their predecessors, Benešić is a bit different in this respect. He used very few existing sources and consulted them only for the sake of comparison. There are two reasons for this. Firstly, despite some comparable works, Benešić's dictionary is typologically unique in Croatian lexicography. Secondly, Benešić's biography reveals a man who relied on his own linguistic knowledge and intuition. Emphasising that he was compiling a dictionary of a literary language which does not have any normative pretensions but only aims to give the examples of usage in literary works and shape the meaning of the words accordingly, Benešić relied on his instinct and experience in dealing with linguistic and cultural matters. For that reason lexicographic solutions in *The Dictionary* are mostly his own. The way in which Benešić's interests intertwine throughout his entire work has been most strikingly described by Tomislav Sabljak:

> To Benešić, a word is not something static – language moves like an actor on the stage and it is important to develop continuously that impression of balance, harmony, rhythm, word order, accent, feelings and ideas as a unique stage expression. The

[36] Since 2008, as agreed with the supervisor of the project of Benešić's *Dictionary* completion, in the new volumes, starting from the entry S, complete bibliographical details are provided for every quotation.

> experience of theatre and stage helped Benešić immensely in his work on dictionaries. Besides already mentioned practical meanings, Benešić also sought poetic characteristics of words, their metaphysical sign (Sabljak 1994: 92).

In addition, Benešić's *Dictionary* describes the language used by Croatian authors in the period between the Illyrian movement and the mid-twentieth century. That is why we can find in it the words which distinguished writers of that time but which contemporary dictionaries, compiled according to lexicographic norms, do not contain (see more in Nikolić-Hoyt 2010: 83).

Finally, the work methodologically comparable to Benešić's *Dictionary* appeared in England some two hundred years earlier and further development of English lexicography was based on it – Johnson's dictionary. Its main principles are still applied, especially in the *Oxford English Dictionary*. Moreover, a common feature of both Benešić and Johnson, the well-foundedness of every lexicographic solution in the language itself, i.e. in the real examples (Nikolić-Hoyt 2010: 81), is also a characteristic of modern dictionaries which are compiled according to the principles of corpus lexicography. Benešić paid special attention to the selection of quotations, i.e. the examples of usage:

> Compiling the dictionary of literary language is not just about entering certain individual words. The most important is to show a particular word with its usage, within a phrase in which that word has a central position. The dictionary will record the word with the attire in which the author had dressed it. It will record the phrase, not just a naked word (Benešić 1985: XXIV).

> ... because words only come to life in quotations, and the word in a quotation is the focus of the idea put forward in it. The word alone, with the explanation of its meaning still does not give the idea of its specificity if that word is not presented within some, albeit longer quotation (Benešić in Sabljak 1977: 11).

Relying therefore solely on his corpus glosses, Benešić only presented those meanings for which he had valid verification. If, for instance, a word in the corpus appeared only in figurative sense, Benešić recorded just such secondary but verified meaning. Hence in his *Dictionary* we find certain meanings, frequently metaphorical and metonymic, which words can have in a specific context. Nevertheless, reading *The Dictionary* and analysing the examples, we quickly notice that Benešić's approach to the treatment of data and the use of metalanguage lacks methodical quality and consistency.

> (1)
>
> **kudjelja,** f., konopljene niti (u Tordinca metaforički: žena).
>
> [**tow**, f. thread of hemp (in Tordinac metaphorically: woman).]
>
> Dade mi čitav tovar kudjelje, da je ispredem [He gave me a whole pile of a tow to twine] *(Šenoa).* – »Otkada u Hajdukovićevoj kući kudjelja zapovijeda, a mač sluša? Naruga se zlobno Ivan« [»Since when in Hajduković's home does a tow command, and a sword listen?« Ivan mocked viciously] *(Tordinac).*

(2)

kaciga, f. šljem (u Polića u prenesenome smislu: redar).

[**helmet**, f. casque (in Polić in a figurative sense: guard).]

U ovo krvavo, prokleto doba, kad hodaju ljudi u kacigama i noževi bliješte [In this bloody, damned era, when people walk in helmets and knives glint] *(Krleža)*. – Mrko ga gleda kaciga zakona, ne zna za proljetnu šalu [The helmet of the law looks at him moodily, he knows no spring prank] *(Polić)*.

(3)

atmosfera, f., parokrug, ozračje, obično u prenesenom značenju: okolina.

[**atmosphere**, f., the air round the Earth, usually in the transferred meaning: environment, ambience.]

... da si ne ojadi ugodne atmosfere, u kojoj je sada živjela, ostavi ona svoga muža ne odgovoriv mu ni riječi [In order not to spoil the atmosphere in which she now lived, she left her husband without answering a word] *(J. Kozarac)*. – Polemike i svađe, što ih je izazvao prvi broj, uzburkale su još više atmosferu [Controversies and disagreements, which were the result of the first number, stirred the atmosphere further] (*Bartulović*). – Sva je atmosfera bila nabijena strahotama [The entire atmosphere was charged with horrors] *(Kaleb)*. – Ali ne ona tišina poslije bure i kiše, kad je zrak svjež i čist, niti ona pred oluju, kad je atmosfera puna elektriciteta, već ona vajna tišina puste sobe, iz koje se netko iselio [But not such silence as after the wind and rain, when the air is fresh and clear, nor the silence as before the storm, when the atmosphere is full of electricity, but this sham silence of an empty room from which somebody moved out] *(Kolar)*.

(4)

baterija, f. [**battery**, f.]

više topova (4–8) iste vrste [several large guns used together].

Tom je baterijom Ivan u odlučivi čas najvećom hladnokrvnošću obližnju jednu

bateriju rusku demontirao [With that battery had Ivan, at the decisive moment and with the utmost composure, dismantled a nearby Russian battery] *(Bogović)*.

u prenesenom smislu [in the transferred sense].

Društvo se nabrzo razigra, osobito kad su zagrmile prve baterije šampanjca [The group started to play in no time, especially after the thunder of the first batteries of champagne] *(Becić)*.

(5)

ćuskija, f., *željezna motka, ovdje u prenesenom značenju (prost, pijan kao batina, kao ćuskija).*

[**jemmy**, f., *a long narrow piece of metal, here in the transferred meaning (rude, drunk as a jemmy).*]

Vi se đaci rugate oficirima, da su ćuskije i šta ti još sve ne znam [You, schoolboys, mock the officers, calling them jemmies and God knows what else] *(Matoš)*. – Više puta napijem se kao ćuskija i ni đavola mi nije [I often get drunk as a jemmy and nothing ever happens to me] *(Kosor)*.

(6)

isprašiti, *u metaforama* [**dust off**, *in metaphors*]:

išibati [*flog, lash somebody*]

Uhvati dakle ljuti Bjesomar bijesa za rog, podigne ga u zrak i ispraši ga dobro brezovačom [Terrible Bjesomar took the fury by horn, lifted him up in the air and flogged him well with broom] *(Brlić-Mažuranić)*.

prekoriti, ispsovati [*reprimand, scold*]

Dandolo je grofa molio, da oprosti slučajnu nepriliku, koja izvire iz neznanja gubernatorova – a samoga gospodina Taddea dobro isprašio radi njegove prevelike brige na krivu mjestu [Dandolo begged the count to forgive the unintentionally caused trouble which stemmed from his ignorance – and he scolded Mr Tadde for his disproportionate and misplaced care] *(Nehajev).*

Such examples allow us to conclude that Benešić approached and dealt with every entry and quotation(s) which exemplify its meaning individually, using different labels and terms which he chose arbitrarily as *metaphorically* in example (1), *in the transferred sense* in the examples (2) and (4), *usually in the transferred meaning* in (3), *here in the transferred meaning* in example (5) and *in metaphors* in example (6). We should also mention that Benesic indicated in the explanation of certain entries which writer uses the word in the transferred sense, as in examples (1) and (2), but we do not find the same method elsewhere. Hence, the examples, suggest two conclusions. Firstly, consistency of lexicographic treatment and uniformity in the selection of labels are not the characteristics of this *Dictionary*. Secondly, despite his lack of lexicographic consistency, it is evident that Benesic noticed and separated figurative meanings, indicating that there is some sort of semantic exentesion at work, most frequently metaphors. In this therefore lies the difficulty of reconciling two aspirations that those who are working on the dictionary's completion have: the successors need to finish the dicitionary the way the author would have done but they also need to avoid the mistakes of an inconsistent treatment. In the next chapter we will show potential solutions to the completion of Benešić's *Dictionary*.

4. Continuation in the next century

By the time of its compilation, Benešić's *Dictionary* was conceived as a dictionary of contemporary language which was missing in Croatian lexicography. Regarding the method of excerpting quotations from literary works which serve to exemplify the meaning, it fitted well into European lexicographic tradition which yielded similar works a few centuries earlier. In spite of the fact that literary quotations were used in English lexicography as early as 1598 in Florio's work *Worlde of Wordes* (Landau 1984: 55), English lexicography considers Samuel Johnson to be the pioneer in this sense, since he used the quotations more systematically than his predecessors, recording the quotation for each word and meaning in his dictionary from 1755. Dictionaries of literary language were also known in Spain and Italy[37] in the early seventeenth century, and on the title page of Richelet's work *Dictionnaire françois* (1680) it was written that the dictionary was «drawn frome the usage of the best authors of the French language.» Curiously, French Academy decided to leave out literary quotations from the dictionary opting for a syncronic dictionary of everyday language and not a historical dictionary of literary language (Béjoint 2010: 78). However, literary quotations have ever since remained a significant feature of

[37] *Tesoro* (1611) and *Vocabolario* (1612).

all major dictionaries and they continue to be used for various reasons: to show that a word exists, to show that it is used by notable authors, to show in what sort of context it is used, to introduce extra information, to mention famous literary passages, to serve as vehicles for values and opinions, or to make the dictionary more attractive (Béjoint 2010: 78).

Although Benešić's *Dictionary* is nowadays a historical dictionary for contemporary Croatian lexicography, its valuable characteristics, which were stressed out in the time of its creation, are still important. Namely, Benešić's *Dictionary* has remained a special dictionary which Croation lexicography does not have, a dictionary in which all words are verified by the quotations from literary works and a dictionary which gives a full portrait of one hundred years period in the development of Croatian literature and Croatian literary language from mid-nineteenth to mid-twentieth century. Today, it can serve not only as source of literary inspiration but also as a basis for linguistic reflections on the abundance and diversity of Croatian linguistic treasure once and today (Nikolić-Hoyt 2010: 86).

The issue of its descriptive nature, which we have dealt with in this article, means that the meanings should be shaped according to quotations, or in other words, that it is necessary to note the meanings which a word can have in a specific context, and possibly only in it and only in that quotation because such a meaning is the result of the author's current inspiration. This is most frequently the case with metaphors and metonymies. At first glance it may seem that polysemy in a descriptive dictionary such as Benešić's, in which the meanings are determined by confirmatory quotations, does not represent such a problem as in normative dictionaries which aim to cover all possible meanings of a particular word. Nevertheless, quotations offer countless possibilities of usage due to the creativity and poetic licence of the authors and make it difficult for the lexicographers to fit them in the framework of conventional lexicographic methods. Two terms are used for figurative, i.e. metaphoric and metonymic meanings in the new volumes of Benešić's *Dictionary*: the abbreviation *fig*. (*pren*.) and a phrase *here in the meaning*, depending on the type of the transfer of meaning. The meanings, i.e. the usages which are closely connected to the cited context are explained with *here in the meaning*:

(7)
slamnjača, f. [**pallet**, f.]
dio ležaja ispunjen slamom [*a cloth bag filled with straw, used for sleeping on*].

Imao starac u njoj krevet, običan željezan pandurski krevet sa visokom tvrdom slamnjačom... [In it, the old man had a bed, an ordinary iron cop bed with a raised, hard pallet...] *(Gjalski, Little Tales I, 1894, 162).* – Najprije su namjestili krevet, slamnjaču, perjanku, jorgan... [They first made the bed,then the pallet, featherbed, quilt...] *(Bertić, Female Lots, 1902, 20).* – Teška je to bila noć i preteška. Ležao sam na svojoj slamnjači kao drvo... [It was a difficult night, a really tough one. I was lying on my pallet like a log...] *(Šenoa, Baron Ivica, 1932, 72).*

2. ovdje u značenju *vojnik* [here in the meaning *soldier*].

Mi ne ćemo da budemo više mrtve brojke, jednake kape i slamnjače bez imena [We will no longer be mere numbers, identical caps and pallets with no name] *(Lovinac, Purple Nights, 1914, 107).*

(8)

skakavac, m. [**grasshopper**, m.]

1. zool. *kukac koji visoko skače* [zool. *an insect with long back legs, that can jump very high*].

I tako je taj starčić u razgovoru poput skakavca skakao sada ovamo, sad onamo, bez ikoje suvislosti i logike [And so that little old man jumped in conversation like a grasshopper, hither and thither, without any coherence and logic] *(J. Kozarac, Between the Light and Darkness, 1891, 33)*. – Jer skakavci, ako nalete, izjedu sve, što je zeleno! [For grasshoppers, if they come, eat up everything that's green!] *(F. Mažuranić, From Dawn to Night, 1927, 165)*.

2. ovdje u značenju *dijete* [here in the meaning *child*].

Kad čovjek ima kod kuće ženu i devet skakavaca, onda je za nj prestala radost života [When a man has a wife and nine grasshoppers at home, then there is no joy of life for him any more] *(Novak, Tito Dorčić, 1906, 107)*.

(9)

savinuti se savinem se, ovdje u značenju *pokoriti se, podčiniti se.*

[**bend,** here in the meaning *submit, yield to*].

... al' kad mu iz ustiju zazvonila muževna riječ, savinula se pred njim duša protivnika kao jesenski list od vjetra [... but when a manly word came out of his mouth, his rival's soul bent before him like an autumn leaf in the wind] *(Šenoa, Images and Views, 1934, 257)*. – ... treba se katkada pokloniti, savinuti i previnuti i sa slatkim smiješkom mnogu gorku progutati [...one should sometimes bow, bend, swoop and with a sweet smile swallow many bitter events] *(Kolar, By Quill and Harrow, 1938, 169)*.

(10)

sliti slijem, ovdje u značenju *spojiti.*

[**pour**, here in the meaning *unite, bring together*].

Zato on teži samo za tim, da te pojedince slije u jednu grupu [For this reason he only wants to pour these individuals into a single group] *(Batušić, From Sienna to Haarlem, 1941, 115)*.

The verification for the main meaning of the word exists in the examples (7) and (8), but we see metonymy with the entry *pallet* in the quotation from *The Purple Nights* and metaphor with the entry *grasshopper* in the quotation from *Tito Dorčić*. Since the meanings *soldier* for *pallet* and *child* for *grasshopper* are closely connected to their contexts, we have used the label *here in the meaning.* In the examples (9) and (10) there is no verification of the main meaning for words *bend* and *symphony*. In *The Dictionary of Croatian Language* (2000) the verb *bend* means to *become deformed, hunch-backed* and the noun *symphony* is explained as *a long, complicated piece of music for an orchestra*. In accordance with Benešić's rule which requires that meanings be shaped on the basis of confirmatory quotations, the main meanings of words are not given in this dictionary since we have no contextual verification. Instead, the interpretation which follows from the context of confirmatory quotation is noted after the label *here in the meaning.*

However, the label *here in the meaning* cannot cover all transferred meanings or semantic extensions which appear in this *Dictionary* and this makes the problems of our lexicographic treatment similar to those that the compilers of normative dictionaries have. According to John Ayto, lexicographers are not too sure what to do with metaphor: «it makes us nervous» (1988: 49). The development of the conceptual metaphor theory within cognitive linguistics had an impact on changing the accepted view of metaphor as 'decorative' feature of literary or rhetorical registers. The authors of the conception see metaphor as a fundamental cognitive process that shapes the way we form concepts and give them names. According to Lakoff and Johnson (1980: 2), most of our ordinary conceptual system is metaphorical in nature. The authors of *The Oxford Guide to Practical Lexicography* (2008), Atkins and Rundell believe that insights like these won't necessarily make the practical task of dictionary making any easier, but by helping us to perceive underlying systems in the language, they leave us better equipped to make sense of language data (2008: 291). In lexicography the label *fig*. is rather frequently used for metaphorical extensions. Most often it is used as a label for a given definition following a corresponding literal one from which the metaphor has been transfered (Ayto 1988: 49). Faced with a great number of different labels which he used inconsistently (see Chapter 2) and trying to find a acceptable uniform label, in Benešić's dictionary we opted for the label *fig.* In the new volumes of *The Dictionary* the label *fig*. is used for the transfer of meaning from the sphere of the concrete to the sphere of the abstract meaning. In the examples (11) and (12) the verbs come in two meanings: the first one is concrete and the second one abstract. Accordingly the opposition *concrete : abstract* does not result in a completely new meaning. Instead, the second meaning is just permeated with abstraction.

(11)

sijati sijem [**sow (sowed, sown/sowed**)].

bacati sjeme na zemlju pripremljenu za sjetvu [*to plant or spread seeds in or on the ground*].

Orali su i kopali, sijali pšenicu i ječam, raž i heljdu, repu i kukuruz, okopavali vinograd, kosili sijeno, otavu i otavicu (ako je bila jesen sretna i topla), i to uvijek od jutra do noći da su ih krsta probadala [They would plough and dig, sow wheat and barley, rye and buckwheat, turnip and corn, tend their vineyards, mow the hay and aftergrass (if the autmn was warm and happy), and they would always do it from morning to night so that their lower-backs ached] *(Krleža, Croatian god Mars, 1933, 8).*

pren. [*fig.*]

Luksemburgovac, da nas omrazi svijetu, sije širom kršćanstva klevetu, da su Horvati, da je liga zadavila Jelisavu [Luxembourgian, in order to make us disliked in the world, sows slanders across the Christendom, saying that Croats, the league had strangled Jelisava] *(Šenoa, Curse,1934, 397).* – Raspomamili se gadovi: uzeli prosipati na nas plinske bombe, sijali ih svim cestama i putovima [The bastards had gone wild, started pouring gas bombs on us, sowed them over all roads and paths] *(Goran, Days of Wrath, 1936, 144).* – U toj sobi šeta udes i sije smrt [In that room the doom walks and sows death] *(Matoš, Tired Tales, 1936, 122).*

(12)

sašiti, sašijem [**sew**, sewed, sewn].

spojiti iglom i koncem [*to join something using a needle and thread*].

O, Šmirganz morat će Ružici sašiti za vjenčanje krasnu odjeću s dugom, dugom povlakom [O, Šmirganz will have to sew a beautiful gown with a long, long train for Ružica's wedding] *(Gjalski, At night, 1913, 414).*

pren. [*fig.*]

> ... sve su to takove krpice, kojima je pisac svojemu komadu htio sašiti komičnu haljinicu [... these are all such scraps and snippets out of which the writer wanted to sew a comic dress for his play] *(Šenoa, The Theatre Reports I, 1934, 169)*.

5. Conclusion

Unfinished Benešić's *Dictionary* proved to be an atypical author-specific work which is not structured in accordance with conventional lexicographic methods. The successors to *The Dictionary* have been given the task to finish it as if Benešić himself had completed it. It is thus obvious that the work on the dictionary completion is a long-term and demanding lexicographic project with numerous subtasks including finding solutions to a great deal of dilemmas. In this paper we have presented an aspect that generally causes a lot of trouble to lexicograpy – metaphors or figurative meanings. Benešić's *Dictionary* is specific in a way that the descriptive nature of the dictionary requires that meanings be shaped by the citations which implies that it is necessary to mark specific senses that a word may have in a specific context, which are often metaphorical and metonymical, and a result of a writer's current inspiration.

The examples that we have presented in this paper show that every single entry with associated citations represents a world in itself. Moreover, the examples show that in Benešić's dictionary, owing to the corpus of citations on which it is based, one can find verification of meanings and usages that cannot be found in normative dictionaries. After several years of work on the systematizing of remaining material of Benešić's dictionary, the first of the new volumes, the thirteenth volume (letter S) will be published at the end of 2013. In this paper we have given several examples showing the treatment of figurative meanings where we try both to follow Benešić's work faithfully but also improve it in accordance with the times in which it is being finished. Therefore, completion of Benešić's dictionary will contribute not only to the history of Croatian language and literature, but it will also enrich Croatian lexicography as one of the few descriptive dictionaries based upon a corpus of citations.

References

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

Ayto, J. (1988). Fig. leaves. "Metaphor in dictionaries". In Snell-Hornby, M. (ed.). ZüriLEX '86 Proceedings. Tübingen: Francke. 49–54.

Béjoint, H. (2010). Lexicography of English. New York: Oxford University Press.

Benešić, J. (1985). Rječnik hrvatskoga književnog jezika od preporoda do I. G. Kovačića, sv. 1. Zagreb: JAZU i Globus.

Lakoff, G; Johnson, M. (1980). Metaphors We Live By. Chicago: University of Chicago Press.

Landau, S. (1984). Dictionaries: The Art and Craft of Lexicography. New York: Charles Scribner's Sons.

Nikolić-Hoyt, A. (2011). Completing an unfinished historical dictionary [on-line]. Oxford : Oxford Research Archive. http://ora.ox.ac.uk/objects/uuid:44c37fd5-78bc-4458-adb2-2864e614142d/ datastreams/ATTACHMENT01 [Access date: 15 Sept. 2013].

Sabljak, T. (1994). "Benešić i hrvatski scenski jezik". In Radmilović, M; Batušić, N. (eds.). Julije Benešić i Tito Strozzi: zbornik radova znanstvenih kolokvija Hrvatskoga narodnog kazališta u Zagrebu 1992. i 1993. Zagreb: Hrvatsko narodno kazalište. 89–94.

Sabljak, T. (1977). "Benešićev *Rječnik hrvatskoga književnog jezika*". Kronika Zavoda za književnost i teatrologiju JAZU III: 2. 9–32.

Dialect lexicography: Catalan nineteenth-century dictionaries

Maria-Pilar Perea

1. Introduction

The main purpose of LEXDIALGRAM[38], a project in progress at the Universitat de Barcelona, was to create a cultural portal to disseminate a collection of nineteenth-century Catalan dialect dictionaries and grammars and to allow users to consult them. The project addresses seven specific areas: 1) the digitization of dictionaries and grammars; 2) the creation of computer tools to enter data in lexical and grammar databases; 3) the creation of a corpora of Catalan grammars and search tools to be used for research and comparison; 4) the automatic mapping of lexical results; 5) the lexical and grammatical analysis of materials from synchronic, diachronic and comparative points of view; 6) the study of sociolinguistics and standardization; and, finally, 7) the study of the contact between different languages.

Using simple and complex search criteria, the project's development of computer technology for both the treatment and dissemination of its materials will facilitate users' point of access.

2. Catalan dialectology in the nineteenth century

Because of his work in writing and promoting the *Diccionari català-valencià-balear* (DCVB), at the beginning of the twentieth century, Antoni Maria Alcover is generally regarded as the father of Catalan dialectology. Parallel to the dialect research that was being conducted in other parts of Europe, however, several studies on dialect had already been completed in Catalonia during the nineteenth century, even though none was explicitly labelled as dialectology. And although other parts of Europe saw the birth of comparative linguistics, experimental phonetics and dialectology in this period, the Catalan-speaking territories only began to develop them as specific scientific disciplines in the twentieth-century.

[38] http://www.ub.edu/lexdialgram/

In his brief section on philological studies in Catalan before 1900, Sever Pop (1950: 340-341) observes that the first works of scholarship were championed by Manuel Mila i Fontanals, Josep Balari i Juvany, Marià Aguiló and, finally, the Roussillon archivist Julia Bernat Alart. Pop also notes that Catalan dialect studies did not adopt a scientific character until the early twentieth century. Surprisingly, even in the section which deals with dialect surveys, Pop does not refer to the works of Alcover, which were written at the very beginning of the last century and which have been incorporated in our portal because of the valuable linguistic information they contain.

In the nineteenth-century, the study of dialects in Catalonia was conducted within other linguistic disciplines and this is especially visible in works of that period on lexicography, grammar and spelling (Perea, 2003). Other text types which often addressed linguistic variation were geography dictionaries, travel books and personal letters.

Three reasons led to the emergence of dialect dictionaries, grammars and spelling-books during this period.

The first was the regional fragmentation of the Catalan-speaking territories, which originated in the loss of their sense of linguistic unity. This explains the proliferation of dictionaries and grammars in Catalan, Valencian, Majorcan, Minorcan, etc., each produced by its own writers in their own dialect quite independently of the others.[39]

The second reason was the role that Castilian and, to a lesser extent, Latin played in many of the vocabularies and grammars of this period, being used to translate Catalan.

The third reason leading to this emergence was the ideological incidence of Romanticism and the *Renaixença*,[40] the two cultural movements that prompted the scholarly collation of all varieties of folklore, popular culture and linguistic data, especially the spoken language, so that the country as a whole might return to its roots. This impetus would explain the appearance of numerous publications (the literature produced by hiking associations, for example)[41] that closely examined every aspect of traditional life in a given geographic area and provided word lists and explanations of grammatical structures whose interest to the reader would be their local nature.

[39] See an example of each typology: *Diccionari de la llengua catalana ab la correspondencia castellana i llatina* (V. Pla, Barcelona, 1839-1840), de Pere Labèrnia; el *Diccionario valenciano-castellano* (J. Ferrer de Orga, Valencia, 1851), de Josep Escrig; el *Diccionari mallorquí-castellá*, de Pere Antoni Figuera (Imp. i Llibreria d'Esteve Trias, Palma de Mallorca, 1840); y el *Diccionari menorquí-español-francês-llatí*, de Antoni Febrer i Cardona (Maria Paredes (ed.), Barcelona: IEC, 2001).

[40] The Renaixença was the late romantic revivalist movement in Catalan language and culture that took place in the early nineteenth century.

[41] See for example the content of the journals *L'Excursionista*, *Anuari de la Associació d'Excursions Catalana*, *Butlletí de l'Associació d'Excursions Catalanes*, *Memorias de l'Associació d'Excursions Científicas*, *Butlletí de l'Associació Catalana d'Excursions* or the *Bolletí del Centre Excursionista de Catalunya.*

3. The portal for nineteenth-century works on Catalan dialect lexicography and grammar

The absence of atlases and of specifically dialectological literature in the Catalan-speaking territories of the nineteenth century[42] might suggest that no description based on the analysis of linguistic variation was contemplated during this period, that such research has only begun in earnest in the early twentieth century and that, consequently, the quality of related works from any earlier period must be poor and narrow in breadth.

The hypothesis of our project, however, is that dialect studies were also written during the nineteenth century and that they are substantial enough to merit a portal. As mentionned above, these materials may have been hidden behind other linguistic disciplines, such as lexicography, grammar, and folk studies or anthropology, but they nonetheless provide relevant and representative information on the status of the Catalan language and its dialect variety in that period of history.

The LEXDIALGRAM project is designed to interactively disseminate these materials online so that they may be compared with each other and with twentieth-century lexical and grammatical data. This will help to extend the knowledge of Catalan dialect lexicography and grammar and enhance the manner in which these are studied and analysed. Finally, special importance has been given to monographic works which, in some cases, had remained unpublished until now.

3.1. List of works

Although 40 dictionaries and grammars written during the nineteenth century have already been catalogued, only 24 lexical and grammatical works, chronologically ordered, are present in the portal.

3.1.1. Dictionaries

1. Ros, C. (1739) *Breve Diccionario Valenciano-Castellano*, Valencia: Josep Garcia.
2. Ros, C. (1764) *Diccionario Valenciano-Castellano*, Valencia: Imprenta de Benito Monfort.
3. Fuster i Tarongí, J. P. (1827) *Breve vocabulario valenciano-castellano*, Valencia: Imprempta de José Gimeno, 142 p.
4. Lamarca, L. (1828) *Ensayo de un diccionario valenciano-castellano*, Valencia: J. Ferrer de Orga, 55 p.
5. Figuera, P. A. (1840) *Diccionari mallorquí-castellà*, Palma: Impremta y Llibreria de Esteva Trias, 626 p.
6. Anònim: *Diccionario Mallorquín y Castellano*. Ms. 1275 de Montserrat, mid-19th century, 50 fol. a 2 col.

[42] While in France dialectology became a university subject in the late nineteenth-century, in Catalonia and in Spain as a whole it did not acquire this status until the 1950s.

7. Escrig, J. (1851) *Diccionario valenciano-castellano*, Valencia: Imprenta de J. Ferrer de Orga.

8. Amengual, J. J. (1858-1878) *Nuevo diccionario mallorquín-castellano-latín*, 2 vol., Palma: Imprenta y librería de Juan Colomar, 748 p i 592 p.

9. Rosanes, M. (1864) Miscelánea que comprende un vocabulario valenciano-castellano, València: J. M. Ayoldi, 140 p.

10. Hospitaler, J. (1869) *Vocabulario castellano-menorquín y vice-versa*, Maó: Imprenta de Miguel Parpal, 292 p.

11. Mercé Marçà, L. (1879) *Diccionari valencià castellà*, Càlig.

12. Alcover, A. M. (1881) *Mostra de diccionari mallorquí* [Edición de M. P. Perea, Barcelona: Publicacions de l'Abadia de Montserrat, 361 p.]

13. Alcover, A. M. (1881-1886) Llista de noms mallorquins que replegava n'Antoni M. Alcover quant era estudiant.

14. Miralles, J. (s. d.) Diccionario Castellano y Mallorquín. De algunos terminos que tienen alguna analogia con la ciencia Veterinaria.

3.1.2. Grammars

1. Amengual, J. J. (1836) *Gramàtica de la llengua mallorquina*, [2a ed.], Palma: Imp. de P. J. Gelabert, (1872), 226 p.

2. Puiggarí, P. (1852) *Grammaire catalane-française, à l'usage des français…*, Perpignan: J. B, Alzine, 134 p.

3. Soler, J. (1858) *Gramática de la lengua menorquina*, Maó: Imprenta de D. J. Fábregues y Pascual, 128 p.

4. Nebot i Pérez, J. (1894) *Apuntes para una gramática valenciana popular*, València: Impremta Ripollés.

5. Pais i Melis, J. (1899) *Grammatica del dialetto moderno d'Alghero*, 1899 [edición de Pasqual Scanu, 1970].

6. Foulché-Delbosc, R. (1902) *Abrégé de grammaire catalane*, Barcelone: Imprimerie et Librairie "L'Avenç".

7. Palomba, G. (1906) *Grammatica del dialetto algherese odierno*, Sasari: Tipografia G. Montorsi.

8. Forteza, T. (1915) *Gramática de la lengua catalana*, finalizada en 1898, aunque editada en 1915, Palma: Esc. tipogr. prov., 563 p.

3.1.3. Spelling works

1. Un MAHONÉS [Joan Ramis i Ramis] (1804) *Principis de la lectura menorquina*, Mahò: Impr. de la Vda. de Fàbregues.

2. F. A. M. S. M. (1812) *Nueva ortografia de la lengua mallorquina*, Palma: Impr. de Sebastian Garcia.

3.2. Methodology

After the digitization of the documents, the dictionaries and the grammars were uploaded to the LEXDIALGRAM portal in pdf format. The next step was the creation of computer tools to enter data in lexical and grammar databases. In the future, we intend to map the lexicon of the lexical databases from the dictionaries.

The fairly simple structure of the lexical database comprises 16 fields. Note, however, that hardly any of the dictionaries in the database can be examined for information in every field.

01. Headword
02. Entry
03. Input_language
04. Definition_language
05. Grammatical category
06. Definition (1, 2...)
07. Examples
08. Meanings (1, 2....)
09. Idioms
10. Cross-references
12. Castilian translation
13. French translation
14. Italian translation
15. Latin translation
16. Dialect variety

Most of the dictionaries are bilingual and generally do not include definitions, although they do provide lexical equivalents in another language. Some of them also offer examples, idioms, geographic references about the use of a certain word and even that word's documentary source. In the database, the data in the field "Headword" was entered last because this word needed to unify formal differences that may occur in the different dictionaries. In general, the headword is usually documented in dictionaries. With regard to spelling, our rule of thumb is to adopt the form of the word as it appears in Antoni Maria Alcover and Francesc de B. Moll's *Diccionari català-valencià-balear* (http://dcvb.iecat.net/). However, if this word refers to a more general entry, the more general entry is adopted. This, for example, is the case of the entry *anclusa* ('anvil'): although it appears in the DCVB as a spelling variant, the database refers the user to *enclusa* ("ANCLUSA *f.,* var. ort.: V. enclusa") and this is the headword that has been assigned.

Dictionary	Headword	Entry	Idiom	Translation
Lamarca	Enclusa	Anclusa		Yunque.
Lamarca	Enclusa	anclusa	Una en la anclusa y atra en lo martell	Una en el clavo y ciento en la herradura.
Fuster	enclusa	enclusa		yunque.
Rosanes	Enclusa	Anclusa (entre ferrers)		Yunque ó bigornia.

When the word is not documented, we choose the use of a coherent spelling.

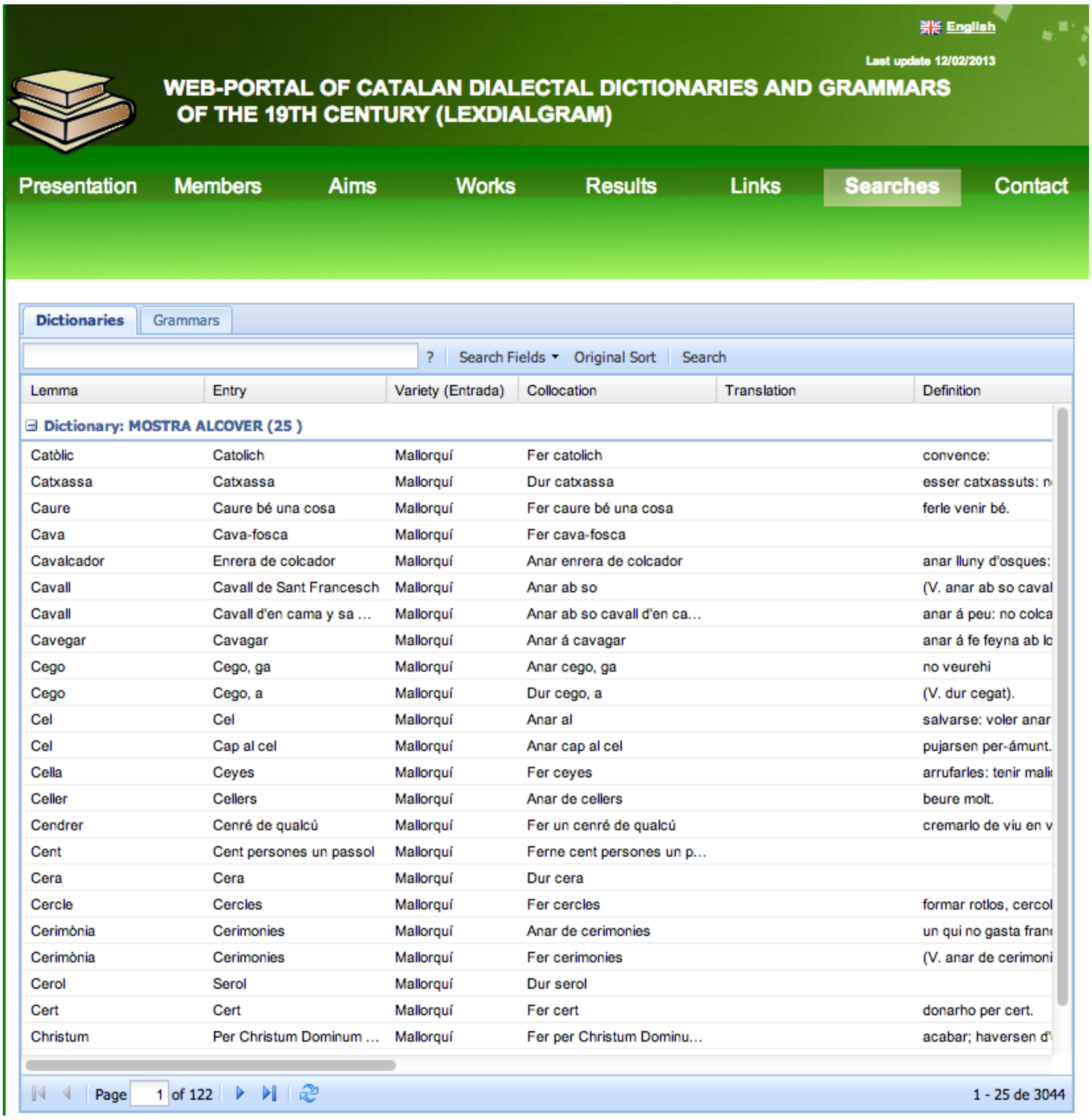
Dictionary	Headword	Entry	Idiom	Definition / Translation
Mostra	onomatopoeia Crec-crec	Crech-crech	Anar crech-crech	anar mortal. to die off
Fuster	opinionàtic	opinionatich		hombre ligero de seso. 'stupid'
Fuster	opinionàtic	opinionatich		hombre de varias opiniones. 'man of several opinions'

The headword can also include Spanish forms:

Dictionary	Headword	Entry	Observations	Idioms	Definition / Translation
Lamarca	Quadrillo 'checkered'	Cuadrillo	(en las medias)		Cuadrado.
Mostra	Reparo 'objection'	Reparo ab una cosa		Dur reparo ab una cosa	repar en ferho.

The structure of the grammar database is more complex. Initially, the information is entered according to the structure of the work so that the content of the different grammars can be more easily compared, in varying degrees of detail. Particular importance has been given to the examples, since in most cases the work was written in Castilian (or in French or Italian) and only the examples appear in Catalan. In due course we plan to use quantitative lexical analysis to compile a list of words from the grammars which will be connected with the corresponding concordance. Finally, we have tried to enter the materials semi-automatically, but this has always required manual monitoring. Our last step will be to categorize and classify the examples.

The online publication of the digitized materials and the corresponding databases in the portal not only helped to a better dissemination of the works but also favoured different studies derived from consulting the data (Figure 1).



Dictionary: MOSTRA ALCOVER (25)

Lemma	Entry	Variety (Entrada)	Collocation	Translation	Definition
Catòlic	Catolich	Mallorquí	Fer catolich		convence:
Catxassa	Catxassa	Mallorquí	Dur catxassa		esser catxassuts: n
Caure	Caure bé una cosa	Mallorquí	Fer caure bé una cosa		ferle venir bé.
Cava	Cava-fosca	Mallorquí	Fer cava-fosca		
Cavalcador	Enrera de colcador	Mallorquí	Anar enrera de colcador		anar lluny d'osques:
Cavall	Cavall de Sant Francesch	Mallorquí	Anar ab so		(V. anar ab so caval
Cavall	Cavall d'en cama y sa ...	Mallorquí	Anar ab so cavall d'en ca...		anar á peu: no colca
Cavegar	Cavagar	Mallorquí	Anar á cavagar		anar á fe feyna ab lo
Cego	Cego, ga	Mallorquí	Anar cego, ga		no veurehi
Cego	Cego, a	Mallorquí	Dur cego, a		(V. dur cegat).
Cel	Cel	Mallorquí	Anar al		salvarse: voler anar
Cel	Cap al cel	Mallorquí	Anar cap al cel		pujarsen per-ámunt.
Cella	Ceyes	Mallorquí	Fer ceyes		arrufarles: tenir mali
Celler	Cellers	Mallorquí	Anar de cellers		beure molt.
Cendrer	Cenré de qualcú	Mallorquí	Fer un cenré de qualcú		cremarlo de viu en v
Cent	Cent persones un passol	Mallorquí	Ferne cent persones un p...		
Cera	Cera	Mallorquí	Dur cera		
Cercle	Cercles	Mallorquí	Fer cercles		formar rotlos, cercol
Cerimònia	Cerimonies	Mallorquí	Anar de cerimonies		un qui no gasta fran
Cerimònia	Cerimonies	Mallorquí	Fer cerimonies		(V. anar de cerimoni
Cerol	Serol	Mallorquí	Dur serol		
Cert	Cert	Mallorquí	Fer cert		donarho per cert.
Christum	Per Christum Dominum ...	Mallorquí	Fer per Christum Dominu...		acabar; haversen d'

Page 1 of 122

1 - 25 de 3044

Figure 1. Results obtained from a particular search

The materials in the database can be used for lexical and grammatical dialect analysis either from a synchronic or from a diachronic and comparative point of view. Since they contain samples of linguistic variation from different periods of history, they can also be used for the following types of research:

1) the study of sociolinguistics through language and language dialects;

2) the study of grammar and of the creation of standard language forms (note that a standard Catalan form was not created until 1917);

3) the study of discourse analysis (using the prefaces to the grammars);

4) the study of the contact between different languages and comparative linguistics, given that most of the dictionaries are bilingual (Catalan dialect-Castilian; Catalan dialect-Latin; Catalan dialect-French; Catalan dialect-Italian) and that the grammars were generally written in Castilian (or, when published in Roussillon or Alghero, in French and Italian respectively).

4. Dialects in nineteenth-century Catalan dictionaries

A good example of how the database covers lexical material is the comparison it offers between a sample of dialect words in the Valencian dictionaries *Lamarca* (L) (1828), *Fuster* (F) (1827) and *Rosanes* (R) (1864). Note, however, that the three dictionaries are not equal in either quality or breadth of reference.

The words are arranged alphabetically according to the headword:

Acatxar (L, F) ‘to crouch’
Adobar (L, R) ‘to fix up’; ’to dress’; ’to tan’
Agarrar (L, F, R) ‘to grab’; ‘to cling’
Agranada (L, R) ‘sweeping action’
Agró (L, F, R) ‘heron’
Agrunsar (L, R) ‘to swing’
Agullat (L, R) ‘needle fish’
Aguller (L, R) ‘strand of thread’
Aixovar (L, F, R) ‘trousseau’
Aladroc (L, R) ‘anchovy’
Albelló (L, F, R) ‘sewer’
Albergínia (L, R) ‘eggplant’
Albors (L, R) ‘arbutus’
Albudeca (L, R) ‘a kind of watermelon’
Aliacrà (L, R) ‘jaundice’
Anouer (L, R) ‘walnut’
Ansa (F, R) ‘handle’
Arruixadora (L, R) ‘watering can’
Bac (L, R) ‘blow’; ‘fall’
Bacora (L, R) ‘fig’
Bajoca (L, R) ‘bean’
Baldelló (L, R) ‘latch’
Baldraga (L, F) ‘free’; ‘no charge’
Barrumballa (L, R) ‘woodchip’
Bavosall (L, F) ‘bib’
Becada (L, F) ‘a sort of bird’; ‘to get some sleep’
Bescollada (L, R) ‘hit in the neck’
Bolcada (L, R) ‘layette’
Brecar (L, F) ‘to waste away’
Bresquilla (L, R) ‘peach’
Buanya (L, R) ‘grain purulent’
Canonet (L, R) ‘needle case’
Endívia (L, R) ‘chicory’
Esguit (L, R) ‘splash’
Estrena (L, R) ‘Christmas box’
Fardatxo (L, R) ‘lizard’
Farfallós (L, F, R) ‘stammerer’
Lledoner (L, F, R) ‘hackberry’
Llepolia (L, F, R) ‘candy’
Llibrell (L, F, R) ‘basin’
Manifasser (L, R) ‘meddler’
Milotxa (L, R) ‘kite’
Mostela (L, R) ‘weasel’
Paltrot (L, R) ‘kind of sausage’
Panderola (L, R) ‘cockroach’
Pedrapiquer (L, R) ‘stonecutter’
Pigota (L, R) ‘smallpox’
Pilma (L, F, R) ‘poultice’
Tarquim (L, R) ‘mud’
Torcaboca (L, R) ‘serviette’

This sample of 50 words deserves a few comments.

1. Most of the coincidences recorded occur between *Rosanes* and *Lamarca*, although a number of words appear in all three dictionaries. In general, these coincidences are semantic.

2. *Rosanes* is the most complete work, because in addition to its length[43] it contains more examples than the other dictionaries and also groups words according to their semantic field (in the database, see the field "Observations"), which helps to classify the lexicon. Several errors occur because the same word is repeated in different groups.

3. All the words are currently used in the Valencian area and also in some varieties of Northwestern Catalan.

4. Many of the Valencian dialect words have become preferred standard forms (e.g., *aixovar* for 'dowry', *llepolia* for 'candy' and *estrena* for 'Christmas box').

5. Of the 50 words, six in the *Diccionari català-valencià-balear* are not included in the standard Catalan dictionary (DIEC2). These are *albors* for the plant genus *arbutus*, *anouer* for 'walnut', *barrumballa* for 'woodchip', *brecar* for 'waste away', *paltrot* to describe a kind of sausage, and *pilma* for 'poultice'. Once again, this raises the question of whether dialect forms should be included in prescriptive dictionaries.

5. Working with non-existent dictionaries

One of the innovations of the LEXDIALGRAM project is that it includes dictionaries or vocabularies written over the second half of the 19th century which have remained unpublished until now. One of them belongs to the Valencian area and the remaining four are assigned to the Majorcan dialect: The *Diccionari valencià castellà* (1879) by Lluís Mercé Marçà; the *Diccionario Mallorquín y Castellano*, anonymous, mid-19th century, and the works by Antoni M. Alcover: *Mostra de diccionari mallorquí* (1881), *Llista de noms mallorquins que replegava n'Antoni M. Alcover quant era estudiant* (1881-1886), and the *Diccionario Castellano y Mallorquín. De algunos terminos que tienen alguna analogia con la ciencia Veterinaria* (n.d.)

The *Diccionari valencià castellà* by Lluís Mercè Marçà was drafted in Càlig (Valencian area) in 1879. The author was a farmer who was interested in the lexicon of its population and collected 841 words, many of them related to agriculture. This work includes words which, according to the dictionary of reference (the DCVB),[44] are used in other dialectal areas. "Pastecum = portapaz" is an example. The DCVB registers that *pastècum* is used only in Minorca and etymologically comes from the Latin formula *pax tecum*.

[43] This dictionary has the largest number of entries but not the broadest lexical variety (it often includes many inflected forms of verbs, for example).

[44] The *Diccionari català-valencià-balear* is descriptive and for this reason it has become our reference work. The *Diccionari de la llengua catalana* by the Institut d'Estudis Catalans is prescriptive and includes only the standard lexicon.

Different phonetic phenomena, such as metathesis, can be found: "barallugá: menearse, moverse" (*ballarugar*) 'to wag', and samples of words that do not appear in the DCVB: "tora dels cabells: horquilla" 'hair pin'; 'trápula: torbellino" 'whirlwind'.

For reasons of space, I will focus on only two Majorcan works: the unpublished dictionary and the *Diccionario Castellano y Mallorquí. De algunos terminos que tienen alguna analogia con la ciencia veterinaria*, initially attributed to Alcover, but recent research seems to prove that the author was a young priest, Josep Miralles, Alcover's friend, who become bishop of Lerida (1914-1925), of Barcelona (1926-1930) and of Majorca (1930-1947) (see Perea, to be printed a). In fact, it is a selected copy of words more or less related to the title of the work of the *Diccionari mallorquí-castellà* by Pere Antoni Figuera, published in 1840. Miralles must have been in his twenties when he wrote this work, although the motivations of doing this dictionary are unknown. His word selection and his lexicon additions make this dictionary a very interesting work from a sociolinguistic point of view.

The *Diccionario Mallorquín y Castellano* includes 2325 entries. It mainly consists of lists of words with their equivalent Spanish translations and a few definitions. This work provides numerous examples of dialectal pronunciation: "amatlo tenrre: almendruco" (Stan. *ametlló tendre*) 'sweet almond', "aubello: sumidero" (Stan. *albelló*) 'sewer'; "aujup: algibe" (Stand. *aljub*) 'cistern'. It also provides evidence that allow us to advance the date of documentation of some words. Thus, it distinguishes between "arne de roba: polilla" 'clothes moth' and "arne del cap: caspa" 'dandruff'. This last word is documented, in the DCVB in 1913, in *Aygo-forts* by the poet Gabriel Maura.

Some phenomena explained in the *Gramática de la lengua catalana*, written by the Spanish poet and scholar Tomas Forteza (1838-1898) in 1881 but not published until 1915, can be found in the *Diccionario Mallorquín y Castellano*. This grammar was based on the new tendencies of European Romance linguistics, particularly on the French version of Friedrich Diez's *Grammatik der romanischen Sprachen* (1836-1844). When referring to the Latin sounds (p. 18, b), Forteza mentions the phenomenon that has taken place in the case of LI (L + glide [j]); the result, "ll", has become [j] in some areas of Catalonia and in the Balearic Islands. Several examples can be found in the dictionary: "aclaridor de cabeis: Escarpidor ó partidor de cabellos" (Stand. *cabells*) 'hear'; "amagatay: escondite ó escondrijo" (Stand. *amagatall*) 'hiding place'.

Forteza notes certain assimilation processes characteristic of the Majorcan dialect: with relation to the consonant groups RL: *parlar* (*pal·lar*) 'to speak' (p. 66, e), which can be seen in "metlera (ave): merla o mirla" (Stand. *merla*) 'blackbird'.

Finally, the *Diccionario Castellano y Mallorquí. De algunos terminos que tienen alguna analogia con la ciencia veterinari*a has 1500 entries. Despite its title, not all words are related to veterinary science. The structure of an entry consists of the lemma and the definition and its equivalent in Spanish, but often the author forgets to write it down or writes "Id.", although the spelling of the Majorcan word is different from the Spanish.

There are also curious definitions ("Curálotodo, Remey qui se aplica para moltissims máls. Curalotodo" 'cure-all'). Besides certain concepts, which in principle are related to diseases and healing of animals (see "Alevosa. serta malaltia de caballs y Bous. Alevosa ò Ranula" 'tumour'), many medical and anatomical terms are collected: “Contátge. Malaltia que se aprent. Contagio” ‘contagion’; ”Anassarca. Especie de infló. Anasarca” ‘edema’; ”Apostéma. Inflo esterior de matéria. Apostema” ‘apostema’; “Diagnòstic. Lo pertañent à los signos y sintomas de las malaltias. Diagnostico” ‘diagnosis’; ”Diafragma. Musculo nervios entre es pit y es ventre. Diafragma” ‘diaphragm’; ”Destemplársa. Alteració en el pols” ‘irregular pulse’; “Desmenjament. Ynapetencia, desgana, de menjar. Desganarse” ‘not wanting to eat’.

There are many words for trees and plants (“Apit. Apio” ‘celery’; “Datil. El fruyt de fasser. Datil” ‘date’; “Ciuró, fruyt. Garba[n]zo” ‘chickpea’; “Coco. Abre y fruyt. Coco” ‘coconut’; “Cugula Yerba. Avena silvestre ò rustica” ‘darnel’), for animals (“Armadillo, animal cuadruped de Africa” ‘armadillo’; “Armiño, Petit animal blánc y so cab de sa coua negre. Armi[ño.]” ‘ermine’; “Arna, Animalet qui roega sa roba. Polilla” ‘moth’; “Aronélla. Aucéll conegut. Golondrina, andolina” ‘swallow’; “Assor. Ave de repiña. Azor” ‘goshawk’) and other words belonging to different semantic fields (“Diamánt. Pédra preciosa. Diamante” ‘diamond’; “Democracia. Gobe[r]n popular” ‘democracy’; “Dexeble. El qui estudia ab mestre ò mestra. Dicípulo” ‘disciple’), which gives the work a more encyclopedic character.

Unfortunately, this dictionary is in very poor condition due to humidity, many pages are broken, and the writing erased. Still it provides some interesting information from a dialectal viewpoint.

All of these dictionaries offer an excellent report of the state of the Catalan dialectal varieties in the period when they were written. The comparison with similar works written in the nineteenth century, which belong to the same dialectal area, and the contrast with modern solutions show that the information is absolutely correct (see Colón 2013 for an overview of Catalan dialect lexicography). What is more, they allow the phonetic and semantic lexicon evolution across a century to be studied.

The question of the authority of these dictionaries raises the question of the balance between the standard language and the Catalan dialects (see Perea to be printed b). The Catalan language was standardised very late if compared to other languages. The standardisation started in 1917 with the publication of the *Diccionari ortogràfic* by Pompeu Fabra. He basically selected the words from the Barcelona area and for that reason dialects were not well represented. This situation, together with the “prestige” of Barcelona’s variety, and the publication in 1931 of the *Diccionari general de la llengua catalana*, also by Fabra, left the Catalan dialects which were far away from the “prestigious” area (that is to say, North-Western Catalan, Valencian, Balearic, Roussillon and Algheres) in a bad situation. The pressure of standardization relentlessly reduces some idiosyncratic traits of the dialects. Against this tendency there must be not only a proper linguistic policy to secure the use of these varieties but also the speakers must show a personal attitude to preserve and construct their own identity.

6. Conclusions

The recovery and digitization of the works described above have provided us with important data on nineteenth-century Catalan dialects. In some cases, this data was previously unknown or only partially examined. Its recovery can help us to compare and examine information about the past with existing data.

The exploitation of lexical data will also help us learn more about the status of the Catalan language and its linguistic variation in a century when dialect studies as such were largely ignored and which even today has not been sufficiently studied.

Finally, within the lexicon, the study of these materials can allow us to observe how linguistic information has been transferred from one dictionary to another and determine how and to what extent authors have incorporated new data. The expected last aim would be that the knowledge of the traditional and historical lexicon used in different Catalan dialects contributes to enrich the content of the present standard dictionary (DIEC2) instead of being two sides of the same coin.

References

Alcover, A.M.; Moll, F. de B. (1930-1962). *Diccionari català-valencià-balear*. Palma de Mallorca: Moll [http://dcvb.iecat.net/]

Colón, G.; Soberanas, A.J. (1985). *Panorama de la lexicografia catalana*. Barcelona: Enciclopèdia Catalana.

Colón, G. (2013) "An approach to Catalan dialect lexicography". *Dialectologia,* special issue, IV, 49-75 <http://www.publicacions.ub.edu/revistes/dialectologiaSP2013/> [access date 22 december. 2013].

DIEC2 (2007). = Institut d'Estudis Catalans: *Diccionari de la Llengua Catalana,* [2nd edition] [http://dlc.iec.cat/].

Meyer-Lübke, W. (1890-1902). *Grammatik der romanischen Sprachen*. Leipzig: Reisland [French edition: *Grammaire des langues romanes*. Paris H. Welter. (1890-1906)].

Perea, M.-P. (2003). "Els dialectes catalans i el desenvolupament de la dialectologia durant el segle XIX". *Actes del Dotzè Col·loqui Internacional de Llengua i Literatura Catalanes.* Universitat de París IV - Sorbonne 4-10 setembre 2000. Barcelona: AILCC, Publicacions de l'Abadia de Montserrat, 413-433.

Perea, M.-P. (2008). "Tomàs Forteza: capdavanter dels estudis lingüístics i gramaticals, in Tomás Forteza i Cortés, *Gramática de la lengua catalana*, vol. I. Barcelona: Publicacions de l'Abadia de Montserrat, 5-46.

Perea, M.-P. (to be printed a). "Un diccionari mallorquí de veterinària de final del segle XIX a la cerca d'autor". *Randa,* 71.

Perea, M.-P. (to be printed b). "Fabra i Alcover: hi ha alguna cosa més a dir-ne quant a l'ús del lèxic?". *IV Col·loqui Internacional 'La lingüística de Pompeu Fabra,* Tarragona 2013.

Pop, S. (1950) *La dialectologie: aperçu historique et méthodes d'enquêtes linguistiques*. Louvain: J. Duculot.

II

Dictionaries for special purposes

Minimal definitions and lexical agreement: project of a dynamic dictionary

Elena de Miguel

1. The project *Diccionario electrónico multilingüe de verbos de movimiento*: general outline

In this paper, I will present the project *Diccionario electrónico multinlingüe de verbos de movimiento* (Multilingual Electronic Dictionary of Motion Verbs, DICEMTO in its Spanish acronym), which has been developed by the group UPSTAIRS at Universidad Autónoma de Madrid since 2009[45]. The main goal of this dictionary is to register systematically and uniformly the different meanings expressed by motion verbs in several languages, as a result of their combination in different contexts.

This proposal belongs to a line of research which aims to identify the sub-lexical information contained in the inner structure of words. This information largely determines the contexts wherein a word can show up and its possible meanings (literal and metaphorical).

The starting hypothesis is based on the Generative Lexicon (henceforth GL; Pustejovsky 1995) theoretical model, which assumes the existence of a repertoire of sub-lexical features that determine the properties of words and license their meaning extension and emptying in different contexts. The context is also determining, insofar as word features form an underspecified entry, which will be fully specified according to the features of the words with which it combines in the syntax.

Therefore, this dictionary is intended to be *dynamic*, since it aims to include, in each entry, the (dynamic) process of meaning construction. In order to do so, each verb entry provides minimal definitions contained in a set of features that will allow deducing each particular context-dependent meaning, on the basis of specific lexical agreement operations. The proposed processes are (generative) meaning-makers,

[45] UPSTAIRS is the acronym of "Unidad de Estudio de la Palabra: Estructura Interna y Relaciones Sintácticas" ('Word Study Group: Inner Structure and Syntactic Relations'), research group HUM F-047 at Universidad Autónoma de Madrid of which I am the head researcher.

and they are responsible for the licensed word combinations (both literal and metaphorical) and the unacceptable ones. Thus, there are three basic assumptions which the planned dictionary lies on: (a) the *underspecification* assumption, or the lack of specification of lexical entries, which enables them to intervene in different syntactic structures and, consequently, in different operations of semantic composition (Pustejovsky 1995); (b) the *decomposition* assumption: it establishes that the word meaning is decomposed into different sub-lexical features (eventive, qualia, etc.)[46] — their encoding constitutes its inner or sub-lexical structure, which is not transparent, but it is visible thanks to its syntactic behavior; and (c) the meaning *compositionality* assumption: it presupposes the existence of a set of principles and (sub-lexical feature agreement) mechanisms, which are able to generate many meanings from the minimal definition of a word, when the word is combined with others in the syntax.

The words that have agreeing features give rise to interpretable meaningful constructions, as in the case of *cuadro* 'painting' and *pintar* 'paint' in (1a). The words that have non-agreeing features, as in the case of *cuadro* and *comer* 'eat', give rise to unacceptable expressions, as in (1b):

(1) a. El cuadro fue pintado por Velázquez.
'The picture was painted by Velázquez'
b. *El cuadro fue comido por Velázquez.
'The painting was eaten by Velázquez'
c. El sol fue pintado por el niño.
'The sun was painted by the kid'
d. El niño ha comido un plato.
'The kid has eaten a dish'

Lexical feature agreement mechanisms[47] license or rule out some combinations, and they also rescue other combinations which were impossible in principle: for example, *sol* 'sun' and *pintar* 'paint' in (1c) and *plato* 'dish' and *comer* 'eat' in (1d) form interpretable expressions because a mechanism allowing *sol* to be recategorized as an [ARTISTIC CREATION] and *plato* as a [FOOD CONTAINER] operates in both cases.

In addition, the existence of sub-lexical features and feature agreement mechanisms allows explaining why (2a) is interpreted as 'The picture which Velázquez painted'[48] and (2b) as 'Velázquez painted many paintings', since both *pintar* and *cuadro* fully agree; why (2c) is an unacceptable expression (since it is redundant); why (2d) is an acceptable expression but (2e) is not: because *cuadro*, unlike *plato*, lacks the feature [CONTAINER], which allows predicating, in the absence

[46] I will return to these concepts and terms later, in section 3.3.2.

[47] I propose this term drawing on Bosque (2004).

[48] There are other interpretations: since *cuadro* refers to an object involving a representation, it includes in its definition, in addition to the feature [ARTISTIC CREATION], the features [ICONIC OBJECT] and [PHYSICAL OBJECT]. This is the reason why (2a) can mean, in addition to 'The painting which Velázquez painted', either 'The painting in which Velázquez turns up' or 'The painting which belongs to Velázquez'. This ambiguity used to be explained in terms of thematic roles.

of complements, that the container's capacity has been reached or saturated; and why (2f) is an acceptable expression, while (2g) only has allows for an hyperbolic interpretation: even though both *fuente* 'platter' and *plato* 'dish' bear the feature [CONTAINER], their underspecified definitions include different functions: [TO CONTAIN FOOD] for *plato* and [TO SERVE OR DISTRIBUTE FOOD] for *fuente*:

(2) a. El cuadro de Velázquez.
'Velázquez's painting'
b. Velázquez pintó mucho.
'Velázquez painted {much/a lot}'
c. *El cuadro fue pintado.
'The painting was painted'
d. El plato está lleno.
'The dish is full'
e. *El cuadro está lleno.
'The painting is full'
f. El niño se comió un plato.
'The kid ate (up) a dish'
g. El niño se comió una fuente.
'The kid ate (up) a platter'

The simplification of lexical entries, favored by the importance ascribed to lexical agreement mechanisms in the construction of new meanings, constitutes an interesting hypothesis for the lexical acquisition and processing studies, and also for inclusive lexicographic proposals. It is one of the tenets of the DICEMTO project.[49] DICEMTO focuses on the study of motion verbs, a very productive semantic field particularly in a contrastive approach. Due to their prototypical ability to empty or extend their basic meaning, motion verbs usually take part in periphrases and idioms. Their definition usually occupies many lines in traditional dictionaries: e.g. *andar* 'walk' has nineteen meanings according to the DRAE. Our project claims that this multiplicity of meanings is not desirable, both from a theoretical and an applied perspective, because relevant theoretical generalizations are lost, and because an explanation for lexical acquisition becomes even more difficult to achieve.[50] This multiplicity of meanings is avoidable, or can at least be reduced, if the dictionary provides entries composed of minimal content features with the ability to produce different meanings when lexical feature agreement processes operate, according to the hypothesis previously outlined.[51]

[49] In line with the research on lexical structure I have developed for the last fifteen years, in the framework of Pustejovsky's GL; cf. for instance De Miguel (2000, 2004b, 2009, 2011, 2012a, 2012b, 2013) and the references quoted in these works.

[50] On the other hand, lengthy definitions are harder to look up for a standard user and they do not guarantee that the searched meaning will be found, since the contextual and semantic potential of the verbs is oftentimes endless.

[51] This proposal is put forward in De Miguel (2009, 2012a, 2013). For motion verbs, cf. De Miguel (2004a, 2012b). Provisional results of the dictionary project are exposed in detail in Batiukova & De Miguel (2013) and De Miguel & Batiukova (2013). De Miguel (2014) presents the dictionary within a general framework, by comparing it to other lexicological and lexicographical projects.

Thus, the goal of this project is to establish the minimal definitions of verbs and to account for the relations that they establish with their arguments and adjuncts in different contexts, which trigger the materialization of different parts of the underspecified definition. This underspecified definition encompasses the potential content which can either be activated or remain hidden.[52] To reach this point, the research has proceeded from syntax to lexicon, which is typically illustrated with how the controversial interaction of *lexical* and *syntactic aspect* has been dealt with. I address this question in the following section.

2. The compositional nature of lexical meaning vs. the lexical nature of compositional aspect

2.1. Aspect in the late 20th century linguistic studies

Aspect was considered crucial in lexicon-syntax interface studies in the mid and late 1980s. Specifically, there was a point where theoretical lexicalist studies — those which defend that the lexicon determines the syntax— started arguing that semantic functions or thematic roles (θ-roles) assigned by predicates to their arguments are not syntactically relevant lexical-semantic primitives; rather, they are the outcome of the relation between the arguments and the structure of the event denoted by the predicate. From this point of view, the θ-roles derive from the relation of the arguments with aspectual information, which is considered as basic. At that point, aspect, which was a relatively neglected notion in the western linguistic tradition, became a necessary category to explain many different grammatical facts (cf. De Miguel 1992).

The grammatical category *aspect*, which encodes lexical-semantic information related to event organization, has many different manifestations in different languages, and even in the same language: it can be realized lexically (as in the opposition between *nacer* 'be born', punctual, and *sesear* 'pronounce the /θ/ as the /s/', frequentative); morphologically (*renacer* 'be reborn', iterative, vs. *nacer* 'be born', punctual); and syntactically (*pintar (cuadros)* 'paint paintings', atelic, vs. *pintar un cuadro* 'paint a painting', telic) (cf. De Miguel 1999). In some of the first theoretical contributions, it was proposed that the aspectual information is conveyed by an event argument (Davidson 1967, and, subsequently, Higginbotham 1987 and De Miguel 1992). In other works, this information is included in a different level of the lexical-semantic representation or in the lexical-conceptual structure of event-denoting words (Tenny 1998; Jackendoff 1987; Grimshaw 1990). Later, neo-structuralist or constructionist models within the Generative Grammar framework

[52] There are other proposals which aim to establish minimal word definitions, like the *Proyecto de definiciones mínimas* (Bosque & Mairal 2013a, b), or studying the contribution of the relationships between words to meaning construction, like the *Diccionario combinatorio dinámico* (Almela, Cantos & Sánchez 2011, 2013). In line with the theoretical model it is based on, DICEMTO assumes that both the lexical information and the context contribution are determining factors: this is the reason why the GL has sometimes been considered a lexicalist model and some other times a constructionist one (cf. De Miguel 2009 in this respect).

— according to which construction determines meaning— claimed that the aspectual information is encoded in one or more functional heads, whose features determine the argument realization in different structural positions and the resulting interpretation of the event (Ramchand 2008, among others).

Anyway, either as a lexical-semantic primitive that determines the syntax or as a structural notion that affects the interpretation, aspect is rescued by formal grammar studies. It plays a crucial role in the definition of predicative words and it changes the account of many linguistic processes: nominalization, attribution, secondary predication, and the distinction between unaccusative and unergative verbs, among others (cf. De Miguel 1999).

2.2. The compositional nature of lexical aspect

One of the phenomena usually explained in aspectual terms is the periphrastic passive in Spanish. This construction seems to require perfective verbs, such as *robar* 'steal' in (3a), and rejects non-perfective verbs, such as *tener* 'have' in (3b):

(3) a. El cuadro de *El grito* fue robado (por un fanático).
'*The Scream* painting was stolen by a fanatic'
b. *El cuadro de *Inocencio X* fue tenido en 2013 (por el Prado).
'The *Inocencio X* painting was had in 2013 (by the Prado Museum)'
(< El Prado tuvo el cuadro de *Inocencio X* en 2013;
'The Prado Museum had the *Inocencio X* painting in 2013')

Nevertheless, this restriction is not enough to account for contrasts like (4a-b), where the same verb behaves differently in the periphrastic passive depending on its complement: if the complement delimits the verb (as in *un poema de Machado* 'a Machado's poem'), the passive is possible; if the complement does not delimit it (as in the case of the uncountable noun *poesía* 'poetry'), the passive is unacceptable, which forces us to revise the requirement: it is not the verb but rather the predicate that has to be perfective.

(4) a. Ha sido leído un poema de Machado en la clausura del curso.
'A Machado's poem has been read in the closing session'
b. *Ha sido leída poesía en la clausura del curso.
'Poetry has been read in the closing session'
(< Alguien ha leído poesía en la clausura del curso;
'Someone has read poetry in the closing session')

Contrast (4a-b) leads us to a classic problem pointed out in aspect studies, the so-called *lexical aspect paradox*: what we call lexical seems rather syntactic. Indeed, linguists often notice that it is difficult to classify verbs according to their lexical aspect, since it can change depending on the complement, the subject, certain adverbial modifiers, and other factors (cf. De Miguel 1999). This phenomenon is often attributed to the *compositional nature of lexical aspect*, which implies a paradox (and a contradiction in terms). Consider the following examples:

(5) a. Ana (se) comió un plato de tallarines en la cafetería de la universidad.
'Ana ate (up) a dish of noodles in the university cafeteria'
b. Ana (*se) comió en la cafetería de la universidad.
lit.: 'Ana ate (up) in the university cafeteria'
c. Ana (*se) comió tallarines.
'Ana ate (up) noodles'
d. Ana (*se) comió pasta.
'Ana ate (up) pasta'

A verb like *comer* 'eat' can denote an accomplishment (in Vendler's 1967 terms) if combined with a DP as complement, like in (5a). By contrast, if the same verb appears without complement, as in (5b), if the complement is in plural and without determiner, as in (5c), or if it is a mass noun, as in (5d), the event is an activity. Consequently, only (5a) accepts the aspectual *se*, which is a clitic compatible with delimited events exclusively.[53]

On the other hand, examples in (6) illustrate how a punctual verb like *salir* 'leave, come out', which denotes an achievement in (6a) and which is incompatible with *aún* 'still', behaves as a durative event if the subject is collective (6b), mass (6c), or plural (6d), which are cases where it can be combined with *aún*:

(6) a. Juan está saliendo (*aún) de la fiesta.
'Juan is (still) leaving the party'
b. El ejército está saliendo (aún) de la ciudad.
'The army is (still) leaving the town'
c. Está saliendo (aún) agua de la habitación.
'The water is (still) coming out of the room'
d. Están saliendo (aún) invitados de la fiesta.
'Some guests are (still) leaving the party'
e. Juan está saliendo (aún) del país.
'Juan is (still) leaving the country'
f. Juan está saliendo (aún) de la depresión.
'Juan is (still) coming out of depression'

However, *salir* combined with an individual subject (*Juan* in 6e, f) also accepts *aún* if the adjunct identifying the place which is left has a spatial or a temporal extension which confers duration to the event, such as *el país* 'the country' or *la depresión* 'the depression'.

In (7) it is once again illustrated that adjuncts can modify the event properties: in (7a), *a una edad temprana* 'at a young age' focuses on the moment when the change of location or state happens, whereas *por la ventana* 'through the window' in (7b) encodes the point through which the change of location is brought about. The consequence is that (7a) accepts the delimiting clitic *se*, whereas (7b) does not:

(7) a. Doña Inés (se) salió del convento a una edad temprana.
lit.: Doña Inés left the convent at a young age.
'Doña Inés {came out of the convent/stopped being a nun} at a young age'

[53] For properties and nature of this culminative clitic, cf. De Miguel & Fernández Lagunilla (2000).

b. Don Juan (*se) salió del convento por la ventana.
'Don Juan left the convent through the window'

Motion verbs examined in DICEMTO typically show the impact of arguments and adjuncts on the construction of the event:

(8) a. Juan va del valle a la montaña.
'Juan goes from the valley to the mountain'
b. La carretera va del valle a la montaña (≈ 'Hay carretera del valle a la montaña')
'The road goes from the valley to the mountain' ('There is a road from the valley to the mountain')
c. El corredor llega a la meta. (≈ 'El corredor ha llegado a la meta')
'The runner reaches the finish line' (≈ 'The runner has reached the finish line')
d. La carretera llega hasta la montaña. (≈ 'La carretera está en la montaña')
'The road reaches the mountain' (≈ 'The road is in the mountain')
e. El niño llega al botón del ascensor. ('El niño {ha llegado al botón del ascensor/es así de alto}')
'The kid reaches the lift button' (≈ 'The kid {has reached the lift button/is this tall}')
f. La policía rodea el edificio. (≈ 'La policía está {rodeando/alrededor} del edificio')
Lit. 'The police surrounds the building' (≈ 'The police is {surrounding /around} the building')
g. La valla rodea el prado. (≈ 'La valla está alrededor del prado')
'The fence surrounds the field' (≈ 'The fence is around the field')

The subject *Juan* in (8a) participates in a dynamic event; by contrast, the subject *la carretera* 'the road' in (8b) cannot, which triggers the static reading of the verb *ir* 'to go'. The same holds for the verb *llegar* 'to arrive', which denotes a dynamic event when the subject is *el corredor* 'the runner', in (8c), and a static event in (8d): the subject *la carretera* lacks dynamism and it cannot be involved in a change of location. Consequently, it coerces the verb into losing the phase of change (from 'not being here' to 'being here'); as a consequence, *llegar* turns into a state verb. Furthermore, (8e) is ambiguous because *el niño* 'the kid' (like *el corredor* 'the runner') denotes a dynamic entity (therefore it can participate in motion events), but it is also interpreted as an entity endowed with spatial extension, like *la carretera*; in these case, the adjunct *al botón del ascensor* 'the lift button' indicates the upper limit of the subject's extension (i.e. its height) (cf. De Miguel 2004a). The examples in (8f, g) illustrate a similar contrast: with the subject *la policía* 'the police', in (8f), the event can be either dynamic or static; the latter is the only possible reading with the subject *la valla* 'the fence', in (8g).

Thus, the data in (6-8) show the existence of interaction between the aspectual content of the verb and the interpretation of the event, triggered by arguments and adjuncts: this is what is called *compositional nature of lexical aspect*.

These data reveal one more crucial fact as far as our research is concerned: syntactic explanations have to include not only event information, but also other sub-lexical features, such as [INDIVIDUAL / COLLECTIVE], [POINT / EXTENSION], or

[BUILDING / INSTITUTION], as shown in (7), where the occurrence of each adjunct does not (only) depend on the aspectual information. (7a) is ambiguous, and (7b) is not, because of the sub-lexical information of *convento* 'convent', which denotes a kind of building, but also the activity developed in that place (with a meaning similar to 'institution'). These features of *convento* combine with the features of the adjunct and they trigger different event interpretations: on the one hand, if it is predicated that the event *salir del convento* 'come out of the convent' takes place *a través de la ventana* 'through the window', the only possible reading is that it is a punctual event of leaving the building; on the other hand, if the event *salir del convento* is combined with *a una edad temprana* 'at a young age', there are two possible readings: leaving the building or leaving the activity developed there ('to be a nun'), since the adjunct does not focuses on a point (*la ventana* 'the window') through which the change of location occurs, but rather on the subsequent state ('to be in a new state: either out of the building or out of the institution').

2.3. The lexical nature of compositional aspect

The contrasts observed in (6-8) depend on whether the subject is individual or collective (*Juan* vs. *el ejército* 'the army'), count or mass (*Juan* vs. *agua* 'water'); whether it is endowed or not with dynamism (*el corredor* 'the runner' vs. *la carretera* 'the road') or whether it is interpretable as a dynamic entity or as a static entity with extension (*el niño* 'the kid' vs. *la policía* 'the police'); whether the complement refers to a building or to an activity (*el convento* 'the convent'); whether the adjunct indicates a point or a spatial extension (*por la ventana* 'through the window' vs. *el país* 'the country'), etc. They suggest that building the verbal (and not only the aspectual) meaning is a compositional process; it takes place in a context and it is triggered by the interaction of different features encoded in the sub-lexical structure of the word: those features mentioned above and others like [±CONTAINER], [±INSTRUMENT], [±PREEXISTING], etc. This idea allows resolving the hypothetical lexical aspect paradox: it is compositional, but it is regulated by lexical agreement processes which either license certain constructions or they rescue other, a priori unexpected constructions, like *pintar el sol* 'paint the sun' in (1c), *comer un plato* 'eat a dish' in (1d), or *llegar la carretera* 'the road reaches' in (8d), depending on the sub-lexical features of the words involved. Therefore, we can invert the terms of the claim and talk about the *lexical nature of compositional aspect* (cf. De Miguel & Fernández Lagunilla 2007). This reformulation is not a mere word play; rather, it constitutes a general hypothesis about the building of verbal meaning and its impact on the syntactic behavior, within the framework of a theory which assumes that lexical information is mapped onto the syntax and that syntax materializes some lexical possibilities and leaves others as a mere potentiality.

From this point of view, periphrastic passive constructions in Spanish should be revised. As shown above in (3a, b) — now repeated as (9a, b)—, periphrastic passives require perfective predicates, but the restrictions are not only aspectual, as shown by (9c):

(9) a. El cuadro de *El grito* fue robado (por un fanático).
'*The Scream* painting was stolen by a fanatic'

b. *El cuadro de *Inocencio X* fue tenido en 2013 (por el Prado).
'The *Inocencio X* painting was had in 2013 (by the Prado Museum)'
(< El Prado tuvo el cuadro de *Inocencio X* en 2013;
'The Prado Museum had the *Inocencio X* painting in 2013')

c. El cuadro de *Las meninas* fue pintado *({por Velázquez / en 1616 / al óleo / para agradar al rey}).
'*Las meninas* painting was painted ({by Velázquez / in 1616 / in oil / to please the king})'

Indeed, the aspectual restriction does not explain the ungrammaticality of (9c): even though the predicate is perfective (*pintar el cuadro de "Las meninas"* 'paint *Las meninas* painting'), it is unacceptable unless one of the bracketed constituents (argument or adjunct) is materialized (*by Velázquez, in 1616, in oil*, or *to please the king*).

This behavior of some passive constructions led Grimshaw & Vikner (1993) to assume the existence of a mixed argument-adjunct constituent: it can be a *by*-complement bearing the agent θ-role (*por Velázquez* 'by Velázquez'), but also a non-selected constituent, like a temporal adjunct (*en 1616* 'in 1616'), a manner adjunct (*al oleo* 'with oil'), or a goal adjunct (*para agradar al rey* 'to please the king'). This proposal requires, on the one hand, the assumption of a hybrid constituent ("necessary adjunct"); and, on the other hand, it does not explain why this constituent is not necessary in some other passives, such as (9a), as shown in De Miguel (1992).

I argued recently (De Miguel 2012a, 2013, among others) that the different behavior of passives in (9a) and (9c) can be explained by taking into account certain non-aspectual information encoded by words: namely, *cuadro* 'painting' refers to an object which starts to exist through the event of being painted, so the passive in (9c) is unacceptable because it is redundant, unless another constituent makes it informatively relevant, like {*por Velázquez / en 1616 / al óleo / para agradar al rey*}. By contrast, *cuadro* is compatible with the predicate *ser robado* 'be stolen' because it is a [PHYSICAL OBJECT] noun and its definition does not encode that its function is to be stolen. Hence, the passive in (9a) is not redundant, it is acceptable, and it does not require any additional specifications; this is the reason why the *by*-complement is optional. This is a kind of analysis which maps the lexical information onto the syntax, insofar as sub-lexical features of words determine their combinations.

2.4. The underspecified definition of verbs: a basic schema specified in context

The hypothesis about the existence of underspecified verb definitions and lexical agreement mechanisms, triggering new senses from their combination with the sub-lexical features of arguments and adjuncts, allows explaining some facts which are apparently unexpected for classic aspectual studies, such as verbs that may belong to different aspectual classes. This is the case of *ver* 'see', which can be ascribed to any of the four Vendler's (1967) aspectual classes, depending on whether it is used with or without a complement, as in (10b,c,d) and (10a), respectively; in the former case, the sub-lexical features of the noun referring to the perceived object determine the type of the event: (10b) is a durative and endless event, an activity; (10c) is a non-durative event with an end, an achievement; finally, (10d) is a durative event with an end, an accomplishment:

(10) a. Sofía ve$_{\text{[state]}}$ muy bien. ('Sofía tiene muy buena vista')
'Sofía can see very well' ('Sofía has good eyesight')
b. Sofía ve el futuro$_{\text{[activity]}}$. ('Sofía adivina el futuro')
'Sofía can see the future' ('Sofía can guess the future')
c. Sofía vio un pájaro $_{\text{[achievement]}}$.
'Sofía saw a bird'
d. Sofía vio una película $_{\text{[accomplishment]}}$.
'Sofía saw a film'

Only when the verb expresses an accomplishment, the passive is acceptable, as the grammaticality contrast between (11a, b) and (11c) illustrates:

(11) a. *El futuro es visto por Sofía.
'The future is seen by Sofía'
b. *El pájaro fue visto por Sofía.
'The bird was seen by Sofía'
c. La película *La grande bellezza* fue vista por muchos espectadores.
'The film *The great beauty* was seen by many spectators'
d. *La película fue vista encima de la televisión por Sofía.
'The film was seen on the television set by Sofía'

Nevertheless, in addition to an object with temporal extension (*La película "La grande bellezza" dura más de dos horas* 'The film *The great beauty* lasts more than two hours'), the noun *película* 'film' denotes the physical medium which contains it (i.e., a [PHYSICAL OBJECT]); in this case, the event of seeing is punctual and the passive (11d) is rejected, as in (11b). This phenomenon shows the interesting interaction between the sub-eventual features of the verbs and the sub-lexical information provided by the arguments and adjuncts of the verbal predicate.
Almost three decades after the aspect emerged strongly in formal theoretical linguistics, it is still present in morphological and syntactic analyses.[54] However, its importance has been decreasing in recent years, since event information has been related to other concepts, some of which have a more primitive nature. Apart from passives, other grammatical constructions, like the middle voice or the secondary predication, which could not be explained in aspectual terms, have been recently accounted for by looking at the different types of information located in the inner structure of the words combined with the verbs (cf. for instance De Miguel 2004b; De Miguel & Fernández Lagunilla 2004; Batiukova 2009). Studies adopting this approach take a step from syntax to lexicology, and this is the framework of the lexicological and lexicographical project underlying DICEMTO.

[54] For example, Bisetto (2014) presents an analysis of the formation of Italian adjectives ending in –*bile* in aspectual terms.

3. The Diccionario electrónico multilingüe de verbos de movimiento (DICEMTO)

DICEMTO is a multilingual electronic dictionary project for motion verbs. The Spanish language is the basis of the dictionary, but it also includes translations into German, Arabic, Chinese, Slovene, Finnish, French, English, Italian, Japanese, Portuguese, Russian and Romanian.[55] The project is being developed by a team of nineteen experts: renowned specialists and PhD students.[56]

3.1. The goals of DICEMTO

The first goal of the project is applied: it aims at compiling a new electronic multilingual dictionary for motion verbs which offers two kinds of search: (a) the search intended for standard users interested in resolving practical questions, related to the encoding or decoding of motion verb constructions; (b) a search oriented to specialists interested in cross-linguistic study of the verbal encoding of motion.

As far as the second kind of search is concerned, the project has two theoretical goals: (a) to confirm the universal nature of the concept of motion, using the data collected in the dictionary; (b) to establish a list of parametrically constrained lexical-semantic features which determine the semantic and syntactic behavior of motion verbs, mainly by taking into account the extension and emptying mechanisms in the framework of the GL model.

3.2. The theoretical tenets of DICEMTO

This dictionary has been conceived as a theoretical lexicographical project. In other words, its design and development profits from some of the advances offered by theoretical linguistics, in order to describe and explain the semantic and syntactic behavior of lexical units.[57]

As shown in section 2, the meaning of the verb changes depending on its interaction with its arguments and adjuncts. As mentioned in section 1, this behavior can be explained by assuming that verbs have underspecified definitions, which can be specified when their sub-lexical features combine with the features of their arguments or adjuncts, by means of lexical agreement mechanisms. From this point of view, verbs do not have one literal meaning and a set of other, metaphorical

[55] The web site of the dictionary provides glossed translations of different uses of the verb *venir* 'come' into these languages (also into Armenian and Slovakian, which were included in the previous project): www.uam.es/gruposinv/upstairs/diccionario/index.htm.

[56] The list of participating researchers is found in www.uam.es/gruposinv/upstairs/equmiembros.htm.

[57] As mentioned in footnote 8, De Miguel (2014) reviews some of the basic assumptions on which the dictionary is based, and compares it with other ongoing lexicological and lexicographical projects. Batiukova & De Miguel (2013) and De Miguel & Batiukova (2014) present this assumptions and the structure of the entries in detail.

meanings; rather, all the possibilities are predicted in the minimal definition and are generated because the word allows it.

This hypothesis is confirmed by several facts: firstly, it is usually complicated to establish the canonical meaning (and the aspectual class) of verbs in the absence of context, as shown for *ver* 'see' in (10). Secondly, the so-called meaning modifications, extensions or emptyings are quite usual and relatively regular: metaphorical and metonymic examples, like *La carretera va del valle a la montaña* 'The road goes from the valley to the mountain', in (8b), or *El niño ha comido un plato* 'The kid has eaten a dish', in (1d), occur systematically in many languages. Besides, not any extension or emptying is possible, as shown by (1b): **El cuadro fue comido por Velázquez* 'The painting was eaten by Velázquez' or **La fuente del día está muy buena* 'The platter of the day is very good' vs. *El plato del día está muy bueno* 'lit. The dish of the day is very good' [='the daily special is very good']. This phenomenon suggests that the so-called metaphorical meanings are not special; rather, they are virtually contained in the word definition and are generated by means of the same process which gives rise to the so-called literal meanings. Therefore, an apparent extension of the literal meaning is only a virtual meaning materialized by the context.

Thus, words are not atomic entities, but rather entities with different sub-lexical features which determine their combinatorial potential. According to Cohen's (1986) metaphor, words are content sacks whose shape and form are adjusted to the content of the adjacent sacks. As defended in De Miguel (2009), those sacks have to be porous to enable the meaning particles to come out of the pores and combine with the adjacent meaning particles in a sort of osmosis. Still talking metaphorically, the feature combination, which takes place when a sub-lexical agreement mechanism operates, triggers either an interpretative collapse or a "lexical reaction". This reaction modifies the initial definitions of words through a process similar to "protoplasmatic kiss", which is the beautiful expression used by Santiago Ramón y Cajal to describe how the information is transmitted by neurons (Sherrington's scientific term is *synapse*).[58]

In sum, the sub-lexical features enclosed in minimal definitions of words have to agree or to adjust, in line with Pustejovsky (1995), which triggers their different senses. The consequences of this dynamic process of meaning building are visualized in word behavior; so, the syntax helps us to identify the opaque sub-lexical meaning, found in the heart of the word, as seen in passives in (3), (9), and (11).

As we know, one of the basic problems of lexicography is how to include in the dictionary the diverse meanings which words activate when combined with other words in context. The assumptions underlying DICEMTO simplify the task of the lexicographer, insofar as it is expected that different meanings can be deduced from the minimal definition, instead of enumerating them in the entry.

If this proposal is correct, it will be able to account for the multiple meanings of the motion verbs dealt with in the projected dictionary. They have been chosen, as

[58] I thank Lourdes González-Pietrosemoli (p.c.) for suggesting me both the beautiful metaphor and the similarities between the information transmission process by nerve cells and the idea of "sacks" transmitting lexical features.

pointed out above, because they typically appear in the different languages in expressions in which they seem to have lost their canonical meaning of motion, as in the periphrastic or metaphorical uses in (12):

(12) a. Sp. Desde entonces se observa como este pino *viene sufriendo* fuertes defoliaciones.
Port. Observa-se, desde então, que este pinheiro *vem sofrendo* de fortes desfoliações.
'Since then it has been noticeable how this pine has been suffering a serious loss of leaf.'
b. Sp. Después puedes enseñarle a usar OpenOffice, que ya *viene instalado.*
En. Afterwards you can teach him to use OpenOffice, which *comes* already installed.
c. Sp. Parece que *viene* tormenta del lado de la sierra.
En. A storm seems *to be coming* from the direction of the mountains.
It. Pare che dalla parte dei monti *stia venendo* un temporale.
Port. Parece que *vem* temporal do lado da serra.
Rom. Se pare că *vine* furtună dinspre munți.

3.3. The structure of DICEMTO

3.3.1. Modules of the dictionary

The dictionary consists of two modules: the "theoretical dictionary" or "theoretical module", which is the core of the project, and the "short dictionary" or "simple dictionary". The theoretical dictionary provides an exhaustive analysis of the meaning of every verb (for the moment, only for Spanish), which determines the information visualized in the "short dictionary" (designed for users interested in obtaining practical information about the verb use). The definitions in the "short dictionary" define the verb minimally in the absence of context: for instance, *venir* 'come' is defined as:

(13) *venir*: 'abandonar un lugar para dirigirse a otro describiendo una trayectoria hacia el hablante', use illustrated with one example.
venir 'come': 'leave a place in order to go to another place, tracing a path towards the speaker', use illustrated with one example.

On the other hand, the "theoretical dictionary" provides a verbal definition incorporated into a highly structured meta-entry, where sub-lexical features are encoded in different specific sub-lexical structures. The information enclosed in the verbal definition, combined with the information provided by co-occurring words, triggers its interpretative potential through lexical feature agreement processes.

3.3.2. The structure of meta-entries in the theoretical dictionary

In previous stages of the project, the design of the dictionary has been completed. It now has a unique meta-entry structure which serves as a template for specific verbal entries, and a list of tags, adapted to the TEI[59] standards, which are used to

[59] Text Encoding Initiative: cf. www.tei-c.org/Activities/Projects/di09.xml.

define the different sub-lexical structures integrated in the minimal verbal definition: the *event structure*, the *argument structure*, the *thematic structure*, and the *qualia structure*.

The level of the event structure (ES) encodes the event type; according to the GL, this level of representation encodes the aspectual features of the predicate in terms of sub-events, which are the essential, temporally and hierarchically organized parts of events. The combination of sub-events gives rise to different kinds of events. In the GL, there are three kinds of events: states, processes, and transitions. Fernández Lagunilla & De Miguel (2000) and De Miguel & Fernández Lagunilla (2004, 2007) modify this classification and propose the existence of eight kinds of ESs for Spanish. DICEMTO follows this classification, whose distinctions are useful to account for the behavior of verbs in periphrasis and non-literal meaning expressions. For example, the ES of the verb *venir* 'come' is the following:

(14) T2 [A3 [A → (P)] A2 [A → (S)]

According to the structure in (14), *venir* denotes a transition (T) between two achievement sub-events (A3 and A2), which, thus, consist of two phases: the head, which denotes the change of location, and a subsequent phase, which can be optionally visualized. Each achievement event, in turn, consists of two phases: on the one hand, the first sub-event of A3 is a simple achievement (A): *Juan vino* ***de Bilbao*** *(a Madrid) ayer* 'Juan came from Bilbao (to Madrid) yesterday', which expresses that Juan {left/stoped being in} Bilbao yesterday; it can be followed by a process (P): *Juan vino a Madrid* ***por Burgos*** 'Juan came to Madrid via Burgos', which expresses that Juan went to Madrid via Burgos. On the other hand, the first sub-event of A2 is a simple achievement (A): *Juan vino a Madrid ayer* 'Juan came to Madrid yesterday', which expresses that Juan {arrived/started being in} Madrid yesterday; and it can be followed by a state (S), which can be highlighted thanks to the aspectual clitic *se*: *Juan se vino a Madrid hace dos meses*, which expresses that Juan has been living in Madrid for two months.

The argument structure (AS) specifies the number of arguments required by the verb, and how they are realized syntactically. The thematic structure (TS) encodes the semantic function or θ-roles of the predicate arguments. The qualia structure (QS) encodes inherent semantic features of the arguments, such as [OBJECT], [EVENT], [INFORMATION], [INSTITUTION], and [CONTAINER]. This information is related to the properties of the real world entity denoted by the nouns. Its linguistic relevance was justified above (in sections 1 and 2), when explaining the difference between *plato* 'dish' and *fuente* 'platter', which are different kinds of [CONTAINER]; *película* 'film', which can be either an [ENTITY WITH TEMPORAL EXTENSION] or the [PHYSICAL OBJECT] which contains it; and *convent* 'convent', which can be either a [BUILDING] or an [INSTITUTION].

3.3.3. Modifications and idioms

The basic premise of this dictionary is that the definitions proposed for motion verbs, by interacting properly with feature lexical agreement processes, account for their use in non-literal meaning combinations. For instance, *venir* 'come', when combined with subjects that cannot be involved in motion (information encoded in their QS), loses the part of its ES which encodes the motion and gets reduced to the initial phase, where the change of state occurs.[60] This is the reason why it comes to mean 'start to occur, arise, emerge' when combined with subjects typed as [TEMPORAL UNITY], like *mañana* 'morning', *noche* 'night', *día* 'daytime', and *primavera* 'spring', or as [EVENT], like *muerte* 'death', *tormenta* 'storm', *lluvia* 'rain', *desgracia* 'misfortune', *problema* 'problem', *auxilio* 'assistance', *calma* 'calm', *soledad* 'isolation', and *satisfacción* 'satisfaction', as in *Ya viene el día* 'The daytime is coming' or in the Spanish example in (12c), *Parece que viene tormenta del lado de la sierra* 'A storm seems to be coming from the direction of the mountains'.[61] This sense is integrated into the section "Modificaciones" ('Modifications'). The user can access the glossed translations to other languages dealt with in the project through the links attached to the Spanish examples.

Periphrastic uses such as those in (12b, c) are also included in the section "Modificaciones" of the theoretical module. In each case, the modification detected in the ES of the verb (and the resulting change of meaning) is indicated. For example, the periphrasis *venir* + gerund is defined as the focalization of the part of the event comprising the beginning and the central point, without referring to the end; and the periphrasis *venir* + participle is characterized as the focalization of the resulting state. These properties are constrained by the main verbs which can appear in these periphrases. Through the attached links, the user can again find translations into other languages.

Lastly, the section "Expresiones idiomáticas" ('Idioms') presents cases which cannot be explained automatically by means of regular processes of emptying or extending of the underspecified definition of the verb; their features require a higher degree of specification, as in *venir a cuento* 'be applicable, relevant' (lit. 'come to story, tale') or *ir sobre ruedas* 'go well, like clockwork' (lit. 'go on wheels'), which are only acceptable in combination with non-human subjects, denoting respectively [INFORMATION] or [CONTENT] and [TOPIC] or [PROJECT]. Although the boundaries between the meaning emptying and extension on the one hand and the transformation into an idiom on the other hand are not always precise, this analysis aims to distinguish quite clearly between both kinds of cases. In this section, too, the user can access through links to the translations into the other languages.

4. Conclusions

To conclude, DICEMTO is a theoretical dictionary project, based on a dynamic, generative, and compositional model of lexical organization. The different meanings of the motion verbs are integrated into lexical meta-entries; their different meaning

[60] This is a lexical agreement process similar to the one making *llegar* 'arrive' lose its change-of-state phase when combined with a non-dynamic subject, as shown above in section 2.2 in the analysis of *llegar la carretera hasta la montaña* 'the road reaches the mountain'.

[61] This change of meaning is licensed by the assumption that events are not atomic entities; rather, they consists of phases or sub-events which can be focused on or remain hidden in context, in line with De Miguel & Fernández Lagunilla (2007). See Batiukova & De Miguel (2013) and De Miguel & Batiukova (2013) for details.

components are distributed among several levels of representation, whose interaction generates the varied meanings displayed by the verb in context.[62] This "radically monosemic" view of the treatment of polysemy has interesting theoretical, descriptive, and cross-linguistic consequences.

References

Almela, M.; Cantos, P. &; Sánchez, A. (2011). "Towards a Dynamic Combinatorial Dictionary: A Proposal for Introducing Interactions between Collocations in an Electronic Dictionary of English Word Combinations". *Procedings of eLex2011*, 1-11. www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-1.pdf

Almela, M.,; Cantos, P. &; Sánchez, A. (2013). "Collocation, Co-collocation, Constellation... Any Advances in Distributional Semantics", *Procedia - Social and Behavioral Sciences*, 95, 231-240. Also in www.sciencedirect.com

Batiukova, O. (2006). "Restricciones subléxicas para la formación de oraciones medias: ampliando la interficie léxico-sintaxis". In De Miguel, E. et al. (eds.). *Estructuras Léxicas y Estructura del Léxico*. Frankfurt am Main: Peter Lang. 329-345.

Batiukova, O. &; De Miguel, E. (2013). "Tratamiento lexicográfico de verbos de movimiento con significado amplio". *Actas del XLI Simposio de la SEL*. [on-line] Valencia: Universitat de València. 439-450. www.uv.es/canea/archivos/Estudios_linguistica_2013.pdf.

Bisetto, A. (2014). "Why Can One Be *irritabile* 'irritable' but Cannot Be **divertibile* 'amuse + able'? On Italian *–bile* Adjectives from Psychological Verbs". *Quaderns de Filologia. Estudis lingüístics*. Vol. XVIII (2013). 25-36.

Bosque, I. (2004). "Combinatoria y significación. Algunas reflexiones". In Bosque, I. (dir.). *REDES. Diccionario combinatorio del español contemporáneo.* Madrid: SM. LXXVII-CLXXIV.

Bosque, I. &; Mairal, R. (2012a). "Definiciones mínimas". In Rodríguez González, F. (ed.). *Estudios de lingüística española. Homenaje a Manuel Seco*. Alicante: Universidad de Alicante. 123-136.

Bosque, I. &; Mairal, R. (2012b). "Hacia una organización conceptual del *definiens*. Capas nocionales del adverbio *arriba*". In Corbella, D. et al. (eds.). *Lexicografía Hispánica del siglo XX. Nuevos proyectos y perspectivas. Homenaje al Profesor Cristóbal Corrales Zumbado*. Madrid: Arco Libros. 125-150.

Casares, J. (1950). Introducción a la lexicografía moderna. Madrid: CSIC.

Cohen, J. (1986). "How is Conceptual Innovation Possible?". *Erkenntnis*, 25. 221-238.

De Miguel, E. (1992). *El aspecto en la sintaxis del español: perfectividad e impersonalidad*. Madrid: Ediciones de la Universidad Autónoma de Madrid.

De Miguel, E. (1999). "El aspecto léxico". In Bosque, I. &; Demonte, V. (eds.). *Gramática Descriptiva de la Lengua Española*). Madrid: Espasa-Calpe. Vol. 2, Chapter 46. 2977- 3060.

De Miguel E. (2000). "Relazioni tra il lessico e la sintassi: classi aspecttuali di verbi ed il passivo in spagnolo". In R. Simone, R. et al. (eds.). *Classi di parole e conoscenza Lessicale. Studi Italiani di Linguistica Teorica e Applicata*. Monographic edition. Vol. XXIX, 2. 201-215. Also in *Círculo de Lingüística Aplicada a la Comunicación*, 2001, www.ucm.es/info/circulo/no8/demiguel.htm

[62] The definitions provided in DICEMTO only include the linguistic information which determines the syntactic behavior and semantic interpretation of the verb; the dictionary does not include encyclopedic information (cf. in this respect Batiukova and De Miguel's works quoted in the footnote 18).

De Miguel, E. (2004a). "Qué significan aspectualmente algunos verbos y qué pueden llegar a significar". In Cifuentes, J. L. &; Marimón, C. (eds.). *Estudios de Lingüística: el verbo. ELUA*. Monographic edition. 167-206.
rua.ua.es/dspace/handle/10045/9773.

De Miguel, E. (2004b). "La formación de pasivas en español. Análisis en términos de la estructura de *qualia* y la estructura eventiva". *Verba Hispanica*,. XII,. 107-129.

De Miguel, E. (2009). "La teoría del lexicón generativo". In De Miguel, E. (ed.). *Panorama de la lexicología*. Barcelona: Ariel. 337-368.

De Miguel, E. (2011). "En qué consiste ser verbo de apoyo". I Escandell, V.; Leonetti, M. &; Sánchez, C. (eds.). *60 Problemas de Gramática (dedicados a Ignacio Bosque)*. Madrid: Akal. 139-146.

De Miguel, E. (2012a). "Properties an Internal Structure of the Lexicon. Applying the Generative Lexicon Model to Spanish". In Sanz, M. & Igoa, J.M. (eds.). *Advances in the Sciences of Language and their Application to Second Language Teaching*. Cambridge: Cambridge Scholars Publishing. 165-200.

De Miguel, E. (2012b). "Verbos de movimiento en predicaciones sin desplazamiento espacial". In Kalenić Ramšak, B. et al. (eds.). *Actas del III Simposio Internacional 'La percepción del tiempo en lengua y literatura'* (Ljubljana, 24-26 november 2011). *Verba Hispanica*. Monographic edition. Vol. XX/1. Ljubljana: University of Ljubljana Publications. 185-207.

De Miguel, E. (2013). "La polisemia de los verbos soporte. Propuesta de definición mínima". In Torner, S. &; Bernal. E. (eds.). *Los verbos en el diccionario. Anexos Revista de Lexicografía*. Vol. 20. 67-109.

De Miguel, E. (2014). "La lexicología en España. Tendencias y proyectos en curso". In García Pérez, R. (dir.). *La lexicologie en Espagne: de la lexicologie à la lexicographie. Cahiers de Lexicologie.* Monographic edition. Vol. 104, 1. 17-44.

De Miguel, E. &; Batiukova, O. (2013). "Diccionario teórico de verbos amplios de movimiento: bases teóricas y desarrollo". In Sánchez Palomino., Mª D. (ed.) (in progress). *Lexicografía Iberorrománica*. Madrid: Arco/Libros.

De Miguel, E. &; Fernández Lagunilla, M. (2000). "El operador aspectual *se*". *Revista Española de Lingüística*. 30, 1. 13-43.

De Miguel, E. &; Fernández Lagunilla, M. (2004). "Un enfoque subeventivo de la relación entre predicados secundarios y adverbios de manera". Revue Romane. 39, 1. 24-44.

De Miguel, E. & Fernández Lagunilla, M. (2007). "La naturaleza léxica del aspecto composicional". *Actas del VI Congreso de Lingüística General. 2004*. Madrid: Arco/Libros-Universidad de Santiago de Compostela. Vol. II, tomo I. 1767-1778.

Fernández Lagunilla, M. &; De Miguel, E. (1999). "Relaciones entre el léxico y la sintaxis: adverbios de foco y delimitadores aspectuales". Verba. 26. 97-128.

Grimshaw, J. &; Vikner, S. (1993): "Obligatory Adjuncts and the Structure of Events". In Reuland, E. &; Abraham, W. (eds.). Knowledge and Language. Vol. II. Lexical and Conceptual Structure. Dordrecht: Kluwer. 145-159.

Pustejovsky, J. (1995). The Generative Lexicon. Cambridge, Mass.: MIT Press.

Ramchand, G. (2008). *Verb Meaning and the Lexicon. A First-Phase Syntax.* Cambridge: Cambridge University Press.

Vendler, Z. (1967). *Linguistics in Philosophy*. Ithaca: Cornell University Press.

Filling gaps in dictionary typologies: ROOTS - a morphological historical root dictionary

Alina Villalva & João Paulo Silvestre

The knowledge of the Portuguese lexicon has still many shortcomings. Some of them are well acknowledged; many other are quite unsuspected. Along the 20th century, Portuguese lexicography lost track of the ongoing research on word diachrony, morphological analysis and lexical semantics, unlike what happened with the lexicographic treatment of other comparable modern languages such as French, Castilian or Italian.

Although Portuguese dictionaries generally include information about etymology and morphological structure, they present it quite inconsistently, since it is generally the output of the accumulation of what can be found in previous dictionaries. This *modus operandi* leaves no room to a systematic analysis of the set of words and word families that form each dictionary's entry list. In fact, most contemporary dictionaries of Portuguese (paper editions and electronic versions as well) obey to conservative models, which tend to incorporate exhaustively the information made available by their predecessors, regardless of the real usage of the words. They devote a very considerable amount of effort to increase the entry list, accommodating neologisms (which are not that frequent) and specialized terms, randomly chosen by dictionary makers. From time to time, major dictionary publishers issue remakes of their own dictionaries for the sake of orthographic updates.

Although users tend to be unaware of their deficiencies, this kind of dictionaries is far from being a useful working tool. In fact, since all previously registered words tend to occur in every new dictionary, without any mention to their usage, users are led to induce that they are equally available, and that is not the case. A different type of dictionary, with a critically selected word list and a thorough lexicographic description, is thus still lacking and still necessary. The project we will present in this paper, ROOTS, has been conceived as a further step to build such a dictionary.

1. The project ROOTS - a morphological historical root dictionary

ROOTS is a research project that aims to produce a prototype of a specialized dictionary that intends to be a useful tool for linguists, lexicographers, translators and language teachers. More specifically, ROOTS is meant to provide a thorough description of word families, considering their presence in the Portuguese lexicon, from a semantic and a morphological point of view, both in diachrony and in their contemporary usage.

The intent to build a thorough morphological historical dictionary for Portuguese is quite striving on the account of feasibility. It requires a program and a step-by-step progression plan. The project ROOTS embodies an initial stage of this program. It is an exploratory project that aims to design a lexicographic model and to test it with roots of simple words[63]. These words are made of unanalysable roots and morphological specifiers (thematic suffixes and inflectional suffixes). Once these simple words are etymologically and semantically described, the model will accommodate their derivatives, thus yielding a dictionary of word families. So, the main goals of this project are the clarification of semantic and morphological issues in the evolution of the Portuguese lexicon, and the assessment of the communicative adequacy of the words that are registered in current Portuguese dictionaries, particularly by signalling obsolete words. Furthermore, if the model is replicated with other languages, multilingual dictionaries will become easier to compile and to validate.

The identification of a set of simple roots is crucial to better understand the lexicon since dealing with a smaller set of units potentiates coherence in the treatment of the data. Furthermore, although no such information is independently available, this project is ground on the assumption that the large amount of words that forms the lexicon of a language (namely Portuguese), regardless of their longer or shorter existence in the lexicon, are made out of a relatively small set of roots.

No existing dictionary provides the type of information that ROOTS is designed to include. In fact, most general dictionaries accumulate information from previous dictionaries and they often present inaccurate information concerning etymology and morphological structures. Even those featured as the best (European and Brazilian) Portuguese dictionaries lack systematic morphological description, historical review and usage assessment. As far as specialized dictionaries are concerned, although a better solution would be desirable, *Corpus Lexicográfico do Português* fulfils the needs that should be addressed by an heritage dictionary. Etymological dictionaries have insufficient information and they mirror the state of the art of the mid 20th century, even if they were produced later. Finally, two

[63] Complex lexicalized words behave simple words. The treatment of this kind of words is left to a latter stage of the program, for the sake of methodological choices.

specialized dictionaries deserve to be mentioned, although they are obviously outdated (either concerning the coverage, or the methodology adopted). The first one is *Dicionário de Raízes e Cognatos* (Goes 1921), which is based on 19th century lexicographic sources – it is mainly a list of neoclassic loans. The second one is the most important Portuguese morphological dictionary, compiled in Brazil. Unfortunately, it gathered lexical information from untrustworthy sources, such as general language dictionaries (cf. Heckler, Back, Massing 1984-1988). They both offer interesting data, but their consultation also requires extensive critical reading.

2. ROOTS: lexicographic model

This project aims to build a specialized dictionary containing a critical selection of a corpus of simple words, their lexicographic description and a usage labelling focused on contemporary Portuguese.

2.1. Critical word list selection

Being a prime source of lexical information, dictionaries can provide a good documentary basis for the collection of a corpus of simple words, and a corpus of the derivatives that contain the same root. In fact, the background for ROOTS is established by a canon of Portuguese dictionaries, from the 16th to the early 20th century, selected for qualitative reasons, since they all played an important role either for the quality of the information they provided or for the normative role that they assumed. At this early stage, we work with a digital corpus comprising the following pre-modern dictionaries: Cardoso 1569, Barbosa 1611, Bluteau 1712, available online (see *Corpus Lexicográfico do Português*, hence CLP), and Figueiredo 1913, also available online (see *Dicionário Aberto*). In the future, this corpus should be extended to include Morais Silva (namely the 1823 edition)[64].

Roots and derivatives included in our dictionary are strictly those that these sources legitimate. None of the dictionaries produced since the beginning of the 20th century are useful to collect simple words. Typically, the items they add are new derivatives, compounds or loans that can be left out of our search for the time being. Nevertheless, we always consult general contemporary dictionaries such as DLPC, Infopedia, Priberam (European Portuguese) and Houaiss 2009 (Brazilian Portuguese), in order to confirm that these words still have a register in these word lists.

Two written text *corpora* are systematically used: *Corpus de Referência do Português Contemporâneo* (CRPC) allows us to document the occurrence of words in the 20th century; *Corpus do Português* (CdP) is the database that gathers more

[64] For the time being, Morais Silva can be consulted online at www.brasiliana.usp.br/pt-br/dicionario/.

texts from the 15th century to the 19th century, which is useful information to establish the time span of each word.

All simple words are described, following the guidelines of Villalva & Silvestre (2014). The lexicographic description of each word includes an etymologic survey, a morphological analysis and a usage labelling.

2.2. Etymological survey

The etymological survey for each root can make use of Machado (1967) or Cunha (1986), which are the reference etymological dictionaries for Portuguese. However, the information they provide often requires further research. The comparison of Portuguese roots with their cognate equivalents in other languages helps to elucidate cases of loan, which are quite commonly found. Therefore, we regularly consult lexicographic *corpora*[65]. Furthermore, trying to fulfil what is expected of an historical dictionary, we need to sort chronologically the new meanings of polysemic words, through a semantic analysis of textual occurrences.

2.3. Morphological analysis

Morphological analysis intervenes to ensure that we appropriately select simple words, like *alto* 'high, notable' (and complex lexicalized words, like *altar* 'altar', at a later stage), that is, words that contain the root that will define a word family (i.e. *alt-* and *altar-*, respectively). Furthermore, morphological analysis will allow elucidating how the members of a word family (derivatives and modified words) relate to each other: some hang directly on the root, other depend from a simple word, and the remaining words relate a complex word:

(1) *alt(o/a)* $_{\text{Adjective Root}}$ → *altiva* $_{\text{Adjective (fem)}}$ → *altivamente* $_{\text{Adverb}}$
'notable' 'arrogant' 'arrogantly'

Morphological hierarchies of this kind are subordinate to the semantic analysis, but they also help to consolidate it.

[65] Namely Tesouro Informatizado da Lingua Galega (TILG); Tesouro Medieval Informatizado da Lingua Galega (TMILG); *Nuevo Tesoro Lexicográfico de la Lengua Española* (NTLLE); *Le Trésor de la langue française informatisé* (ATILF); *Tesoro della Lingua Italiana delle Origini* (TLIO).

2.4. Time span and usage labelling

One of the main goals of the project is to identify obsolete words, despite being registered in contemporary dictionaries. In the cases where a derived word does not have a significant occurrence in the CRPC database (inferior to 100, considering that the database has more than 300 million words) or when it is not documented by different textual sources, we will label it as obsolete. Thus, for each word (each meaning of each word, in fact) we establish a tentative time span, according to the registers found in CdP and CRPC. This information is far more stable then the usual identification of a 'first' occurrence, which systematically needs to be revised.

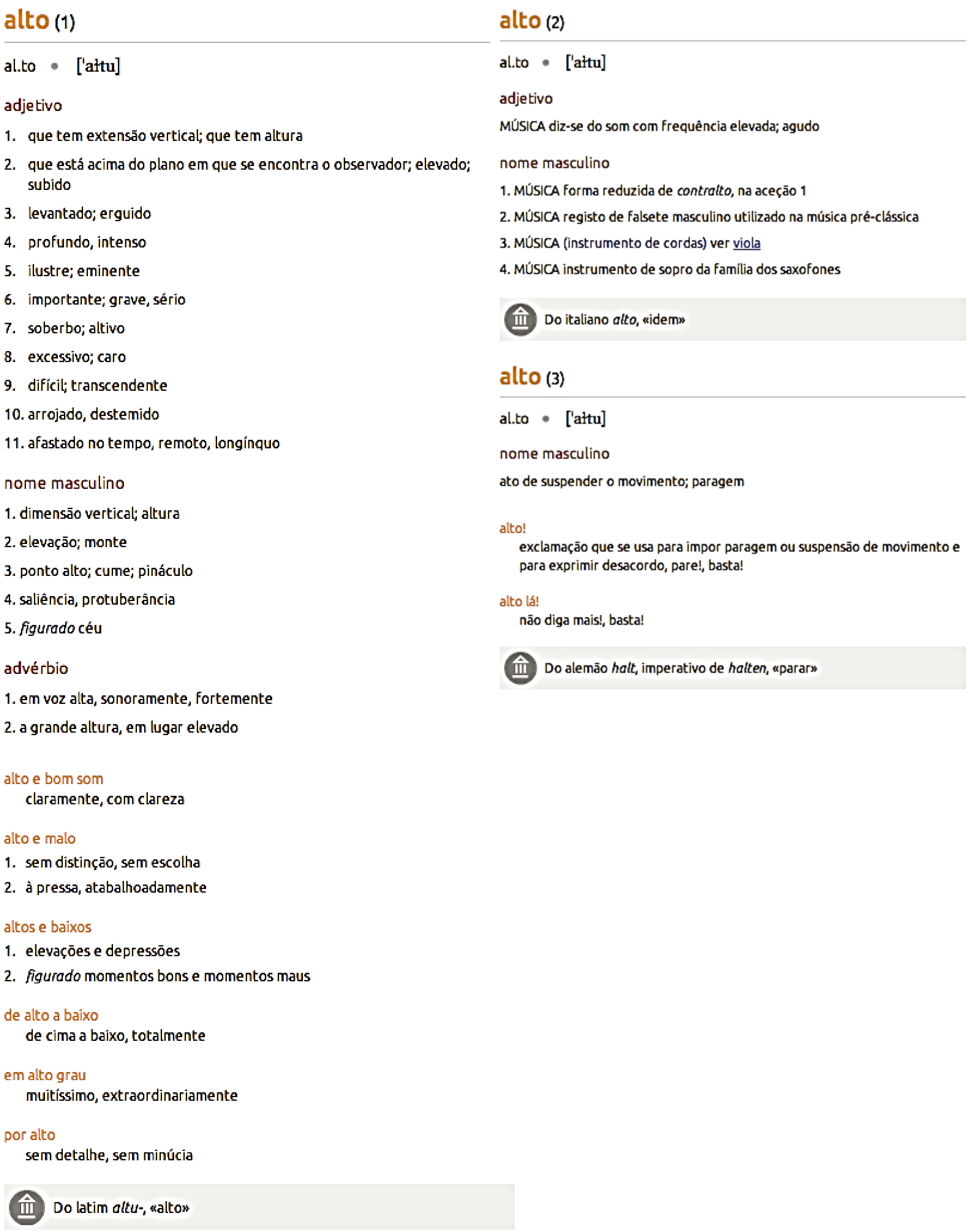
3. Case study: the root *alt*-

The first stage of ROOTS is devoted to adjectives. So far, we have studied a small number of adjectives that enabled us to test the model of description: *bravo* 'brave, angry', *esquisito* 'weird', *largo* 'large' and *comprido* 'long'[66]. We have intentionally selected words that have a long existence in Portuguese, which usually enables semantic mutations. Contrastive analysis is quite useful is these cases. In this paper, we will present the study of the root *alto* 'high, notable' as a case study and an example of ROOTS work plan, methodology and lexicographic model.

Alto is a very frequent word in Portuguese[67]. *Infopedia* provides the following information:

[66] See Villalva & Silvestre (2011) for *bravo*, Silvestre & Villalva (2014) and Villalva & Silvestre (in print).

[67] Davies & Preto-Bay (2008) lists the top 5000 most frequent words in the *Corpus do Português*. *Alto* is at the rank 185.



alto (1)

al.to • ['aɫtu]

adjetivo

1. que tem extensão vertical; que tem altura
2. que está acima do plano em que se encontra o observador; elevado; subido
3. levantado; erguido
4. profundo, intenso
5. ilustre; eminente
6. importante; grave, sério
7. soberbo; altivo
8. excessivo; caro
9. difícil; transcendente
10. arrojado, destemido
11. afastado no tempo, remoto, longínquo

nome masculino

1. dimensão vertical; altura
2. elevação; monte
3. ponto alto; cume; pináculo
4. saliência, protuberância
5. *figurado* céu

advérbio

1. em voz alta, sonoramente, fortemente
2. a grande altura, em lugar elevado

alto e bom som
claramente, com clareza

alto e malo
1. sem distinção, sem escolha
2. à pressa, atabalhoadamente

altos e baixos
1. elevações e depressões
2. *figurado* momentos bons e momentos maus

de alto a baixo
de cima a baixo, totalmente

em alto grau
muitíssimo, extraordinariamente

por alto
sem detalhe, sem minúcia

Do latim *altu-*, «alto»

alto (2)

al.to • ['aɫtu]

adjetivo

MÚSICA diz-se do som com frequência elevada; agudo

nome masculino

1. MÚSICA forma reduzida de *contralto*, na aceção 1
2. MÚSICA registo de falsete masculino utilizado na música pré-clássica
3. MÚSICA (instrumento de cordas) ver viola
4. MÚSICA instrumento de sopro da família dos saxofones

Do italiano *alto*, «idem»

alto (3)

al.to • ['aɫtu]

nome masculino

ato de suspender o movimento; paragem

alto!
exclamação que se usa para impor paragem ou suspensão de movimento e para exprimir desacordo, pare!, basta!

alto lá!
não diga mais!, basta!

Do alemão *halt*, imperativo de *halten*, «parar»

In this entry, we find a musical term (*alto* 2), which we will not consider here, and a homonym, which is in fact a different word (*alto* 3), and that is why we will disregard it as well. We will also skip the expressions included in 1, since they deserve a description on their own right, but not here. Thus, we will consider the adjective, the noun and the adverb *alto*, from *alto* 1, which are presented with an equivalent status, all of them being anchored in the Latin *altu-*. This is probably not the intended interpretation of this entry and in fact it is not difficult to find out that no *altu*- noun or *altu*- adverb ever existed in Latin. We could alternatively deduce that these three Portuguese words were derived from the Latin adjective, but no demonstration of that hypothesis has been found and it is quite absurd to admit it, since conversion processes that turn adjectives into nouns and adjectives into adverbs are quite frequent in contemporary Portuguese. The list of meanings (let us focus on the adjective for now) is also quite difficult to interpret: the first meaning uses a derivative (i.e. *altura* 'height') to explain its input, but the most difficult part is to decide what kind of hierarchy presides this ordering (frequency, antiquity, etc.). Some of these meanings are difficult to recognize in contemporary Portuguese (cf. *alto* = *profundo* 'deep') or independently of a specific collocation (cf. *alto* = *caro* 'expensive' vs. *preço alto* = 'expensive').

Another issue concerns the location of complex words that contain the root *alt*-. Alphabetically ordered paper dictionaries offer us the possibility to easily trace words beginning with *alt*- but the same is not true for roots that have been prefixed; some electronic dictionaries have a word list searching tool that emulates searching facilities of paper dictionaries, but some don't, and a few dictionaries allow word searches with wildcards that facilitate the location of the root in non-initial positions. For Portuguese, one such dictionary is the electronic version of the 1913 edition of Cândido de Figueiredo. The availability of this kind of search tools is far from being the solution we need. In fact, no available dictionary for Portuguese includes a morphological encoding that would allow to trace words containing a given root. They just allow to search graphic strings. The output of this search will be a word list that includes words containing the target root (e.g. *altivo* 'arrogant') and words containing a homographic string of characters (e.g. *esmalte* 'enamel'), which have to be discarded.

Figueiredo registers 340 words that contain the sequence *alt* in initial or medial position[68], but not all of them contain the root *alt*-. A huge time consuming manual selection will set apart words that are related to the root *alt*- from those that randomly contain the sequence <alt>. If we accept the definitions proposed by this lexicographer, we obtain a set of 92 words somehow related to the root *alt*-.

[68] This sequence can not occur in final position, since no Portuguese words end in <t>.

alta $_{\mathrm{N}}$ alta-roda altabaixo altabrava altaforma altamente altanado altanar-se altanaria altaneiramente altaneiro altania altar altar-mor altareiro altarista altavela alteação	altear alteável alteza altibaixa, o(s) altícomo alticornígero altifalante altiloquência altiloquente altiloquia altilóquio altíloquo altimetria altimétrico altímetro altimurado altinopolense altinopolitano altiplano	altipotente altirrostro altíssimo$_{\mathrm{ADJ}}$ altíssimo$_{\mathrm{N}}$ altissonância altissonante altissonantemente altíssono altista altitonante altitude altívago altivamente altivar altivez altiveza altivo altivolante altívolo	alto alto-alegrense alto-alemão alto-alentejano alto-beirão alto-comissário alto-duriense alto-falante alto-forno alto-minhoto alto-navarro alto-relevo alto-ribatejano altor altosa altura burro-alto contralto cornialto	enaltar enaltecer enaltecimento exaltação exaltadamente exaltado exaltador exaltamento exaltar pernaltas pernalteira pernalteiro pernalto pernaltudo planalto ribalta sobreexaltar superexaltado

A subset (marked in grey in the table above) is formed by 55 morphological compounds that will not be considered here, since they have involve at least another root (e.g. *planalto* = *plan* + *alt*). Thus, only 39 words are eligible members of the *alt-* root family:

alta $_{\mathrm{N}}$ altamente altanado altanar-se altanaria altaneiramente altaneiro altania	altar altareiro altarista alteação altear alteável alteza	altíssimo$_{\mathrm{ADJ}}$ altíssimo$_{\mathrm{N}}$ altista altitude altivamente altivar altivez altiveza	altivo alto altor altosa altura enaltar enaltecer enaltecimento	exaltação exaltadamente exaltado exaltador exaltamento exaltar sobreexaltar superexaltado

Since Figueiredo is presently a centennial dictionary, we crosschecked this list with the word list of more recent dictionaries (*DLPC*, *Infopedia* and *Priberam*). The comparison brought very few additions (four) and it revealed that not so many words have since been excluded. Crosschecking this final list with frequency data from CRPC proved that a significant amount of its members either have no records or they have a very low frequency, suggesting a very peculiar and restricted usage.

The following table comprises 43 words, including those that come from Figueiredo's list, those that appeared in later dictionaries, as well as their registers in all these dictionaries and their frequency in CRPC:

Figueiredo 1913	*DLPC 2001*	*Infopédia 2014*	*Priberam 2014*	*CRPC 2014*
alta $_{N}$	✓	✓	✓	6.684
altamente	✓	✓	✓	8.211
altanado	✗	✓	✓	0
altanar	✓	✓	✓	0
altanaria	✓	✓	✓	14
altaneiramente	✗	✗	✓	7
altaneiro	✓	✓	✓	43
altania	✗	✓	✗	0
altar	✓	✓	✓	1417
altareiro	✗	✗	✓	0
altarista	✗	✗	✓	0
alteação	✗	✓	✓	0
✗	*alteado*	✗	✓	28
✗	*alteador*	✓	✓	1
✗	*alteamento*	✓	✓	55
altear	✓	✓	✓	23
alteável	✗	✓	✗	0
alteza	✓	✓	✓	297
altíssimo	✓	✓	✓	1299
altíssimo $_{N}$	✓	✓	✓	132
altista	✓	✓	✓	166
altitude	✓	✓	✓	2.086
✗	✗	*altitudinal*	✗	0
altivamente	✓	✗	✓	53
altivar	✗	✗	✗	0
altivez	✓	✓	✓	180
altiveza	✗	✗	✗	0
altivo	✓	✓	✓	174
alto $_{ADJ}$	✓	✓	✓	50.604
altor	✗	✗	✗	0
altosa	✗	✓	✓	0
altura	✓	✓	✓	70.768
enaltar	✗	✗	✗	0
enaltecer	✓	✓	✓	675
enaltecimento	✓	✓	✓	50
exaltação	✓	✓	✓	1095
exaltadamente	✗	✓	✓	19
exaltado	✓	✓	✓	896
exaltador	✗	✓	✓	4
exaltamento	✓	✓	✓	5
exaltar	✓	✓	✓	533
sobreexaltar	✓	✓	✓	0
superexaltado	✗	✗	✗	0

The analysis of this data allows us to conclude that contemporary dictionaries have a quite stable entry list since the beginning of the 20th century. Around half of the words in this table (22 out of 43) are present in all of these four dictionaries. According to frequency information from CRPC, two of them are very frequent words (cf. 2a); six other are quite frequent (cf. 2b); eight words have a low frequency (cf. 2c); five have a very low frequency (cf. 2d); and two of them are inexistent in this corpus (cf. 2e).

(2)	a.	*altura*	CRPC = 70.768
		alto/a(s)$_{\mathrm{ADJ}}$	CRPC = 50.604
	b.	*altamente*	CRPC = 8.211
		alta$_{\mathrm{N}}$	CRPC = 6.684
		altitude	CRPC = 2.086
		altar	CRPC = 1.417
		altíssimo$_{\mathrm{ADJ}}$	CRPC = 1.299
		exaltação	CRPC = 1.095
	c.	*exaltado*	CRPC = 896
		enaltecer	CRPC = 675
		exaltar	CRPC = 533
		alteza	CRPC = 297
		altivez	CRPC = 180
		altivo	CRPC = 174
		altista	CRPC = 166
		altíssimo$_{\mathrm{N}}$	CRPC = 132
	d.	*enaltecimento*	CRPC = 50
		altaneiro	CRPC = 43
		altear	CRPC = 23
		altanaria	CRPC = 14
		exaltamento	CRPC = 5
	e.	*altanar*	CRPC = 0
		sobreexaltar	CRPC = 0

Furthermore, five words from Figueiredo's list have been excluded in subsequent dictionaries. None of them occurs in the contemporary corpus, which may suggest that they are obsolete words, but tagging words as obsolete requires other tools, as we will see later. This status is probably adequate for four of them (cf. 3a), but not for the last one (cf. 3b). Since *superexaltado* is a superlative adjective, which we do not expect to typically occur in written texts, its absence in CRPC is probably due to the design of the corpus and not to the absence of the word in contemporary European Portuguese. Notice that, unlike those words in (3a), *superexaltado* is a complex compositional word, a kind of word that dictionary makers may decide to include or leave out of the word list.

(3)	a.	*altivar*	CRPC = 0
		altiveza	CRPC = 0
		altor	CRPC = 0
		enaltar	CRPC = 0
	b.	*superexaltado*	CRPC = 0

Six other words have been excluded only by the DLPC. Three of them (cf. 4a) have no records in the contemporary corpus, which should be interpreted as in the previous case: they may be obsolete words. The remaining two (cf. 4b) are quite infrequent in the corpus, but they are compositional complex words, which typically have a low frequency usage: speakers when needed form them, and their interpretation is fully predictable. As we said above, dictionary makers can choose to include them in the word list, or not, but the decision must be systematically respected. DLPC is not systematic, since it excludes *exaltadamente* and *exaltador*, but it includes *exaltado* and *enaltecimento*, for instance, as all the other dictionaries do, and it also includes *altivamente*, which is excluded only by *Infopedia* (cf. 4c). Words of this kind are not obsolete words:

(4)	a.	*altanado*	CRPC = 0
		altosa	CRPC = 0
		alteação	CRPC = 0
	b.	*exaltadamente*	CRPC = 19
		exaltador	CRPC = 4
	c.	*altivamente*	CRPC = 53

Furthermore, five words from Figueiredo's list fail to be present in DLPC and one of the other dictionaries: *Infopedia* in (5a) and *Priberam* in (5b). These cases are not very different from those just mentioned: either they are obsolete or they don't need a dictionary register. The problem is that all these dictionaries make random choices of what to include and what to let out.

(5)	a.	*altaneiramente*	CRPC = 7
		altareiro	CRPC = 0
		altarista	CRPC = 0
	b.	*altania*	CRPC = 0
		alteável	CRPC = 0

Finally, only four different words appear in more recent dictionaries. We might think that these words were neologisms, because these dictionaries are quite recent (or recently updated), but they are not and they are also quite infrequent in the corpus. Once again, these are complex compositional words that do not need to be present in general dictionaries:

(6)	*alteado*	CRPC = 29
	alteamento	CRPC = 55
	alteador	CRPC = 1
	altitudinal	CRPC = 2

Now that we have a list of words that contain the root *alt*-, we must decide if they are all members of the same word family, or not. Word family is a concept that needs to be handled with care for two main reasons[69]. The first one is related to the fact that there is no consensual understanding of what a word family might include (e.g. compound words, inflectional paradigms, etc.). In this project, word families include simple words (e.g. *alto* 'high'), documented compositional derivatives (e.g. *altura* 'height'; *altivez* 'pride') and documented compositional modified words (e.g. *altíssimo* 'very high'). Consequently, word families in ROOTS exclude (cf. 7a) lexicalized derivatives that will head a family of their own (usually these words are loans) and (cf. 7b) words that are cognates but in contemporary Portuguese they are morphologically unrelated (most frequently, the root has a different form) – they will also head a family of their own:

(7) a. the derivative lacks a base form:
*altano → *altaneiro* 'proud'
no such suffix is available in contemporary Portuguese:
*alt**[*itude*] 'altitude'
*alt**[*ar*] 'altar'
no such prefix is available:
*[*ex*]*altar* 'to exalt'
words that have undergone a semantic shift:
altosa 'long wool'

b. *alça* 'handle' ← alçar 'to elevate' < Lat. **altiāre*
outeiro 'hill' < Sp. otero < Lat. *altarĭu*

These exclusion criteria reduce the *alt*- word family to a set of 23 words:

***alt*- word family**	**other word families**
alta $_{\text{N}}$	*altaneiro*$_{\text{ADJ}}$
altamente	↳ *altaneiramente*$_{\text{ADV}}$
alteação	*altanaria*$_{\text{N}}$
alteado	*altania*$_{\text{N}}$
alteador	*altanar*$_{\text{V}}$
alteamento	↳ *altanado*$_{\text{ADJ}}$
altear	> *altar*$_{\text{N}}$
alteável	↳ *altareiro*$_{\text{N}}$
alteza	↳ *altarista*$_{\text{N}}$
altíssimo	*altitude*$_{\text{N}}$
altíssimo $_{\text{N}}$	↳ *altitudinal*$_{\text{ADJ}}$
altista	*altosa*$_{\text{N}}$

[69] See Bauer and Nation (1993) for further discussion.

altivamente	*exaltar*$_{\text{V}}$ ↳ *exaltação*$_{\text{N}}$ ↳ *exaltado*$_{\text{ADJ}}$ ↳ *superexaltado*$_{\text{ADJ}}$ ↳ *exaltadamente*$_{\text{ADV}}$ ↳ *exaltador*$_{\text{N}}$ ↳ *exaltamento*$_{\text{N}}$ ↳ *sobreexaltar*$_{\text{V}}$
altivar	
altivez	
altiveza	
altivo	
alto$_{\text{ADJ}}$	
altor	
altura	
enaltar	
enaltecer	
enaltecimento	

The second issue related to the concept of word family is that sharing a root is not the only relevant feature wrt to the relationships these words have with each other (cf. Williams 1981). In this project, we decided to show the hierarchy of morphological relationships, in order to distinguish words that are derived or modified from the root (cf. 8a) from words that are derived or modified from previously derived or modified words (cf. 8b):

(8)	a.	*alto* → *altura*	'high → height'
		alto → *altíssimo*	'high → very high'
	b.	altivo → *altivez*	'arrogant → pride'
		altissima → *altissimamente*	'very high → very highly'

The output of this morphological hierarchy is as follows:

(9) *alt*$_{\text{ADJR}}$	
↳ *alto/a*$_{\text{ADJ}}$	CRPC = 50.604
↳ *altamente*$_{\text{ADV}}$	CRPC = 8.211
↳ *alta*$_{\text{N}}$	CRPC = 6.684
↳ *altíssimo/a*$_{\text{ADJ}}$	CRPC = 1.299
↳ *altíssimo*$_{\text{N}}$	CRPC = 132
↳ *altivo/a* $_{\text{ADJ}}$	CRPC = 174
↳ *altivamente*$_{\text{ADV}}$	CRPC = 53
↳ *altivez*$_{\text{N}}$	CRPC = 180
↳ *altiveza*$_{\text{N}}$	CRPC = 0
↳ *altivar*$_{\text{V}}$	CRPC = 0
↳ *altista*$_{\text{ADJ}}$	CRPC = 166
↳ *altura*$_{\text{N}}$	CRPC = 70.768
↳ *alteza*$_{\text{N}}$	CRPC = 297
↳ *altor*$_{\text{N}}$	CRPC = 0
↳ *altear*$_{\text{V}}$	CRPC = 23
↳ *alteação*$_{\text{N}}$	CRPC = 0
↳ *alteado/a*$_{\text{ADJ}}$	CRPC = 29
↳ *alteador*$_{\text{N}}$	CRPC = 1
↳ *alteamento*$_{\text{N}}$	CRPC = 55
↳ *alteável*$_{\text{ADJ}}$	CRPC = 0
↳ *enaltar*$_{\text{V}}$	CRPC = 0
↳ *enaltecer*$_{\text{V}}$	CRPC = 675
↳ *enaltecimento*$_{\text{N}}$	CRPC = 50

Now that the word family is defined, we must evaluate the usage of each of its members. Frequency figures (from CRPC) shed some light on the set of words that are more or less generally used, but used or rarely used words require further research, since they do not form an homogeneous set. This survey can help to set apart the set of words that have a real existence from those that have a merely fictitious survival in dictionaries, as a result of editorial options rather than a linguistic validation.

Words that were used in past synchronies, but are no longer in use will be marked as obsolete (cf. 10a); words that have a dictionary register but no matches in the corpus will be marked as undocumented (c. 10b). The first set includes old words and also, quite frequently, phonetic or morphological old variants of words that are still available; the second set includes possible words and morphological duplicates.

(10)	a.	*altiveza* 'pride'	cf. *altivez* 'pride'
		altor 'hight'	cf. *altura* 'hight'
	b.	*altivar* 'to make arrogant'	
		alteável 'that can be elevated'	
		alteador 'that elevates'	
		alteação 'elevation'	cf. = *alteamento*
		enaltar 'to make high'	cf. = enaltecer

The last step is probably the most decisive to the design of our dictionary prototype. Simple words often have a polysemic character, often related to the fact that they have long been present in the lexicon. New meanings may accumulate with old meanings, and some old meanings may also become obsolete. The morphological hierarchy presented in (9) does not incorporate semantic information, which is an essential component of this lexicographic model. The list of meanings that can be found in the source dictionaries often includes proper meanings and meanings that depend on collocations, as well as meanings that correspond to common usage and other that don't. Hence, we decided to extract the relevant meanings of each simple word that can be documented in a text corpus, namely *Corpus do Português*.

Once all the meanings of a given root have been deciphered, we can link each complex word to the relevant meaning and, thus, doublecheck the semantic analysis as well as the chronological marking. In the case of *alt-*, the output of the morphosemantic analyses, and a brief etymological description, is as follows:

alt- < **Lt.** ***altu-***$_{\mathbf{ADJ}}$ (de *altu-* ‘alimentado, crescido’, particípio de *alo*, *-ere* ‘alimentar’)

1.	***alto/a***$_{\mathbf{ADJ}}$	**profundo ‘deep’**	13th - 18th
	altíssimo/a$_{ADJ}$	muito profundo ‘very deep’	17th
	altamente$_{ADVloc}$	profundamente ‘deeply’	15th - 17th
	altura$_{N}$	profundidade ‘depth’	13th - 19th
	alto$_{N}$	mar profundo ‘deep see’	14th –
2.	***alto/a***$_{\mathbf{ADJ}}$	**grande ‘tall’**	13th –
	↳ *alto*$_{ADV}$	(num tom) elevado ‘loudly’	16th –
	↳ *altamente*$_{ADVintens}$	muito ‘very’	17th –
	altíssimo/a$_{ADJ}$	muito alto ‘very tall’	19th –
	alto$_{N}$	cimo ‘top’	15th –
	alta$_{N}$	cimo ‘top’	13th –
	alta$_{N}$	subida ‘rise’	19th –
	↳ *altista*$_{ADJ}$	em subida ‘rising’	20th –
	alta$_{N}$	permissão ‘consent’	18th –
	alteza$_{N}$	qualidade do que é alto ‘height’	14th, 18th
	altura$_{N}$	qualidade do que é alto ‘height’	15th –
	altear$_{V}$	tornar mais alto ‘make high(er)’	19th –
3.	***alto/a***$_{\mathbf{ADJ}}$	**ilustre ‘notable’**	14th –
	altíssimo/a$_{ADJ}$	muito ilustre ‘very notable’	17th –
	↳ *altíssimo*$_{N}$	deus ‘god’	15th –
	altamente$_{ADV}$	de um modo ilustre ‘in a notable way’	14th –
	altivo$_{ADJ}$	superior (positivo) ‘superior (positive)’	16th –
	↳ *altiveza*$_{N}$	qualidade do que é superior (positivo) ‘quality of being superior (positive)’	17th
	↳ *altivez*$_{N}$	qualidade do que é superior (positivo) ‘quality of being superior (positive)’	17th -
	↳ *altivo*$_{ADJ}$	superior (negativo) ‘superior (negative)’	17th –
	↳ *altiveza*$_{SUBS}$	qualidade de ser superior (negativo) ‘quality of being superior (negative)’	16th – 17th
	↳ *altivez*$_{SUBS}$	qualidade de ser superior (negativo) ‘quality of being superior (negative)’	18th –
	alteza$_{N}$	qualidade do que é ilustre ‘quality of being notable’	14th – 18th
	↳ *alteza*$_{N}$	pessoa ilustre ‘notable person’	15th –
	altura$_{N}$	‘quality of what is notable’	14th –
	enaltar$_{V}$	tornar notável ‘to make notable’	19th
	enaltecer$_{V}$	tornar notável ‘to make notable’	19th –

So, the adjective *alto* has three meanings. Meaning 1 and 2 have older registers, but only the second remains available. Meaning 3 corresponds to a metaphorical diversion from meaning 2. Notice that the first meaning of *alto*, which is ‘deep’ is lost in contemporary usage (although we can not deduce this from CRPC frequencies, which do not differentiate meanings). The only exception is *alto*$_{N}$, meaning ‘deep see’, which is lexicalized, and still available. Finally, derivatives may be exclusive from a given meaning (cf. *alto*$_{N}$ ‘deep see’, from meaning 1; *altista*$_{ADJ}$

'rising', from meaning 2; *enaltecer*$_V$ 'to make notable', from meaning 3), but they may also be polysemic, along the lines that the root is (cf. *altura*$_N$ 'depth', 'height', 'importance').

Conclusion

To our knowledge, ROOTS is an innovative project, which includes a new lexicographic model. Its main purpose is to provide an adaptive tool for the description of the core lexicon of a given language. Portuguese is a good case study, because there is an obvious deficit on such kind of specialized dictionaries, but it is also interesting because of its position in the space of romance languages. In fact, the lexicographic model designed for ROOTS facilitates the detection of contrastive issues, such as the direction of loans and semantic divergencies.

References

Bauer L.; Nation, P. (1993) Word Families. *International Journal of Lexicography*, 6, 4 (253-279).

CdP - Corpus do português: 45 million words, 1300s-1900s. Online: www.corpusdoportugues.org [10/09/2014].

CLP - Corpus Lexicográfico do Português. Universidade de Aveiro - Centro de Linguística da Universidade de Lisboa Online: clp.dlc.ua.pt

CRPC - Corpus de Referência do Português Contemporâneo. Centro de Linguística da Universidade de Lisboa. Online: www.clul.ul.pt/pt/recursos/183-reference-corpus-of-contemporary- portuguese-crpc)

Figueiredo, C. de (1913). *Novo Diccionário da Língua Portuguesa*. Porto: Typ. da Empr. Litter. e Typographyca.

Goes, C. (1921) *Dicionário de Raízes e Cognatos da Língua Portuguesa.* Belo Horizonte: Paulo Azevedo & Cia.

Heckler, E.; S. Back, E.R. Massing (1984-1988) *Dicionário Morfológico da Língua Portuguesa*. São Leopoldo: Unisinos.

Houaiss, A.; Villar, M. (2009). *Dicionário Houaiss da Língua Portuguesa*. Rio de Janeiro: Objetiva.

Infopédia. Dicionário da Língua Portuguesa da Porto Editora. Online: www.infopedia.pt/lingua-portuguesa/ [10/09/2014].

NTLLE - Nuevo Tesoro Lexicográfico de la Lengua Española. Online: ntlle.rae.es/ntlle/SrvltGUILoginNtlle [10/03/2014].

Silvestre, J. P.; Villalva, A. (2014). A morphological historical root dictionary for Portuguese. In A. Abel, C. Vettori & N. Ralli, eds. (2014) *Proceedings of the International Congress EURALEX XVI: The User Focus*. Bolzano (967-978).

Silvestre, J. P.; Villalva, A. (2015). Mutations lexicales romanes: *esquisito*, *bizarro* et *comprido*. *InnTrans: Innsbrucker Beiträge zu Sprache, Kultur und Translation*. Band 7. Frankfurt am Main: Peter Lang Verlag (149-165).

TILG - Tesouro Informatizado da Lingua Galega. Online: http://ilg.usc.es/TILG/

TLFI - Le Trésor de la langue française informatisé. Online: atilf.atilf.fr/tlf.htm

TLIO - Tesoro della Lingua Italiana delle Origini. Online: http://tlio.ovi.cnr.it/TLIO/

TMILG - Tesouro Medieval Informatizado da Lingua Galega. Online: http://ilg.usc.es/tmilg/

Villalva, A.; Silvestre, J. P. (2014) *Introdução ao Estudo do Léxico. Descrição e Análise do Português.* Petrópolis: Vozes.

Villalva, A.; Silvestre, J. P. (2011) De *bravo* a *brabo* e de volta a *bravo*: evolução semântica, análise morfológica e tratamento lexicográfico de uma família de palavras. *ReVEL* 9, 17. Online: www.revel.inf.br/files/artigos/revel_17_de_bravo_a_brabo.pdf [09/10/2014]

Vocabulário Ortográfico do Português.
Online: www.portaldalinguaportuguesa.org/vop.html [10/03/2014].

Williams, E. (1981) On the notions 'lexically related' and 'head of a word'. In *Linguistic Inquiry* Vol. 12, No. 2 (245-274).

Is a language dictionary of proper names feasible?

Jean-Louis Vaxelaire

1. Introduction

When I was a child, my teachers used to tell me that dictionaries were the most useful tools we had to learn a language. It is the practice of translation that revealed to me that dictionaries were not always infallible, especially when we have to deal with the outer reaches of the lexicon such as slang words, borrowings or proper names. For several reasons, monolingual dictionaries do not integrate all possible words and focus on the core of the lexicon. To offset these losses, there are specialized dictionaries, such as slang dictionaries, etc.

The case of proper names in France is interesting because although they are generally excluded from general lexicography, there are proper names dictionaries such as *Petit Robert des noms propres* (a separate volume from the language dictionary *Petit Robert*) or the last part of the *Petit Larousse*, which means they are not completely forgotten. It is also true that some American dictionaries have included proper names but, as Zabeeh (1968: 35) points out, dictionary entries for names, in both general and specialized dictionaries, typically provide their etymological meaning or descriptions of famous bearers of the name, not a real description of sense. We will see that the problem is exactly the same with French proper names dictionaries, they are not dictionaries in the strict sense but rather some kind of encyclopedia.

If we look further into the issue, it soon becomes clear that as proper names are generally excluded from language dictionaries, this is also sometimes the case with derivatives or phrases that contain proper names. Thus, the expression "Ce n'est pas le Pérou" is not defined in the well-known dictionaries Larousse or Robert, either in the language part (*Peru* cannot be a headword and it is difficult to pick out the expression under the *être* entry) or in the proper names part where all we can learn about Peru is that it is a country in South America with a certain number of inhabitants and a certain number of square kilometers.

I will discuss in this paper the French situation, but it is in varying degrees equivalent to that of most countries. A whole area of linguistic information is rejected from dictionaries though it is no less important than many terms dictionaries do define.

2. An unfilled need

It would be useful to integrate these elements (or their derivatives) for students, for all who are interested in French culture and for translators, because these words also carry with them elements of culture. It seems to me difficult to read the French press without the help of a tool that takes this factor into account. Thus, during the summer of 2013, when former President Nicolas Sarkozy had problems with judges, his followers compared his legal proceedings with *Outreau* and *Dreyfus*. If the name *Dreyfus* appears in proper names dictionaries where his relations with the law are generally described, the name *Outreau* only exists as the name of a town in these dictionaries. The link between these two names is that they are the symbols of (old and recent) errors made by the French justice system. There are very many occurrences of *Outreau* in the French press nowadays, the name will maybe disappear in ten or twenty years, but right now, it would be useful to find out from a dictionary that it refers to a miscarriage of justice.

Similarly, in this example from *Le Canard Enchaîné*, "Bref, une soirée publicitaire très privée. Et très Saint-Tropez!" (05/12/07), the French understand that the use of the name of a town in this adjectival position is due to the reputation of Saint-Tropez as the meeting place of the newly rich and the jet set. If this reputation is known on several European countries, this is not necessarily the case in Asia or South America.

In the following example, taken from a journalist's blog, the author is making a play on words with a French expression:

> Arrive (Renault-)Dacia, qui leur offre, excusez du peu, de retrouver leur dignité - sous la forme d'un salaire de folie de, tiens-toi bien, 285 euros mensuels.
>
> Byzance.
>
> http://vivelefeu.blog.20minutes.fr/, 25/03/2008

The ironic context (285 euros cannot be an exorbitant salary since it is well below the French minimum wage) tells the reader that we should not read *Byzance* in a literal sense (the ancient city of Byzantium). So we reconstruct the phrase "Ce n'est pas Byzance" which means that it is not something exceptional. The main lexicographical problem with this type of phrase is that the proper name carries most of the semantic load, so the phrase is ignored by the language dictionaries that do not include *Byzance* as a headword and forgotten in dictionaries of proper names which have more information to give than a definition of this phrase.

While speaking about Alain Rey, Meschonnic (2010: 24) writes that Rey should not be described as a "puits de science[70]" but as "a mountain, a Himalaya". It is possible to make a referentialist reading thanks to a proper names dictionary, however what is important is not whether the Himalaya is actually the highest mountain range on the planet, but that in our culture this is the paragon of a gigantic mountain. If someone were to discover tomorrow that due to a technical

[70] "a fount of knowledge" but the word *puits* means "well", i.e. a kind of hole, the opposite of a mountain.

error, another mountain range in Antarctica is higher than the Himalaya, we could still continue to use this metaphor. We all know that the sun does not really rise in the morning but we will not drop the word *sunrise*, why would it be different with proper names?

In a last example, *Le Canard Enchaîné* describes a political meeting "dans une ambiance plus proche de la foire d'empoigne que de *La petite maison dans la prairie*"(28/08/13). There is no antonym for *foire d'empoigne* ("free-for-all") in French and as *Little House on the Prairie* is the symbol of naive good feelings, it could be seen as a kind of antonym. Of course, this is not typically French as European culture is more and more intertwined with the globalized Hollywood culture.

3. Inescapable questions

I should repeat that in this paper when I'm talking about names, I do not limit myself to proper names as such but also to all their derivatives, etc., because they are all elements that tend to be excluded from French dictionaries[71].

Two projects are possible to erase this problem,: the first one is to try to improve the existing dictionaries, the second is to create a new specialized dictionary.

If one wants to create a dictionary type that does not yet exist, several questions must be asked: why don't they exist today? Can this role be fulfilled by other dictionaries (one or several combined)? Will the public be interested by this tool? There are obviously other relevant questions, but I will limit myself to these.

1) The answer to the first question will seem paradoxical since proper names dictionaries already exist in several languages, but you cannot judge a book by its cover, and it is not because it is written *dictionary* that it has to be one. In a previous work (Vaxelaire, 2005b), I analyzed the French proper names dictionaries and came to the conclusion that they are closer to encyclopedias and have little in common with dictionaries. For instance:

1. The classification is not strictly alphabetical, the various *Martin* are classified according to their date of birth rather than with their first name.
2. There are only a few phonetic transcriptions and almost only for French names that do not follow the norm, so it is impossible to know what is the pronunciation of the foreign name *Phuket*. All we can do is trying to guess whether we must pronounce the first two letters sound [f] or [p]. Similarly, is the letter U a [y] as in French, a [u] or closer to what exists in English like [ʌ]?
3. There is little or no information about gender and number.
4. It is often impossible to find out about the use of a determiner. Thus, the two articles *RTBF* and *RTL* follow one another in the *PLI* without any indication that the first works without an article (*RTL*), which is not the case of the second (*la RTBF*).
5. There are very few phrases containing proper names, those identified relate to unfamiliar referents (*aller à Canossa*, *franchir le Rubicon*), while those containing

[71] I write here French because I have shown in an article (forthcoming) that Italian dictionaries are less reluctant to include antonomasias such as *Rambo* in the sense of "violent man" or *Waterloo* as a "disastrous defeat".

the names of famous people or places are left out (*le quart d'heure de Rabelais, ce n'est pas le Pérou*).

6. There are no entries for morphemes derived from proper names, even though they are extremely productive as *Mc-* or *-gate*.

7. There are no entries for names referring to nonexistent or non-ontological references. When I say non-existent referent, I am not thinking about fiction (referents exist in the world of fiction), but of assumptions scientists made like *Neith* the satellite of Venus which was in reality an optical illusion. Regarding the second case, there is no entry like *Jean = French male name* or *Jeanne = French female name*. If we connect the previous point with this one, we can see that like the encyclopedia, the proper name dictionaries are interested in an identifiable referent, and not a name in the linguistic sense.

8. Another point that connects these dictionaries to the encyclopedia is the presence of various common nouns, but encyclopedically defined. It is still surprising to read articles *anarchisme* or *hyperréalisme* in a book that is supposed to define proper names.

The choices, such as what gets into the dictionary, are another proof of the encyclopedic ambitions of authors: it will be more difficult to see the name of singers and starlets appear in one of these books than that of a recognized architect or painter. The name of *Michael Jackson* for instance had to wait until millions and millions of his records had been sold to finally appear in French dictionaries. If this option is defensible in an encyclopedia (one can imagine that the architect or painter will leave a mark in history but the starlet that everyone talks about today will be forgotten in thirty years), it is less acceptable in a dictionary that aim to meet immediate needs.

The list could be longer – for example, no lexical relationships while some proper names have nicknames that approximate synonymy[72]. Even though, the few studies that have been done so far say there are many gaps in the linguistic treatment (Lecomte-Hilmy, 1989, Sarfati, 2000, Vaxelaire, 2005b).

Why doesn't exist a language dictionary of proper names? Because lexicographers have a theoretical *a priori* about names. Names are not presented in proper name dictionaries as linguistic elements but as carriers of a referent, real or fictional. Lexicographers do not seem to imagine that readers would like to learn how to pronounce *Maastricht*, they believe that everything they are interested in is to know the number of inhabitants of the city and its characteristics compared to other cities.

The study made by Mireille Elchacar (2011) shows, however, that there are indeed proper names in language dictionaries (eg = *l'Empereur = Napoléon* in the subentry *empereur*) because it seems impossible to describe the lexicon of a language while setting aside a significant part of it. I did further research with the name Napoléon and within the first edition of *Petit Larousse Illustré* (1905), there are 46 occurrences

[72] Though this is less and less popular, some French first names have quite different hypocoristic forms like *Françoise / Fanchon* and *Nathalie / Nanou* that cannot be guessed alone.

(some of them concerning Napoléon III). In a more recent edition (2009), there are still 24 occurrences. Some of them, come with no surprise (*grognard*, *conquête*), but others are startling (*train*, *cirque*). I also found another nickname (i.e., a proper name) of Napoléon with *l'Aiglon*.

2) I have already partly answered the second question at the beginning of this paper: proper names were initially rejected from language dictionaries because there are just too many, and lexicographers have validated this theoretical point of view explaining that proper names are not a part of the language, following the majority of linguists who see names as signifiers without a signified or elements that, like a pointing finger, have the sole function of designating a referent. The signifier without a signified is absurd from Saussure's point of view and we can understand texts that contain proper names of people, places or political parties though we do not know the referents. The proper name is therefore not an extra-linguistic tool but an assembly of morphemes. Although the argument about the quantity of proper names is admissible, we also know that dictionaries have never included the entire lexicon of a language, it would then be possible to choose the most culturally salient in the reference corpus.

What is even more annoying is that derivatives can undergo this ostracism that affects proper names. The *Petit Robert* gives lists of adjectives derived from proper names outside its nomenclature as if these adjectives are necessarily different from others.

3) As encyclopedias, a proper names dictionary is designed for native speakers, and speakers that know their language well enough to don't be interested by simple facts like gender or pronunciation. Furthermore, an encyclopedia is a sum of knowledge, not a tool, though in my mind the dictionary is just a tool to help for linguistic needs. Translators, all who learn a language, those who want to better understand the culture can be interested by such a tool.

I once read a book for writers that listed proper names by country, adding their gender[73]. When we have to deal with a language we do not know at all that kind of information. Take for example Korean: it is difficult to know if the person who writes is a man or a woman judging from his/her name. Writing that *Gaspard* is a male name and *Oceane* a female name is far from being unnecessary if we are dealing with non-French people.

The cultural dimension also involves differences within the same language. As far as French is concerned, a statement related to French political life will not always be easily understood in Quebec or francophone Africa. One can imagine the equivalent between England and the United States or Portugal and Brazil. Certain proper names are more salient in one country than another and it is more than likely that they do not convey the same semantic load.

[73] Sherrilyn Kenyon : *The Writer's Digest Character Naming Sourcebook*, Cincinnati, Writer’s Digest Books, 2005.

4. A language dictionary of proper names

For a language dictionary of proper names to be effective, the first point is to draw up the nomenclature. The choice of entries can only be left to the subjectivity of lexicographers[74], the criterion of the number of occurrences in the selected corpus must be taken into account, the aim is not to offer a piece of the scholarly culture of countries but to treat linguistic facts, and the name of a Hollywood star is more likely to be used in an antonomasic way than the name of a composer of twelve-tone music as talented as s/he is. It seems vital to establish a reference corpus (press, literature, audio recordings when their treatment will be facilitated) to get a clear view of the reality of a language lexicon.

The medium to be used is a crucial element as I have already mentioned the limitations imposed on the nomenclature by paper dictionaries, it is clear that these issues are no longer important if the dictionary appears in an electronic format, the number of entries is only limited by the work the contributors can accomplish. Beyond the issue of the number of entries, switching to electronic format offers the user new possibilities for consultation (eg the "interpretative approach" described by Piotrowski, 1996). The failure of the majority of current electronic dictionaries can be explained because they do not sufficiently exploit the technical capabilities that are available, most of the time these dictionaries are nothing more than a revised version of the paper dictionary. In other words, they have adopted a medium that provides new opportunities but retaining the old lexicographical practice rather than renew them.

One of the features offered by an electronic dictionary is to simplify the lexical links between words and their derivatives, especially as they are not always obvious for proper names (it is not easy to know that inhabitants of the *Hauts-de-Seine* are named *Alto-séquanais*). It is possible to go further with these links, displaying those between endonyms and exonyms, who are becoming increasingly problematic in the French press, for instance the name *Göteborg* is more and more often replaced by *Gothenburg*, the English form of the Swedish town.

It is a pious platitude to say that in modern lexicography we must propose a unified treatment of the entries. In the case of proper names, why should we only give the pronunciations that are supposed to be complex? A foreigner must learn for example that in French the final *-s* of *Paris* is silent, which is not necessarily the case in his native language or in the international English that he may speak, but that this letter is not silent in some proper names like *Duras*. A dictionary cannot afford to provide information in a dispersed form as do dictionaries of proper names. Although it seems obvious to classify *Le Caire* (Cairo) as a masculine name, it is almost as obvious that *père* is a masculine name and yet no dictionary of language fails to indicate this information.

We read very often in a language such as French that proper names are used without a determiner. If this is generally true for anthroponyms (there are dialectal variations where the use of the definite article before a name is accepted), this is not

[74] French lexicography's main idea was expressed by Rey-Debove (1971 : 31), the importance of a proper name in the sociocultural system outweighed its frequency, the professional activity of a person will influence the appearance of his or her name in the nomenclature.

the case for all types of names, it is essential to indicate whether a proper name needs or not to be preceded by a determiner. For example, when I was working in Cyprus, I noticed that most of the students called their country “la Chypre” on the model of “la France” and "la Turquie". However, this is not the case as this country name works without a determiner in French like *Israël*, *Andorre*, *Madagascar* and some others. As there is no clear rule[75], the lexicographer would not waste his or her time to give this information.

It is not necessary to detail all the categories that are required,. I~~i~~n brief, a good dictionary should contain information on pronunciation, gender, morphosyntax, derivation and lexicalization, which is not the case of French proper names dictionaries.

5. The problem of the definition

The most difficult issue is the definition. In classical theory, the proper name “désigne et ne signifie pas” (Rey, 2008: 73), which is a mistake. If I stand before you saying “I'm Peter” or “I'm Paul”, you'll be inclined to believe me. By contrast, if I announce that “I'm Sandra” or “I'm the Socialist Party”, you'll be more skeptical. Why? Because *Peter* and *Paul* contain the semantic feature /masculine/ that the other two examples do not have. Even if they would not have more than a single seme, they are therefore not semantically empty.

It is true that the meaning of proper names is generally very limited, which is problematic for a dictionary. If I simply define *Lisbon* “name of a city, capital of Portugal”, the public is accustomed to encyclopaedic proper names dictionaries may be embarrassed. One can also wonder whether a dictionary that contains only definitions such as the one I gave earlier of *Jean* (Chapter XXX) may be of interest to readers.

There is a way to enrich this strictly linguistic part by what we can call *cultural*. There are several definitions of what is the cultural perspective in lexicography. For Ronald Shusterman (Lecercle & Shusterman, 2002 : 119), we need to come out of the dichotomy *language / encyclopedia* to embrace both of them. In its *Dictionnaire culturel en langue française* (2005), Alain Rey opposed the cultural and the encyclopaedic because the last one “traite par exemple les noms-concepts « fruit » et « fleur » selon la science botanique, supposée universelle ; le premier décline les valeurs symboliques, littéraires, affectives... de ces deux entités selon les différentes civilisations. Les deux types de dictionnaires tentent d'aller au-delà de la langue, mais l'espace de l'encyclopédie est extralinguistique et donc en partie fictif, alors que celui du dictionnaire culturel est interlinguistique, restant dans le langage naturel” (2008 : 139). The educational specialist Robert Galisson also opposes *cultural* and *encyclopaedic* regretting that language dictionaries contradict their theoretical assumptions by leaving too much space to encyclopedia at the expense of culture. Galisson (1991: 122) gives the example of the French word *dragée*

[75] We are told that the size of the country is the main criterion but Madagascar is much larger country than Luxemburg, though in French this is “le Luxembourg” and “Ø Madagascar”.

(“sugared almond”) whose composition (“almonds coated in sugar”) is given without any reference to its role in baptisms. This encyclopaedic information is secondary for the student who is learning French, the cultural part (the fact that we offer these *dragées* at baptisms) is much more important in the context of learning a language.

I share Galisson's view, when we are describing the cultural part of lexicon, we are not as Rey writes beyond but at the heart of language, unless someone thinks that languages are just objects for the grammarian. When we are translating, these cultural issues generally arise, and that may be problematic. A good formal proficiency of a foreign language is not enough to be a good translator, and proper nouns are an inevitable part of culture.

As we have seen at the beginning of this text, the point is even more complicated when it comes to proper names because of the widely shared theoretical assumptions. For instance, Alain Rey thinks that in the case of proper names there is an “identité du type langue et de l'occurrence” (in Elchacar, 2011: 38). I have proven in previous works (Vaxelaire, 2005a & 2008) that semantically there is precisely a large difference between type and occurrence. I use the classical distinction in French linguistics between *meaning* and *sense*, the first one referring in broad outline to the type (what we can read in a dictionary) and the second to the occurrence (what we can read in a text). What is typical of proper names is that they do not have a lot of meaning but that they can gain much more sense in various contexts. It is when this sense crystallizes that we can create an antonomasia (in other words, this semantic content we call sense becomes a part of the meaning and therefore the proper name becomes a common noun). To give an example, it is difficult to understand the metaphor that makes the public prosecutor of Nice when he says: “Je n’ai pas l’impression d’être procureur de la République à Chicago” (*Le Monde*, 23/12 / 09) with a proper names dictionary. In this type of work, we only learn that Chicago is a “city in the United States, active port, home to modern architecture, etc.”. A request with the name *Chicago* in the CD-ROM version of *Petit Robert* leads us to the entry *gangster* “Membre d'un gang. ➙ **bandit.** *Les gangsters de* Chicago *dans les années trente. Film de gangsters.*” In the online version of the *Trésor de la langue française* we find traces of *Chicago* in a *ville* subentry: “*Ville carrefour de qqc.* Ville point central de quelque chose. *Car Marseille est aujourd'hui la ville-carrefour du trafic de la drogue. C'est un* Chicago*, dans un autre genre*”. We can understand with these examples that in French culture the name *Chicago* is associated with concepts such as gangsterism and drug trafficking. An encyclopedia could give us the murder rates or real statistics about drug trafficking in Chicago. This is not what is interesting here because the public prosecutor is not speaking about the real Amercian town but about a French concept whose name is *Chicago*, and this concept is about everything the police and justice is fighting for.

To write an article where these facts are explicitly written however brings us to a problem that dictionaries such as *Merriam-Webster* with the second meaning of *Philippine* in 2005 and the *Oxford English Dictionary* in 2007 with the neologism *McJob* (a low-paying job with no hope of progression) were confronted with. McDonald's executives had asked the name to be withdrawn on the grounds that the jobs the company offered were better than what was described. Coming back to my example, it is conceivable that the mayor of Chicago would dislike a definition associating the name of his city and gangsterism and ask it to be removed because this equation refers to the 1930's. It is probably difficult to distinguish the real city and the concept without dropping the naive notion that language is a strict reflection of reality.

The problem with names of persons is identical or even stronger. If everyone accepts that the name *Hitler* is a kind of synonym of "absolute evil", it will be much more difficult with names of living persons. Thus, if a journalist from *Le Canard Enchaîné* (15/09/99) can write: "ce n'était plus du Bedos, mais du Delon!", this is because Alain Delon is well known for his self-centeredness (when someone speaks about himself at the third person, it is often related to a Delon kind of act), but it seems impossible during the lifetime of the actor to associate his name and *egocentricity* in a dictionary. Death is maybe not the only barrier. In the Norwegian TV series *Dag*, the main character announces: "Yesterday I was Gandhi, today I am Bobby Fischer!" (s03e04, broadcast in 2013). If it is not difficult to describe *Gandhi* as a symbol of love thy neighbor, it is more complicated to write that Fischer was known for being temperamental and hateful towards Jews. Any negative statement (and there are a lot in antonomasias) can create difficulties. There is a gap between the point of view of language, where I stand, and the reality of these people and places, but I am not sure if a clear warning at the beginning of the dictionary would be enough.

6. Conclusion

The need for a language dictionary of proper names is clear from the point of view of anyone who wants to learn a language or to increase his/her knowledge of the culture that this language conveys. An advertising executive would say there are a lot of things you do not know you need them until you have them and this dictionary is one of these things.

The main idea behind this project is that the interest of a proper name does not lie only in its reference, it can convey cultural elements, the proof is given by the creation of antonomasias. But, because of theoretical prejudices, a lot of linguistic elements are removed from dictionaries and the proper names dictionaries have not been designed to fulfill this need.

There are classical linguistic elements to be given like gender, but also more cultural elements that show why the name *Mengele* for example has been used very often though Mengele was not a main character in Nazi hierarchy. Just as I said with the "large country criterion" in French, there is not necessarily an appropriateness between reality and what a language actually uses. The choice of the reference corpus is essential to enrich the content of articles and prove that the definitions are objective (though the corpus will not necessary be objective).

Though it is theoretically feasible and practically useful, a language dictionary of proper names would anyway meet some barriers because dealing with proper names is for some people not exactly the same as dealing with other kinds of words. In ancient mythologies, the name is a part of the person and it seems that this issue is not clearly resolved. If I write that Alain Delon is the symbol of egocentricity, this is not for personal reason, but only because I have dozens of quotations that say it. This point is probably the biggest hurdle to clear.

References

Anderson, J.M. (2007). *The Grammar of names*. Oxford: Oxford University Press.

Galisson, R. (1991). *De la langue à la culture par les mots*,. Paris: CLE international.

Galisson, R.; André, J.-C.(1998). *Dictionnaire de noms de marques courants — Essai de lexiculture ordinaire*. Paris: Didier Érudition.

Lecercle, J.-J.; Shusterman, R. (2002). *L'emprise des signes — Débat sur l'expérience littéraire*. Paris: Le Seuil.

Lecomte-Hilmy, A. (1989). "Du statut linguistique des noms propres dans cinq dictionnaires français", *Cahiers de lexicologie.* 54, n° 1. 8-32.

Piotrowski, D. (1996). "Opérations hypertextuelles et formes lexicographiques", *in* Piotrowski, D. (ed.). *Lexicographie et informatique : autour de l'informatisation du* TLF. Paris: Didier-Erudition. 319-336.

Rey-Debove, J. (1971). *Étude linguistique et sémiotique des dictionnaires français contemporains*. The Hague-Paris: Mouton.

Sarfati, G.-E. (2000). "Le statut lexicographique du nom propre : remarques méthodologiques et linguistiques". *Mots*. 63. 105-124.

Vaxelaire, J.-L. (2005a). *Les noms propres — Une analyse lexicologique et historique*,. Paris: Honoré Champion.

Vaxelaire, J.-L. (2005b). "Nom propre et lexicographie française". *Corela*. Numéro thématique. http://corela.edel.univ-poitiers.fr/index.php?id=1239

Vaxelaire, J.-L. (2008). "Étymologie, signification et sens". In Durand, J.; Habert, B.; Laks B. (eds.). *Actes du Congrès Mondial de Linguistique Française* – CMLF'08. 2187-2199.

Zabeeh, F. (1968). *What is in a name?: An inquiry into the semantics and pragmatics of proper names*. [on-line] The Hague: Martinus Nijhoff.

Petit Larousse Illustré. Paris: Larousse, 2009.

Petit Larousse Illustré, 1905. [on-line] http://dictionnaire1905.u-cergy.fr/ Access date: 13/05/2013]

Le Petit Robert. Paris: Le Robert. 2011.

Le Petit Robert des noms propres. Paris: Le Robert. 2011.

A dictionary of abbreviations in linguistics: Towards a bilingual, specialized, single-field, explanatory dictionary

Ivo Fabijanić

1. Introduction

Behind the process of looking for a full form of abbreviations lies a whole range of problems worth considering and solving. The aim of this paper is to provide some theoretical and methodological models in preparing both a macro- and micro-structural framework for compilation of a future English-Croatian, specialized and explanatory dictionary of abbreviations in linguistics, one which has never been developed before. The dictionary would cover the core areas of linguistics (phonology, morphology, syntax, semantics and pragmatics) and its interdisciplinary areas as well (e.g. sociolinguistics, psycholinguistics, neurolinguistics, ethnolinguistics, computational linguistics, biolinguistics, etc.).

When trying to find a full form of an abbreviated word, majority of users usually take a general, bilingual, and explanatory dictionary, hoping to find satisfactory information about its "meaning" or *expansion* (for *expansion* cf. Kompara 2012).

The practice of providing expansions for abbreviations solely, in our opinion, goes back to the beginnings of lexicographic practice when abbreviations were (and still are) mostly used as an economical tool in describing and explaining a wider context a headword might be understood within. In most specialized and non-specialized on-line dictionaries of abbreviations, such as *Abbreviations.com*[76], attempting to be "[...] the world's largest and most comprehensive directory and search engine for acronyms, abbreviations and initialisms on the Internet [...]", abbreviations are organized in different categories (e.g. academic and science, business and finance, community, computing, etc.). Entries in it are explained only by their expansions and the categories they fall into. *Abbreviations.com* is a monolingual (English to English), specialized, multi-field, explanatory dictionary.

Abbreviation Finder – Dictionary.com[77] "[...] provides simple descriptions of acronyms and shorthand from many areas of life, including computer science, sports, social media, conversation and industry. Entries include explanations of the

[76] http://www.abbreviations.com/

[77] http://dictionary.reference.com/abbreviations/

context of abbreviations as well as the direct meaning [...]". According to its typological features, *Abbreviation Finder – Dictionary.com* is also monolingual, specialized, multi-field, explanatory dictionary. *Acronym Search*[78] has the aim of providing "[...] the best free resource possible [with] over 50,000 acronyms and abbreviations in many categories [like] chat, computer, military, finance, accounting, airports, sports [...] and more". As it is a software system (i.e. web search engine), designed to search for information on the WWW, search results are generally presented in a line of results. This dictionary can also be classified as a monolingual, specialized, multi-field, explanatory dictionary.

All Acronyms Dictionary[79] is "[...] a free human edited website with more than 1,000,000 [...] acronyms and abbreviations [with] the main purpose to have a convenient lookup web tool for those who need to quickly find an acronym definition or need to abbreviate a particular word or phrase". In *All Acronyms Dictionary*, acronyms, initalisms, alphatbetisms and other abbreviations are described only by their expansions, which means that it can as well be classified as a monolingual, specialized, multi-field, explanatory dictionary.

The *OED*[80] provides a apecialized list of "[...] the most common abbreviations used in the *OED*", which "[...] are only abbreviated in certain contexts, esp. when used as a subject label or in a work title".

A somewhat different approach to abbreviations is found in *Acronyms and Abbreviations*[81] powered by the Acronym Finder, "[...] the web's most comprehensive dictionary of acronyms, abbreviations and initialisms [which] allows users to decipher acronyms from a database of over 600,000 entries covering [...]" different categories. In its description we found that it "[...] exists purely to unravel the bewildering range of acronyms that impact daily life [...]". The difference in approach is understood by the fact that it deals not only with the expansions, but also with various more comprehensive information about the meaning, sources, references, and some related links.

It goes without saying that this typological classification list of on-line dictionaries is far from being fully comprehensive; its only intention is to illustrate some references, their basic typological characteristics, and the approaches to abbreviations utilized in them.

1.1. Classification of abbreviations

Abbreviations are not anymore seen as marginal types of word formation in English. However, the process of their creation has long been neglected as they do not belong to regular morphemic processes (Fandrych 2008a). The usage of abbreviations is not reserved for the formal register; instead, they are equally used in both highly specialized technical jargon and in informal types of Internet or mobile-phone communication, denoted as *Netspeak* and *Textspeak* (Fandrych 2007; Bieswanger 2007; Crystal 2004).

[78] http://www.acronymsearch.com/

[79] http://www.allacronyms.com/

[80] http://public.oed.com/how-to-use-the-oed/abbreviations/

[81] http://acronyms.thefreedictionary.com/

One of the distinctive traits of this word-formation group is the lack of consistent categorisation and typology, as well as fixed boundaries between the respective types of abbreviations. López Rúa (2004) states that there are numerous disagreements over what an acronym in general is and what proper abbreviations, clippings and blends are. Another discrepancy within this field is the lack of consistent usage of a certain term for certain abbreviation type. Among the authors cited in this work, there is a great disagreement over what the terms initialism, alphabetism and acronym represent, and what they should be used for (cf. Plag 2003; Algeo 1991; Jackson and Zé Amvela 2005; Fandrych 2008a, 2008b; Stockwell and Minkova 2001; Harley 2006; López Rúa 2006; Cannon 1989; Crystal 1995). Similar discrepancies regarding the differentiation between clippings, blends and clipped compounds may be seen when comparing Fandrych (2008), Plag (2003), and Algeo (1991).

The classification of abbreviations used here largely relies on López Rúa's (2006) work. *Abbreviations* are divided into *simple* and *complex abbreviations*, with the former encompassing only *proper abbreviations*, which appear in the written medium only, and the latter, including blends, clippings and initialisms further divided into alphabetisms and acronyms. Complex abbreviations appear in both written and spoken medium. According to López Rúa, an *initialism* is "*[...]* the result of selecting the initial letter, or occasionally the first two letters, of the orthographic words in a phrase and combining them to form a new sequence" (2006: 677). The pronunciation of initialisms depends on various linguistic factors, but there are two main modes of their pronunciation – as a word, i.e. *prototypical acronyms*, and as a series of letter names, i.e. *prototypical alphabetisms* (Ibid.). There are some abbreviations which can be pronounced and graphically presented both ways (*VAT/Vat* < 'Value Added Tax'), and some hybrid abbreviations that are pronounced as a combination of the two ways (*CD-ROM* < 'compact disc read-only memory'). *Clippings* are abbreviations created through a "*[...]* process by which a word-form of usually three or more syllables is shortened without a change in meaning or functions" (Ibid., 676). Although clippings express informal connotations and familiarity (Plag 2003) or even serve as euphemisms (Fandrych 2008b), which ultimately leads to different stylistic properties of the word, the meaning more or less remains intact. As far as *blends* are concerned, they are created by "*[...]* joining two or more word-forms through simple concatenation or overlap and then by shortening at least one of them" (López Rúa 2006: 677).

We find this taxonomy appropriate because it clearly distinguishes certain abbreviations types. This in particular refers to the terms for initialisms, alphabetisms and acronyms, which are often used interchangeably or wrongly dubbed as abbreviations or shortenings (e.g. in Plag 2003, Jackson and Zé Amvela 2005). The term initialism denotes an abbreviation created through the usage of initial letters, which applies to both alphabetisms and acronyms. The term alphabetism denotes an abbreviation pronounced as a series of letters of the alphabet, i.e. letter by letter, while the term acronym, coined in 1943, has been generally accepted to denote abbreviations pronounced as whole words.

2. Aims and objectives

One of the main aims of this research is to propose an entry structure of the future dictionary which would treat abbreviations in a more appropriate way. By appropriateness, for one thing, we mean the way(s) an entry is organized. Majority of abbreviation dictionaries are unidirectional and explanatory and do not deal with other information a user would find interesting and useful, such as information on diachronic, diatopic, diamesic, diastratic and diaphasic variations. Here it is necessary to say that not all variations are of the utmost importance for a dictionary of abbreviations, but some, like diachronic, diamesic and probably diatopic variations, would prove to be valuable in making its scope more comprehensive and informative.

We shall now briefly consider the main objectives in compiling the dictionary, i.e. its main characteristics. Due to the fact that the spelling of the headwords would, for the most part, consist of capital letters, we believe that the alphabetical arrangement of lemmata would be the most suitable form of presentation for this dictionary. The term bilingual implies a dictionary that would explain abbreviations in both English and Croatian, which means it would be bidirectional: firstly explaining an English abbreviation in English and secondly, providing a translation of the expansion in Croatian, as well as the translation of information on the diamesic level. The feature of being specialized is understood within the frame of a scientific and technical dictionary which would firstly serve as a lexicographic tool for professionals in the field, synchronically representing the specialized lexicon as it exists. Secondly, it would also be aimed at both wide, general, non-specialized and specialized audience. The singularity of the subject-field indicates the purpose of the dictionary to be informative and encyclopeadic in content for that particular scientific field. The specialized corpus of the dictionary identifies the domain-specific collection of abbreviations in linguistics from specialized publications in the field. As for the final explanatory feature, the entries would present the following data: spelling, pronunciation, typology of abbreviations, full form of source phrases (expansions), orthographic and morphological characteristics of both abbreviations and their expansions, and the origin of headwords.

2.1. Two-sidedness of the lemma

Contrary to previous lexicographic practice of dealing with the micro-structure in a dictionary of abbreviations, we feel that headwords in such a dictionary should be considered differently from those in a general-purpose dictionary, primarily because of the specific formational processes inherent to abbreviations. In other words, their headword form should be seen as two-sided, which means that if we are to understand not only the meaning of an abbreviated form, but its final orthographic form as well, we have to take into account both sides – the abbreviated form and its source phrase or expansion (e.g. *ACL* – 'Association for Computational Linguistics'). The left side of the lemma represents its final abbreviated form while the right side its expansion. The right side of the lemma is not merely its meaning, but is also its morphemic structure (although its formation is sometimes considered to belong to non-morphemic or non-morphematic word-formation processes, cf. Fandrych

2004). Having taken into consideration its two-sidedness and various characteristics of forms found in the right side of the lemma, we shall see that this would necessarily lead to a unique designation of their distinctivness. For the purpose of their differentiation, a system of exclusive classification and subclassification of abbreviations proposed in earlier research (cf. Fabijanić, Malenica 2013) is improved in this research.

2.2. Narrower and broader sense of the lemmata

Abbreviations in the linguistic terminology will be classified according to two main criteria – *narrower* and *broader sense*. The narrower sense of their creation refers to those formed by using initial letters of each element in the expansion (mainly alphabetisms), and pronounced either by individual names of letters or as a word. The broader sense implies the ways and processes of formation, more or less different from the orthographic norms (mainly hybrid forms, acronyms, blends and clippings featuring some orthographic changes), in consequence of which, one or more initials are used for various smaller elements of the expansion (smaller than words, yet bigger than initials). Due to this, initials for graphemes, compounds, affixes, grammatical and lexical words found in the final form of an abbreviation, as well as different orthographic changes, such as ellipsis, conversion, metathesis, addition, etc., will be analysed and (sub-)classified. By considering their different expansion combinations and comparing those different combinations with the final abbreviated forms, we will be able to understand the real nature of their (non-) morphemic structure and the possibility of differentiating the initialisms according to various criteria.

2.3. The multi-level approach

A tentative solution for the lexicographic presentation of abbreviations is based on Fandrych's (2008a) theoretical framework, i.e. the multi-level approach for the analysis of submorphemic word-formation processes. The multi-level approach is comprised of three main aspects: 1) *Structure and Modes of Production*, i.e. the structural aspects and word-formation potential, word class, medium and origin; 2) *Cognitive Aspects*, i.e. semantic, semiotic and motivational aspects, lexicalisation and institutionalisation, and 3) *Functional Aspects*, i.e. stylistic and sociolinguistic aspects, pragmatic and text-linguistic aspects. We believe that the application of this interdisciplinary approach will give a fuller and more transparent picture of various orthographic, morphological, semantic, stylistic and functional processes involved in the production and uses of abbreviations.

As for the initial step in compiling this specialized dictionary, the first aspect of the multi-level approach, i.e. the structure and modes of production, will be supplemented by our novel aspects within the structure and modes of production, i.e. on both orthographic and morphological level. It has been observed that "[t]here is evidence for a remarkable **word-formation potential**, in the form of multiple formations...[and] there are also a number of **non-morphematic** word-formation processes which are multiple formations." (Fandrych 2008a: 74-75; emphasis in original). The word-formation potential in this work is understood exactly as the reapplication of the bi-aspectual criterion used in the creation of abbreviations, i.e. the criteria of narrower and broader sense. If one of the criteria is being used more frequently than the other, most frequently narrower over broader sense, it obviously represents the greater word-formation potential of abbreviations.

2.4. The entry and its elements

Firstly, an entry would consist of a headword and all its possible variants. Together with the data about its origin and the medium the abbreviation was noted in, the information on the word class (where applicable) will be also provided. Basic form would be succeeded by the information on pronunciation and the type of word-formation. A whole set of novel descriptors will be used when dealing with the type of word-formation (i.e. the modes of production). The descriptors will inform the user about the originality and regularity of the final word-forms, i.e. whether they were formed according to the orthographic norms of abbreviations or with some evident exceptions to the norms.

For the purpose of their differentiation we proposed a system of exclusive classification and subclassification of abbreviations (Fabijanić, Malenica 2013). Miscellaneous realisations of abbreviations are generally diversified into two main groups: those considered as realised in narrower sense and those in broader one. Abbreviations in narrower sense are exclusively explained with an *LLL* descriptor for initials used in their formation (a characteristic three-letter descriptor was chosen due to the most frequent number of graphemes found in most initialisms). Abbreviations in broader sense are represented with a whole set of additional different letters or initials (written either in capital or small letters) added to a three-letter descriptor: e.g. *l* for initials made of small letters, *P* refers to initial affixes (mostly prefixes), *N* stands for a numeral, *S* refers to a syllable, and *W* stands for a word used in a hybrid form of an abbreviation. Orthographic changes, evident by comparing initialisms (alphabetisms or acronyms) with their expansions, are explained by other descriptors: *E* for ellipsis, *C* – conversion, *M* – metathesis, and *A* for addition of a word or a diacritic sign not normally found in expansions. Comprehension and consequently classification of abbreviations depends on the degree of their (non-)coordination with the common orthographic norms.

The following examples of different abbreviation forms (an alphabetism, an acronym, a clipping, a blend and a hybrid form) present the micro-structure of a future dictionary entry, i.e. the structure in which all elements of the first aspect (cf. § 2.3. *The multi-level approach*) were taken into consideration.

ADS [*ˈeɪˈdiːˈes*] **T**: *alph.* | **E**: *American Dialect Society*/Američko dijalektalno društvo | **M**: written/ pisani, spoken/govorni | **D**: *LLL* | **O**: *ADS* was founded in 1889 with the intention of creating a dictionary of American dialects. (ELL)/Američko dijalektalno društvo osnovano 1889. s namjerom stvaranja rječnika američkih narječja.

ACE [*eɪs*] **T**: *acr.* | **E**: 1) *Automatic Content Extraction*/Automatsko ekstrahiranje sadržaja, 2) *Australian Corpus of English*/Korpus australskoga engleskog | **M**: written/pisani, spoken/govorni | **D**: 1) *LLL*, 2) *LLL E = prep.*; **O**: 1) The *ACE* program is a successor to → *MUC* that has been running since a pilot study in 1999. (ELL)/*ACE* program provodi se još od pilot-projekta iz 1999., a naslijedio je MUC., 2) The corpus of

Australian English compiled at Macquarie University using texts published in 1986. (HEL) / Korpus australskoga engleskog sačinjen na Macquarie sveučilištu iz tekstova objavljenih 1986.

AUX, Aux, aux [-] **T**: *clip.*|**E**: *Auxiliary*/pomoćni|**M**: written/pisani|**D**: *Lll*|**O**: *Lat.* auxiliaris - 'giving aid' (RDLL) /*lat.* pomoćni - 'koji pomaže'.

AFRILEX [ˈ*æfrɪˌleks*] **T**: *blend*|**E**: *AFRIcan association for LEXicography*/Afričko leksikografsko udruženje|**M**: written/pisani, spoken/govorni|**D**: *LSF E= noun, prep.*|**O**: *AFRILEX* was founded in 1995 and strives to promote all aspects of lexicography on the African continent. (ELL) / Organizacija osnovana 1995. u svrhu promocije svih aspekata leksikografije na afričkom kontinentu.

ALGOL [ˈ*ælgɒl*] **T**: *hybr. (syll+s.abb.)*|**E**: *ALGOrithmic Language*/algoritamski jezik|**M**: written/ pisani, spoken/govorni|**D**: *SSL*|**O**: Programming computer language appeared in 1958. (RDLL) / Programski računalni jezik nastao 1958. godine.

Each entry consists of seven structural elements, five of which are represented by abbreviations (simple abbreviations): *T*, *E*, *M*, *D*, and *O*. The entry starts with a headword and its variant forms, all written with either small or capital bold letters, and is followed by its pronunciation with the use of IPA symbols. The first abbreviation (*T*) stands for the type of abbreviation explained and described, and relies on our subclassification of abbreviated formations (cf. § 2.5., 3.). *E* stands for the expansion elements that abbreviations were formed from. *M* represents the medium a headword was used in – written, spoken or both media. *D* represents a previously determined set of descriptors used in the classification of abbreviations according to the expansion elements compared with the resultant final form. *O*, the last user abbreviation, refers to the origin of the headword, i.e. to the information about the date of its emergence, the author/-s who devised it or the language it was borrowed or adapted from. The information given in the elements about the medium and the origin is also provided in Croatian.

2.5. Previous research on orthographic features of abbreviations, morphological features of expansions, and the correlation with the recent study

The previous research of abbreviations in English medical terminology (Ibid.) has justified the use of the mentioned classification. Most of the alphabetisms were formed according to the criterion in narrower sense and formed of initials written with capital letters which were taken from every single element found in the expansion. Acronyms were fewer and the ratio of those formed in narrower and broader sense was not as regular as for alphabetisms, i.e. in favour of those formed in narrower sense. The ratio for alphabetisms was 124 : 53 (i.e. they are 2,3 times more frequent) and for acronyms – 10 : 11, which means that they are almost equally represented. As far as the ratio of hybrid forms is concerned, the irregularity is even more evident – 2 : 9 in favour of broader-sense formations (i.e. those formed by narrower-sense criterion are 0,2 times less frequent).

The research of abbreviations in linguistics has presented us the following statistics: the ratio for alphabetisms is 159 : 109, in favour of narrower sense (they are 1,4 times more frequent than those formed in broader sense), for acronyms – 23 : 31 (0,7 times lesser frequency of those formed in narrower sense), for simple abbreviations the ratio is 19 : 1 for narrower-sense formations, and for clippings – 15 : 6 (i.e. they are 2,5 more frequent than the broader-sense formations). Blending has only been attested through the examples of broader-sense formations. A similar ratio for hybrid formations, in which broader-sense abbreviations are more frequent than narrower-sense ones, is also evident in this research.

3. The corpus of abbreviations in linguistics and the methods of analysis

The abbreviations analysed in this article were taken from different dictionaries of general linguistics and dictionaries of various linguistic disciplines (e.g. phonetics and phonology, lexicography, etc; cf. Sources). The corpus comprises 441 abbreviations, all belonging either to the category of simple or complex abbreviations. There are 268 alphabetisms, 65 acronyms, 5 blends, 19 clippings, 22 simple abbreviations and 62 hybrid forms. The analysis examines the orthographic features and morphological elements used in the formation of abbreviations, such as initials, prefixes, numerals, syllables, splinters and lexical words, as well as various types of orthographic changes (e.g. ellipsis, addition, metathesis) that the abbreviations have undergone, when compared with their expansions.

Our corpus provides the additional information about the analysis for each abbreviation type in the following order: abbreviation, expansion, descriptor of abbreviation form, abbreviation type, source, orthographic changes inherent to an abbreviation and the details of the corresponding orthographic change.

The analysis of abbreviations in linguistics has proven the possibility of classification and further subclassification of abbreviations according to two criteria – narrower and broader sense, applied not only to initialisms (cf. Fabijanić, Malenica 2013) but to simple abbreviations too, as well as to other groups of complex abbreviations like blends and clippings.

If our previous classification of abbreviations in medical terminology is compared with this one in linguistics, we may see that many of the descriptors have been attested in this work. Some new descriptors have also been applied to the analysed abbreviations, mostly because other categories of complex abbreviations were taken into consideration. Our classification and subclassification of abbreviations in narrower and broader sense is shown in the following table, together with the descriptors presenting possible realizations of the analysed types of abbreviations. Numerals in brackets show the total for a corresponding type, while descriptors in bold refer to those not attested in the previous work.

Alphabetisms		Acronyms		S. abbreviations		Blends	Clippings
n. sense	b. sense	n. sense	b. sense	n. sense	b. sense	b. sense	b. sense
LLL(159)	LLL E (85)	LLL (23)	LLL E (24)	LLL (21)	LLL E (1)		
	LLL M (2)						
	PLL (3)						
	PLL E (4)		PLL E (3)				
	P-LL E,A(1)						
	LlL (2)						
							Lll (3)
							lll E (1)
	L-LL (7)						
	L-LL E (1)						
	L-LL E,A (1)						
	L/LL (1)						
	NLL C,M (2)						
			SLL (2)				SLL (12)
							SLL E (1)
						SSL (1)	SSL (1)
						SFL E (3)	
							FLL (1)
			FLL E (2)				
						FFL (1)	
						FFL E (1)	

A new descriptor for the group of alphabetisms is *P-LL*, describing those made of initials, one of which abbreviates a prefix, and a hyphen. In the group of acronyms and in the group of some clippings in broader sense, an *FLL* descriptor is used, which means that they were made of initials from each element of the source phrase and of initials from a splinter extracted from one element only. For the sake of transparency of symbols used in descriptors, an *F* symbol stands for a splinter, i.e. a fracture of a word, due to the fact that an *S* has already been reserved for a syllable present in an abbreviation. Blends are represented only by a subgroup of examples classified according to the criterion of broader sense and there are two new descriptors – *SFL* and *FFL*. The former represents a blend formed of a syllable and a splinter, while the latter represents the one formed of at least two splinters. In the group of clippings in broader sense, there are *Lll* and *lll*. The former represents a clipping made of one initial and of at least two small-letter initials, while the latter represents a clipping made of small letters only.

As far as hybrid forms are concerned, there are eighteen types of hybrid forms subdivided according to miscellaneous combinations and elements used in their formation in ten groups. The following table presents hybrid formations along with short descriptions of combinations of their elements. The first is the group of hybrids formed of alphabetisms and acronyms. The next is the group of hybrids formed of alphabetisms and a lexical word. The third and fourth columns present the descriptors for combinations of alphabetisms and simple abbreviations, i.e. alphabetisms and splinters, alphabetisms, splinters and whole lexical words, respectively. A combination of an acronym and a lexical word is classified in the next column. The sixth group is characterized by the combinations of two simple abbreviations, a simple abbreviation and either a lexical word, clipping or an acronym. In the seventh group, the combinations of clippings with alphabetisms and lexical words are found. The combinations of splinters, simple abbreviations, lexical words and graphemes are subdivided in the eighth group, while the ninth and tenth groups of hybrids contain the combinations of syllables, splinters, simple abbreviations and lexical words, and the combination of the phonetic respelling and an alphabetism, respectively.

There are various other new descriptors in the group of hybrid formations that were not noted in the previous work, such as *LLW* (abbreviations formed of initials and a lexical word), *PLW* (formed of initials for both lexical and grammatical words together with a whole lexical word), *L/LW* (formed of initials for every element and a slant line with the addition of a lexical word), *FLW* (formed of a splinter, initials and a lexical word), *F-LW* (formed of a splinter, alphabetism, a lexical word and the addition of a hyphen), *S-LW* (formed of a syllable, hyphen, initials and a lexical word) and *SFL* (formed of a syllable, splinter and initials), *WLL* (formed of a combining form and an acronym), *W-LL* (combined by an affix and an alphabetism, together with a hyphen).

Hybrids (alph+acr)	Hybrids (alph+w)	Hybrids (alph+s.abb)	Hybrids (alph+spl) (alph+spl+w)		Hybrids (acr+w)	Hybrids (s.abb+w) (s.abb+clpp) (s.abb+s.abb) (s.abb+acr)		Hybrids (clpp+alph) (clpp+w)	Hybrids (spl+s.abb) (spl+s.abb+w) (spl+grphm)	Hybrids (syll+s.abb) (syll+spl+s.abb)	Hybrids (phntc rspll+acr)
b. sense	b. sense	b. sense	n.sense	b. sense	b. sense	n.sense	b. sense	b. sense	b. sense	b. sense	n.sense
LLL E (4)			LLL (1)	LLL E (2)		LLL (2)	LLL E (1)				LLL (1)
LLL E, A (1)											
		LlL (1)									
								LlL E (1)			
		LLl (3)		**LLl** (1)			**LLl** (3)				
		LLl E (1)									
	LLW (8)							**LLW** (1)			
	LLW E (8)				**LLW** E (1)						
	PLW (1)										
					WLL (1)						
	W-LL (1)										
	L-LW A (3)						L-LW A (4)				
	L/LW (1)										
									FLL (1)		
				FLL E (2)					**FFL** E (1)		
									FLW (1)		
				FLW A (1)							
									F-LW A (1)		
										SSL (1)	
										SFL E (3)	

3.1. Alphabetisms in narrower sense

Classification of alphabetisms in narrower sense begins with those formed of initials written with capital letters and taken from every single element in the expansion (either lexical or grammatical word), i.e. one initial represents one lexeme in the expansion. As with the alphabetisms in medical terminology (cf. Fabijanić, Malenica 2013), the class of alphabetisms in linguistics is also the most abundant one. The descriptor representing abbreviations in narrower sense is *LLL*[82].

TYPE: **LLL**	
ALPHABETISM	EXPANSION
AG	Applicational Grammar
ACD	American Collegiate Dictionary
CCDA	Critical Classroom Discourse Analysis
RHWCD	Random House Webster's College Dictionary
IJOAL	Indian Journal Of Applied Linguistics

3.2. Alphabetisms in broader sense

The most distinguishing features of alphabetisms in broader sense are understood through a set of various orthographic and morphological changes. The most notable one is the ellipsis of either grammatical or lexical words in the expansion (or sometimes both).

3.2.1. Ellipsis in alphabetisms

The ellipsis can be of three kinds: either a lexical or grammatical word is omitted from the expansion, or both, and consequently, no initial for the omitted word has been used in an alphabetism. The examples of words being omitted from the expansion in this work are written in italics. In the corpus of analysed abbreviations there are examples of omitted articles (*a*, *the*), prepositions (*of*, *for*, *as*), conjunctions (*and*), nouns (*language*, *approach*, etc.), as well as punctuation marks (hyphens, slant lines and commas).

TYPE: **LLL E**	
ALPHABETISM	EXPANSION
VSO	Verb, Subject, Object
ESL	English *as a* Second Language
CDQ	Context-Dependent Quantifier *approach*
ALLP	Australian Language *and* Literacy Policy
ACTFL	American Council *on the* Teaching *of* Foreign Language

[82] All initials for expansion elements in this study are presented with capital letters.

3.2.2. Affixes in alphabetisms

Alphabetisms with a combination of initials for lexical morphemes and affixes are generally of two kinds: the first with only one initial for an affix, and the second with more than one initial for affixes in the expansion. In this research, only those with one affix (*PLL*), ellipsis and addition (*PLL E*, *A*), and with a hyphen incorporated, have been analysed (*P-LL*). There are also some examples of alphabetisms with initials taken from the combining forms which, in formation of words, have a very similar function. The following prefixes and combining forms have been attested: *hyper-*, *multi-*, *non-*, *morpho-*, *socio-*, *long-* and *well-*. Several examples show the presence of ellipsis, i.e. the omission of a hyphen or a preposition.

TYPES: **PLL**, **PLL E**	
ALPHABETISM	EXPANSION
MP	MorphoPhonemic
SCT	SocioCultural Theory
HTML	HyperText Mark-up Language
LTM	Long-Term Memory
SSCS	Society *of* Socio-Cultural Studies

3.2.3. Small letters in alphabetisms

Small letters in alphabetisms are very rare and they make up only two examples. These are *CmC* (< 'Computer mediated Communication') and *PoL* (< 'Parsimony of Levels'). In both of them the second element of expansions was abbreviated to a small letter. In the former case, that element is a lexical word and in the latter one it is a grammatical word. A descriptor for them is *LlL*, with a small letter representing a lexeme abbreviated in such a manner.

3.2.4. Hyphens and slant lines in alphabetisms

Hyphens and slant lines are also rare and they can be divided into two subgroups according to their formation. In the first one, there are those which retained a hyphen or a slant line in the same position as it was used in the expansion, and in the second, there are those in which hyphens and slant lines in expansions were omitted or were added to them. For the latter reason the abbreviation *A* for *addition* was introduced.

TYPES: **L-LL; L-LL E; L-LL E, A; L/LL**	
ALPHABETISM	EXPANSION
C-F	Context-Free
T-C	Topic-Comment
R-A	Referentially Autonomous *expression*
PS-SS	Phonological Stucture - Syntactic Structure
CBSE-ELT	Central Board *of* Secondary Education - English Language Teaching

3.2.5. Numerals in alphabetisms

Another infrequent modality of formation justifies the necessity for a more detailed orthographic analysis. Namely, there are a few alphabetisms with numerals (*NLL*) in which additional change, i.e. *metathesis* (*M*) was noticed. In alphabetisms *L1* (< 'First Language') and *L2* (< 'Second Language') words were converted (*C*) into numerals, and the order of expansion elements was modified, i.e. reversed, in resultant alphabetisms.

3.3. Acronyms in narrower sense

The acronyms in narrower sense are also formed of initials taken from every word in the expansion. Interestingly, the ratio of those formed in narrower sense to those in broader sense is not in favour of the narower-sense acronyms, which was the case in the research of medical abbreviations, too. This is explained by the fact that acronyms (sometimes called *acrophones* or *phonetic acronyms*) are pronounced as words, not as a serious of letters, and that not all letters from expansion elements are used in their formation. On the basis of the bi-aspectual concept, acronyms are classified into two categories – narrower and broader sense.

TYPE: **LLL**	
ACRONYM	EXPANSION
ACE *[eɪs]*	Automatic Content Extraction
AAVE *[eɪv]*	African American Vernacular English
SALT *[sɒlt]*	Speech Application Language Tags
CELEA *[siːlɪə]*	Chinese English Language Education Association
LAPTOC *[ˈlæptɒk]*	Latin American Periodical Table Of Contents

3.4. Acronyms in broader sense

In broader sense, they can be formed by omitting lexical words (nouns), grammatical words (prepositions, articles, conjunctions), punctuation marks (hyphens, slant lines), or even all three in the expansion or by using initials of affixes and combining forms, and by using initials for syllables and splinters.

3.4.1. Ellipsis in acronyms

Generally speaking, the ellipsis can be of three kinds: the omission of lexical words, the omission of grammatical words, and the omission of punctuation marks. Occasionally, there can be a combination of omissions. In *FLAC* an article was omitted; in *CALLSSA*, prepositions and a conjuction; in *BANA*, punctuation mark and a conjunction, and in *TEFL*, prepositions and an article.

TYPE: **LLL E**	
ACRONYM	EXPANSION
FLAC *[flæk]*	Foreign Language Across *the* Curriculum
CALLSSA *[ˈkɔːlsə]*	Center *for* Applied Language *and* Literacy Studies *and* Services *in* Africa
BANA *[ˈbænə]*	Britain, Australasia, *and* North America
WOZ *[wɒz]*	Wizard-of-Oz *simulation*
TEFL *[tefl]*	Teachers *of* English *as a* Foreign Language

3.4.2. Affixes, syllables, and splinters in acronyms

In the case of *PLL* for acronyms, there are only two affixes, i.e. *multi*- and *non*-, and one combining form – *indo*-. As far as syllables and splinters are concerned (*SLL* and *FLL* descriptors), in *RELC* and *COBUILD*, the syllables *[re]* and *[kəʊ]* were used as initials and in *COBOL* and *PRAESA*, the splinters <PR> and <CO> were used, as well as the omission of prepositions, the article and a noun.

TYPES: **PLL E; SLL; FLL E**	
ACRONYM	EXPANSION
MUD *[mʌd]*	Multi-User Domain
NESB *[nezb]*	Non-English Speaking Background
PIE *[paɪ]*	Proto Indo-European language
RELC *[relk]*	REgional Language Centre
COBUILD *[ˈkəʊbɪld]*	COllins Birmingham University Information Language Database
COBOL *[ˈkəʊbɒl]*	COmmon Business-Oriented Language
PRAESA *[ˈpreɪsə]*	PRoject *for the study of* Alternative Education *in* South Africa

3.5. Simple abbreviations in narrower and broader sense

Simple abbreviations are also examined on the basis of the bi-aspectual concept. In narrower sense (*LLL*), there are 21 simple abbreviations which can be differentiated by the selection of initials from the expansion elements, and in broader sense (*LLL E*), there is only one abbreviation. Namely, the abbreviations can be classified into those formed of one (*A* < 'Adjective'; *H* < 'Hearer'), two or three initials. Those formed of two and three initals are either abbreviated by the use of the first and the last grapheme in the word (*IR* < 'IntervieweeR'), by using the first and two final graphemes (*VCE* < 'VoiCE'), by abbreviating the first and the last consonant in the first syllable of the word, and a consonant in the second syllable (*MSC* < 'MaSCuline'), or by taking the inital of the first and one or two consonants of the second syllable in a polysyllabic word (*SG* < 'SinGular'; *DCL* < 'DeCLarative'). In MPH (< 'MetaPHor'), two consonants of the second (last) syllable were used in forming an abbreviation. There is only one example of ellipsis – *M* for 'Metapragmatic *joker*', in which a noun *joker* was omitted.

3.6. Blends in broader sense

The fact that blends are created by "*[...]* joining two or more word-forms through simple concatenation or overlap and then by shortening at least one of them" (López Rúa 2006: 677), explains the lack of the narrower-sense aspect in their formation. The word-forms joined in a blend are usually splinters or syllables and not initals, which means that, according to the elements, an *LLL* formation is not as expected as in other abbreviations.

In our corpus there are five blends of this kind: three of them are made of syllables and splinters (*SFL*: *AFRILEX* < 'AFRIcan *association for* LEXicography'; *PROLEX* < 'PROgramme *for* LEXicography'; *PROLOG* < 'PROgramming *in* LOGic'), one is made of more than one syllable (*SSL*: *FORTRAN* < 'FORmula TRANslation'), and one by joining at least two splinters (*FFL*: *PATR* < 'PArsing *and* TRanslation').

All blends, except *FORTRAN*, have undergone the ellipsis (*E*) of either lexical or grammatical words.

3.7. Clippings in broader sense

A similar reason of abbreviating splinters, syllables and an initial is the reason for not having a resultant *LLL* formation in a clipping, i.e. a narrower-sense formational aspect. In the broader sense there are nineteen clippings divided in six different subcategories.

3.7.1. Small letters in clippings

There are four examples of clippings with small letters. Three of them are made of an initial and the following two small letters (*Lll*: *Adv* < 'Adverb'), and one is made of small letters with the omission of a word and a hyphen (*lll E*: *wh* < 'wh-word').

3.7.2. Syllables and splinters in clippings

In the case of syllable use in formation of clippings, they make the largest group of such formations and can be classified in three subgroups: those having one syllable in a resultant clipping (*SLL*), one with a syllable and ellipsis (*SLL E*), and one made of two syllables (*SSL*). The examples from the first group are either formed of initials from the whole syllable or of initials from the whole first syllable and the initial from the second syllable. The ellipsis is realised in the omission of a lexical word in front of the clipping, while in the class of polysyllabic formations, there is one example with two syllables written with capital letters. The last example in the table presents a clipping formed of a splinter.

TYPES: **SLL; SLL E; SSL; FLL**	
CLIPPING	EXPANSION
ACC	ACCusative
ACT	ACTive
COP	COPula
REF	*establishing* REFerence
INACT	INACTive
PL	PLural

3.8. Hybrid forms

3.8.1. An alphabetism + an acronym

The classification of hybrid forms is not always bi-aspectual because the elements used in their formation are of various kinds, giving different results. The hybrids belonging to the criterion of broader-sense formation are classified into two subgroups: hybrids described by *LLL E* with the omission of adjectives, prepositions, articles, conjunctions, nouns and punctuation marks (e.g. *IATEFL* < 'International Association *of* Teachers *of* English *as a* Foreign Language'), and by *LLL E, A*, in which both the omission and addition (of a hyphen) take place (*CD-ROM* < 'Compact Disk Read-Only Memory').

3.8.2. An alphabetism + a word

The most numerous ones (24 examples) are those found in the group of hybrids made of alphabetisms and lexical words (alph+w). They are subdivided into seven different subgroups. There are those described by an *LLW* descriptor (e.g. *ATN grammar* < 'Augmented Transition Network grammar') and by *LLW E*, in which omissions of prepositions, articles, conjunctions or punctuation marks take place (*ELT dictionary* < 'English *as a* Foreign Language dictionary'; *IP grammar* < 'Item-*and*-Process grammar'). Addition of a hyphen is present in hybrids described with *L-LW A* (*NP-anaphora* < 'Noun Phrase anaphora'), a slant line with *L/LW* (*ID/LP format* < 'Immediate Dominance/Linear Precedence format'), the presence of the inital for an affix with *PLW* (*MD approach* < 'MultiDimensional approach') and a word (i.e. an affix) used in front of an alphabetism and described by *W-LL* (*hyper-RP* < 'hyper-Received Pronunciation').

3.8.3. An alphabetism + a simple abbreviation

Three subgroups of broader-sense formations make the group of alphabetism + simple abbreviation hybrids. The first is represented by *LLl* in which the initials are combined with the inital and the first small letter of the final syllable in a polisyllabic ultimate word of the expansion (*LCPt* < 'Language Corpus Politics'), or by combining initials with the inital and the final small letter of a monosyllabic word (*LHRs* < 'Linguistic Human Rights'). The second type is the one with the omission of a hyphen combined with the abbreviation (*TRPs* < 'Transition-Relevance Places'), while the third, the *LlL* descriptor, represents a combination of initals, both written in capital and small letters, and of a simple abbreviation (*SaPs* < 'Speech acts Projections').

3.8.4. An alphabetism + a splinter; an alphabetism + a splinter + a word

Seven examples are analysed in this group of hybrid formations. In alphabetism + splinter subgroup, three combinations have been attested: an *LLL* narrower-sense combination (*LGR* < 'Lengthened GRade'), an *LLL E* broader-sense combination (*EGR* < 'Extra *high pitch* GRade') with an inital, a splinter and the omission of lexical words, and an *LLl* combination (*BSAfE* < 'Black South African English') with initials and a splinter made of both capital and small letters. A similar way of formation with a different result is found in the subgroup of alphabetisms combined with splinters (*FLL E*), e.g. *ALLEX* (< 'African Languages LEXical *project*') and *SIGLEX* (< 'Special Interest Group *for* LEXical *resources*') are realised as acronyms, made of initials and a splinter, and both changed by the ellipsis. The last example of the third subgroup *CHILDES database* (< 'CHIld Language Data Exchange System database') combines the initials, a splinter and a word, and is modified by the addition of the word (*FLW A*).

3.8.5. An acronym + a word

There are two hybrid forms in the group of combinations made of an acronym and a lexical word. The *WLL* descriptor stands for a word (i.e. a combining form) combined with an acronym (*PanSALB* < 'Pan South African Language Board'), and *LLW E* for *BASIC English* (< 'British, American, Scientific, International, Commercial English') in which punctuation marks were omitted.

3.8.6. A simple abbreviation + a word; a simple abbreviation + a clipping; simple abbreviation + simple abbreviation; simple abbreviation + an acronym

All four examples of the first subgroup are explained by the *L-LW A*, because they have all been formed by the initial from one expansion element, by the addition of a hyphen and by the use of a whole word (*C-command* < 'Constituent command'). One example in the second subgroup is described by the *LLL* descriptor and belongs to the narrower-sense formations (*HGR* < 'H-GRade'). The other example, which makes the third subgroup, is a combination of a simple abbreviation and an acronym, also featured by *LLL* (*LBOTE* < 'Language Backgrounds Other Than English'). The fourth is represented by *LLl* and examplified by a hybrid form *LPt* (< 'Language Politics'), in which the initial from the first element is combined with the inital and the first small letter of the final syllable in a polisyllabic second element of the expansion.

3.8.7. A clipping + alphabetism; a clipping + a word

One example per each of the subgroups make the following group of hybrids. *ToBI* (< 'Tones *and* Break Indices') is made of a clipping and an alphabetism, and by omitting the conjunction (*LlL E*), which results in an acronym. The other example is *INFL node* (< 'INFLection node'), made of a clipping combined with a noun (*LLW*).

3.8.8. Splinters + a simple abbreviation; a splinter + a simple abbreviation + a word; a splinter + a grapheme

In *EURALEX* (< 'EURopean Association fo*r* LEXicography'), described by *FFL E*, there are two splinters combined with a simple abbreviation, and orthographically changed by the ellipsis (omission of a preposition), while in *LISP language* (< 'LISt Processing language'), described by *FLW*, there is a splinter in a combination of a simple abbreviation and a word from the expansion. In an *F-LW* type, exemplified by *equi-NP deletion* (< 'equivalent Noun Phrase deletion'), a splinter was added (*A*) in front of an alphabetism which was combined with a word. A rather different combination is realised in *BIT* (< 'BInary digiT'), described by *FLL*, because it represents a combination of a splinter and not a simple abbreviation, but a final grapheme of the second word in the expansion.

3.8.9. A syllable + a simple abbreviation; a syllable + a splinter + a simple abbreviation

ALGOL (< 'ALGOrithmic Language') is understood as a hybrid abbreviation formed of two syllables and a simple abbreviation (*SSL*), while in three hybrids described by *SFL E*, a syllable is combined with a splinter and a simple abbreviation, resulting in an acronym, e.g. *ASIALEX* < 'ASIan Association for LEXicography'.

3.8.10. Phonetic respelling + an acronym

One example in the corpus of analysed hybrid abbreviations is formed by a one-grapheme phonetic respelling of the first expansion element combined with an acronym. The example for this hybrid is *XTAG* < 'Extensible Tree Adjoining Grammar', represented by the *LLL*, a narrower-sense abbreviation descriptor.

4. Conclusion

On the basis of the work carried out, we have come to the following conclusions. The aim of producing a dictionary of abbreviations in linguistics would involve the compilation of a bilingual, bidirectional, specialized (domain-specific, technical), synchronic, explanatory, alphabetically arranged dictionary which would be informative and encyclopeadic in content, and serve both non-specialized and specialized audience. The sources which have been used in compiling the corpus of abbreviations in linguistics have proved to be trustworthy, valuable and fundamental in the scope. Having analysed and classified the abbreviations in linguistics, it can be concluded that they can also be classified according to the bi-aspectual criterion of narrower and broader sense. The narrower sense of their creation refers to those formed by initial letters of each element in the expansion, and pronounced either by individual names of letters or as a word. The broader sense is understood as the ways and processes of formation, more or less different from the orthographic norms. The solution for the lexicographic presentation of abbreviations is based on Ingrid Fandrych's Multi-level approach which is comprised of three aspects. The first aspect of the multi-level approach, i.e. the structure and modes of production aspect, will be supplemented by our novel aspects on orthographic and morphological level.

Summing up the results of this research and comparing them with the previous one on abbreviations in medical terminology, we would like to state that both analyses have proven similar results, i.e. most of the alphabetisms were formed according to the criterion in narrower sense, while the ratio of those formed in narrower and broader sense for acronyms, which were fewer in number than alphabetisms, was in favour of broader-sense formations. As far as hybrid-form ratio is concerned, broader-sense criterion is also more evident among them. The direct results of the analysis have attested the possibility of applying previously devised descriptors, as well as some new descriptors, which have emerged in the analsyis of abbreviations in linguistics. A new descriptor for alphabetisms is *P-LL*, for acronyms and some clippings in broader sense it is *FLL*. Blends are represented by two new descriptors – *SFL* and *FFL*. In the group of clippings in broader sense, the new descriptors are *Lll* and *lll*, while in the group of hybrid formations the following new descriptors have been introduced: *LLW*, *PLW*, *L/LW*, *FLW*, *F-LW*, *S-LW*, *SFL*, *WLL*, and *W-LL*. Therefore, it appears possible to say that the concept of classification of abbreviations according to their orthographic and morphological features, and the concept of their lexicographic presentation comprised of multi-level approach are compatible and can be used as a valuable tool in compiling a future dictionary of abbreviations. Finally, as far as the micro-structure of the future entry is concerned, its first level will consist of seven structural elements, five of which are represented and explained by abbreviations *T* (type), *E* (expansion), *M* (medium), *D* (descriptor), and *O* (origin), and two making the headword (spelling), i.e. its variant forms, and pronunciation.

References

Algeo, John (ed.) (1991). *Fifty Years Among the New Words*. Cambridge University Press: Cambridge.

Bieswanger, Markus (2007). “2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space and time-saving strategies in English and German text messages”. In Floyd, Simeon et al. (eds.), *Texas Linguistic Forum*, Volume 50.

Cannon, Garland (1989). “Abbreviations and Acronyms in English Word-Formation”. In *American Speech*, Vol. 64, No. 2.

Crystal, David (1995). *The Cambridge Encyclopaedia of the English Language*, Cambridge University Press: Cambridge.

Crystal, David (2004). *A Glossary of Netspeak and Textspeak*. Edinburgh University Press: Edinburgh.

Fabijanić, Ivo, Frane Malenica (2013). “Abbreviations in English medical terminology and their adaptation to Croatian”. In *JAHR*, Vol. 4, No. 7.

Fandrych, Ingrid (2004). *Non-Morphematic Word-Formation Processes: A Multi-Level Approach to Acronyms, Blends, Clippings and Onomatopoeia*. Unpublished PhD Thesis. University of the Free State: Bloemfontein.

Fandrych, Ingrid, (2007). “Electronic Communication and Technical Terminology: A Reapproachment?”. In *Nawa Journal of Language and Communication*, Vol.1, No. 1.

Fandrych, Ingrid, (2008a). “Pagad, Chillax and Jozi: A Multi-Level Approach to Acronyms, Blends, and Clippings”. In *Nawa Journal of Language and Communication*. Vol.2, No. 2.

Fandrych, Ingrid (2008b). “Submorphemic elements in the formation of acronyms, blends and clippings”, In *Lexis 2*: *“Lexical Submophemics / La submorphémique lexicale”*.

Harley, Heidi (2006). *English Words: A Linguistic Introduction*, Blackwell Publishing: Oxford.

Jackson, Howard, Etienne Zé Amvela (2005). *Words, Meaning and Vocabulary*, Continuum: London.

Kompara, Mojca (2012). “The first Slovene automatically compiled dictionary of abbreviations”. In *Proceedings of the 15th EURALEX International Congress*. University of Oslo: Oslo.

Landau, Sidney I. (2001). *Dictionaries: the Art and Craft of Lexicography*. 2nd edn. Cambridge University Press: Cambridge.

López Rúa, Paula (2004). “Acronyms & Co.: A typology of typologies”, In *Estudios Ingleses de la Universidad Complutense*. Vol. 12, pp. 109-129.

López Rúa, Paula (2006). “Non-Morphological Word Formation”. In *Encyclopedia of Language and Linguistics (2nd Edition)*. Vol. 2., Elsevier: Oxford.

Malenica, Frane, Ivo Fabijanić (2013). “Abbreviations in English Military Terminology”. In *Brno Studies in English*. Vol. 39, No. 1.

Plag, Ingo (2003). *Word-formation in English*. Cambridge University Press: Cambridge.

Stockwell, Robert, Donka Minkova (2001). *English Words: History and Structure*. Cambridge University Press: Cambridge.

Sources

Filipović, R. (1990). *Anglicizmi u hrvatskom jeziku: porijeklo – značenje – razvoj.* Školska knjiga: Zagreb.

Anglicisms in European Languages. Manfred Goerlach (ed.). Oxford University Press: Oxford. 2005.

Concise Encyclopedia of Pragmatics. Jacob L. Mey (ed.). Elsevier: Oxford. 2009.

Hartmann, R. R. K., Gregory James (1998). *Dictionary of Lexicography*. Routledge: London

Encyclopedia of Language and Linguistics, 2nd edition. Keith Brown (ed.). Elsevier: Oxford. 2004.

Roach, Peter (2009) *English Phonetics and Phonology, Glossary: A Little Encyclopaedia of Phonetcs*. Online: http://www.cambridge.org/elt/peterroach

Kristal, Dejvid (1985.) *Enciklopedijski rečnik moderne lingvistike*. Nolit: Beograd.

The Handbook of English Linguistics. Aarts, B., April McMahon (eds.). Blackwell Publishing: London. 2006.

Bussman, Hadumod (1998) *Routledge Dictionary of Language and Linguistics*. Routledge: London.

Trask, Robert Lawrence (2005). *Temeljni lingvistički pojmovi*. Školska knjiga: Zagreb.

Kinship and some lexicographic issues

Cristina Fargetti

1. Introduction

As one of the tasks that a lexicographer proposes, when working with an indigenous language, is the treatment of kinship terms found. This is a complex task for various reasons, one being the difficulty in data notation, considering the varieties found. From the study carried out between Juruna, a language I have been studying since 1989, initially within project of Lucy Seki (Fargetti, 1992, 2007), I propose a discussion of some aspects of this task.

Juruna people, who speak the language with identical denomination, live in seven villages, near to the road BR-80, in the Low Xingu region – Tubatuba, Maitxiri, Pequizal, Paqsamba, Pakayá, Pakajá, Mupadá - and in Diauarum and Piaraçu Posts, in "Parque Indígena do Xingu", state of Mato Grosso, Brazil. The population has many children and is estimated in 400 persons, practically all of them speakers of the indigenous language, and with some kind of bilingualism (Portuguese and other indigenous languages of the region – which has more 16 people, with different languages).

There are anthropological studies about the Juruna kinship system (Lima, 1995, Oliveira, 1970), however, both do not coincide in their results, and have different types of inconsistencies, demanding recollect, and reanalysis. There is also another study (Araújo and Storto, 2002), which compares kinship terminology of Karitiana and Juruna, but with data of Landin, for the first, and Lima (op. cit.), for Juruna. So, in this paper, I try to discuss the problems to a lexicographic work, with this type of documentation, trying to think about some methodological approaches, for a linguist.

2. Anthropological basis

According to Levi-Strauss (2003, 1967 1st ed.), the prohibition of incest is typically human, however, very relative, because what is considered incest depends from culture to culture. This prohibition is the basis of kinship systems, which differ in each nation, as shown in many studies about Brazilian indigenous communities; Silva (2009), for instance, presents Waimiri-Atroari case, that differs radically from

Dravidian (India). In linguistic studies of indigenous languages, sometimes the complexity of the kinship system of the people is not shown, getting up with many questions about the use of the terms presented. Another difficulty is that in small societies, the degree of proximity between persons is very large, so for a woman, for example, one man may be her brother, her cousin and her husband, which sometimes results in terms of different relationship that the researcher has in mind (he may think the speaker said "cousin", but he said "brother"). Several anthropologists use a system of abbreviation for the English descriptions of each term. Ex: MB - "mother's brother" (the term used is not always corresponding to what we call "uncle" because it can extend to F- "father"). There is some discussion of this conventional notation in Viveiros de Castro (1995), which has different studies about Brazilian Indians kinship systems. Levi-Strauss (1945, apud SILVA, 1999) had criticized the excess in the abbreviation, citing Davis and Warner (1945) with the formalism: $\frac{C^{2a}/^{2d}}{\square}\,\underline{\text{o}}\,\frac{SU^{1a}}{8}$ / EGO = husband. What can we say about: FFFZSSD, FFMBSSD, FFFZDDD? Is it possible to describe these relationships? What can we do in the dictionary entries?

As pointed by Silva (1999), kinship systems are: a) a terminological system, with a vocabulary (linguistic phenomenon) and b) an attitude system, with codes of conduct between individuals, due to social networks (cultural phenomenon). So he says that the focus of analyses of Levi-Strauss was the second one, in a structuralist method, looking for abstract systems, leaving aside the semantics of the words, their linguistic part.

There are some manuals that introduce basic concepts of anthropological kinship studies, like Parkin (1997) and Ghasarian (1996). They explain concepts and types of theories in the field. But they aren't sufficient for a linguistic discussion that brings different questions when the researcher needs to recollect or reanalyze data.

The kinship system of Juruna is classified as Dravidian (structural relationship with the Dravidian system from South India), largely observed in the Amazonia. Its typical features are the existence of preferential marriages between cross-cousins and dichotomy between consanguinity and affinity (Lima, 1995). This means that the terminological system has influences of the type of relationships that are possible between persons: for cross-cousins, the parents that are siblings have different gender, and in this case, this type of cousins may marry; but parallel-cousins – with parents that are siblings of the same gender - cannot marry each other, because they are like siblings. Furthermore, in this system, there are differences of treatment between kin that are consanguineous and that ones who are not.

The rule for preferential marriage sometimes is not followed, because there isn't a possible cross-cousin. Then, it could exist for example an intertribal marriage, but never a marriage with a consanguineous relative, a cognate.

It displays distinction between the terms of reference and vocatives. However, sometimes vocatives can be used as reference terms. Ex: **upa** "my father", **baba** "vocative", **upa wï** 'my father came", **baba, wï ane**? "Daddy, did you come?", **baba wï** “my father came".

3. Methodological issues

There are some difficulties with the collecting of data about kin terms. Endangered languages, like indigenous ones, never have database of written texts, because available texts are in small quantities and cannot show this kind of words. In the case of Juruna, there are very few ones written by the language's speakers and some translated from oral narratives. Daily, observing the common use also proved to be less productive, because, perhaps due to the small time of permanency in field, few data could be found. An alternative tried was to ask to a speaker to tell his complete family history. Nevertheless, also in this case, such words had a small occurrence. It didn't cover all the kinship system, since the speaker selected parts of his history which only mentioned his parents, grandparents and siblings. Asking him to talk about other kin could lead to artificial productions, like active elicited data. Some solutions to this problem will be presented in the following section 5.

At the beginning, we know that we cannot search from our own culture, as in everything else. To collect the first data, we make the genealogy of the speaker, starting from descriptions (from the basic relations – father, mother, son, daughter, brother, sister – like "brother of your father") and not from descriptive terms of our culture (like "uncle"), because the kinship system could be classificatory (with same term for "father" and "uncle"). So we can have all kin proper names and then we ask the speaker how he or she calls this person, in reference or vocative way. There are almost always differences between female Ego and male Ego, let's say, between terms that a woman says and terms that a man says, and because of this factor, we have to ask the terms for men and women, not mixing all, in individual interviews. Real (1973), an anthropological manual, from the 1940's, previous to Lévi-Strauss, guides to ask the speaker all the terms that each relative could say to him, inversely. But this is not a good strategy, because it can lead to misunderstandings. It is better to make separated interviews, not mixing the female and male Ego terms.

Another problem is that in small societies, the degree of proximity between people is very large, so for a woman, for example, one man may be her brother-in-law, her cousin and her husband, which sometimes results in terms of different relationship that the researcher has in mind (he thinks the speaker said "cousin", but she said "brother-in-law"). Furthermore, the existence of previous anthropological studies not always help, because data sometimes don't match with our own analysis, occurring gaps and misconceptions.

Therefore, lexicographers need to reanalyze data from the indigenous language, and perhaps without the anthropologist participation. In some cases, there aren't any studies about the matter, and we should begin from zero. This brings the need? of anthropological readings, in a vast area of research, with more than two centuries of tradition.

4. Kinship and dictionary

First of all, we could think about how there could be an indigenous language dictionary. Is it possible to make something different from a vocabulary list? Which type of lexicographic work is possible to construct?

To make a work different than just a list requires a complex research, because indigenous languages, in general, don't have texts database, so the linguist needs to make a database, transcribing and translating oral texts, and helping indigenous teachers in writing, themselves, texts in their language, if it has a standardized orthography. The research needs to dialogue with different areas, trying to understand the indigenous way of thinking: anthropology, ornithology, zoology, botany, astronomy, etc. It's necessary, then, a group of collaborators, indigenous teachers among them, if they have some linguistic knowledge.

As we point in Fargetti (2013:128):

> Uma descrição adequada de itens lexicais no dicionário é uma meta a ser atingida, portanto. É muito ruim encontrar em obras descrições tais como "tipo de constelação", "tipo de ave", na ausência de um equivalente. Em termos de documentação para o futuro, para novas gerações de falantes e para pesquisadores, tais descrições são inadequadas, não permitem o entendimento e a comparação com outras línguas. Obviamente, tais descrições ocorrem em dicionários de línguas indígenas pelo fato de não termos, para regiões específicas do país e mesmo para comunidades de fala, estudos feitos por físicos, ornitólogos, ictiólogos, e outros. [83]

This means that a linguist, studying an endangered language, with almost no documentation and description, needs to make ethnographic research, in a dialogue with different areas, to achieve adequate knowledge about the lexemes he wants to describe in a dictionary.

This makes hard, long and expensive the construction of a general dictionary of the language. But the lexicographer may conclude some mini dictionaries, or vocabularies, with accurate research, about some semantic fields, before concluding the general dictionary. This can bring important knowledge to comparative studies and can help in bilingual education, in the indigenous society.

> In contrast to bilingual dictionaries of major languages, a dictionary of an endangered language does not primarily serve as a tool for translation or foreign-language acquisition, but more typically as a resource for research and a repository of information that is valuable for language revitalization and teaching in the speech community. (MOSEL, 2011:339)

So to whom should the dictionary serve? Answering this question is not always easy, because the answer, the final product, will present the characteristics of the image that the lexicographer has about its consultants, not always appropriate, according to the view of his critics. Because of these two goals, a resource for

[83] "An adequate description of lexical items in the dictionary is a goal to be achieved, therefore. It's too bad to find descriptions in works such as "kind of constellation", "type of bird" in the absence of an equivalent. In terms of documentation for the future for new generations of speakers and researchers, such descriptions are inadequate; do not allow the understanding and comparison with other languages. Obviously, these descriptions occur in indigenous language dictionaries because we do not have, two specific regions of the country and even speech communities, studies by physicists, ornithologists, ichthyologists, and others."

research (including historic-comparative studies) and an empowerment of the language, an indigenous language dictionary must present important grammatical knowledge and discuss all the lexicographic decisions, like the spelling of words (is there some orthography?), phonetic transcription, kind of language variant chosen, and so on. The grammatical knowledge, like word classes, is very important to linguistics studies, and also to meta-linguistic questioning made by indigenous teachers.

In indigenous languages, there is a problem to think in lemmatization: some of them have inalienable nouns, which always must have person marking; for example, we cannot say "foot", but "my foot, your foot, etc", and this possessive is generally an inflectional prefix. If there isn't "foot" in the language, but always an inflected word, with a prefix, how it could be lemmatized in the dictionary? Inflating a letter, at the beginning of all inalienable words, being consistent in a choice of a person, like "my...", always in the same way? Changing the person marking, sometimes "my...", sometimes "your...", "our..." ? Which is the argument to make a choice? It could be possible to find an inalienable word in the dictionary? It could be possible to know the difference with an alienable word? This will be discussed in section 5 below.

Semantic fields are a chance for better detail in obtaining more reliable and better definitions. This requires collaborator researchers in a partnership to study words for birds, food, body parts, physiology, etc. And also requires a more direct participation of speech community, to diminish the asymmetry researcher-consultant, in a dialogue between real interlocutors. In the case of Juruna teachers, some have university degrees, and can have a meta-linguistic thinking. This helps a lot in collecting data for the lexicographical making which is really different from collecting for language structure investigations: we need special questionnaires, tests and analyses for the achievement of the person marking system of a unknown language, for example; but the sentences obtained are not so good like examples in a dictionary, they are not real communication and can sound false to the speaker. As I pointed in Fargetti (2010: 120), when discussing a Juruna bilingual book, about material culture:

> Embora os textos elaborados pelos Jurúna apresentem boas informações sobre cada item, não constituem uma entrada elaborada, com critérios lexicográficos adequados, nem permitem, com a linguagem adotada pelos autores, uma comparação aos elementos da cultura material de outros povos. É preciso pensar na relação discurso indígena x sistematização lexicográfica, tendo em vista, obviamente, o resultado final, o dicionário, que satisfaça às necessidades de documentação dos falantes (o que ficará documentado para o futuro?) e as necessidades da academia.[84]

[84] "Although the texts prepared by Juruna have good information about each item, they do not constitute an elaborate entry, with adequate lexicographical criteria, nor allow, with the language adopted by the authors, a comparison to the elements of the material culture of other people. You need to think of the relationship "indigenous discourse x lexicographical systematization", of course, with a view to the final result, the dictionary, which satisfies the needs of documentation of speakers (what will be documented for the future?) and the needs of the academy."

So, even texts that describe material culture items (thought sometimes as handcrafts) are not completely appropriate, for lexicographic purposes. Some sentences can be used, like examples, but we need some kind of normalized description, comparable within languages and cultures like anthropologists remark in proposed norms of description (important referential works are RIBEIRO, 1988, 1989).

Furthermore, to the dictionary, we need more complex information, knowing if there are social variants or gaps (is there some kind of verbal incorporation? to which words?). So, the direct participation of the speakers in the lexicographic work is important, although sometimes difficult.

There are restrictions in lexicographic work, with an endangered language, and we can talk about:

a) limitations of the researcher - little knowledge of the language , little profundity into issues of language/culture relationship, lack of available time, lack of a team of collaborators and pressure for high productivity;

b) limitations given by the situation of language use - few speakers, speaking into disuse;

c) political constraints , difficulty of entry into indigenous areas (legal impediments and / or academic), non- acceptance of the work by speakers, conflict of interest between academic centers;

d) lack of a digitalized database ;

Therefore, our work is in the medium and long term. A dictionary cannot be developed quickly, particularly with the conditions of research that we have working with an indigenous language: little or no documentation, difficult access to speakers of the language, difficulties of financial support, etc. But we must have respect for indigenous languages, thinking that all people deserve a proper lexicographical work, not superficial ones.

In dictionaries of indigenous languages, sometimes the complexity of the kinship system of the people is not shown, getting up with many questions about the use of the dictionarized terms. In Ramirez (2001,p. 178), a Baniwa dictionary, we can find:

> - **KITSI**- 1. –**kítsini** n. dep. primo paralelo (filho do irmão do pai ou da irmã da mãe), parente clânico: nokítsini *meu primo paralelo*. **Nokísi ñai** *meus primos paralelos, os de meu clã*[85]

There is term for female cousin and other clan denominations, but there is no mention of the difference between masculine and feminine Ego, nor the question of the difference between reference term and vocative term.

The structure of entries requires semantic knowledge of the language enough to not run into the same misunderstandings and prejudices. If the lexicographer parts of his language, trying to find correspondences in the indigenous language, he might get inconsistent data, supplied by an informant willing to get rid of a job that

[85] - KITSI -1 . - kítsini n . dep . parallel cousin (father's brother's child or mother's sister child) , clan relative: nokítsini *my parallel cousin*. **Nokísi ñai** *my parallel cousins, those of my clan*

he barely understands and that will be criticized, at another time, by other speakers of the language, consulting the work done.

Also setting the lemma is not always something easy. For verbs, we have the infinitive form in Portuguese dictionaries, however, in indigenous languages, in general, we do not talk about "infinitive" because there is no evidence for postulating it in these languages. The existence of a verbal mode is explained by comparison with other existing modes and the function or functions that a specific form assume in the language.

Dialectal differences may be another "thorn", but one should keep in mind that every society has its "standard language" unfortunately always dictated by the ruling class, leaving aside other " dialects " , considering them " hillbillies, wrong , etc. " (because thinking so about their speakers). We can warn people against prejudice, but there is nothing to do about the linguistic normalization, because every society has its own.

5. Some entries

In Juruna I have constructed, until now, 39 entries for kinship terms, 10 of them are alienable, and the rest inalienable. These ones need to have a prefix of person marking, a possessive. The difference can be seen below:

-pa . n. [pá](termo de referência; identidade para ego masculino e ego feminino; consanguíneo) 1. F pai, *upa wï* meu pai chegou; 2. FB irmão do pai (também **-pa nana**) 3. MH marido da mãe; 4. MZH marido da irmã da mãe ~ **-pa i'uraha .**n. eFB irmão mais velho do pai ~ **-pa iza.** n. yFB irmão mais novo do pai (cf. **baba**)[86]

baba. n. [⍰bába] (termo de referência e vocativo; identidade para ego masculino e feminino; consanguíneo). 1. F pai , *baba wï* "o pai chegou"; *baba, atene wï* "pai, venha logo" ; 2. FB irmão do pai (**pã** – termo em desuso, mas tido como mais tradicional) (cf. –pa)[87]

The second term is alienable, it doesn't need person marking, but the first term is inalienable, it never could occur without a person marking, a possessive, but the lemma was proposed without any mark of person, using just a hyphen to signalize the need of such a mark. This is a lexicographic decision by an abstract form, to avoid inflating a letter, at the beginning of all inalienable words, if we have had chosen a particular person mark, like "our-inclusive" (in elicitation Juruna speakers almost always say inalienable words as "our foot", "our father", and so on). It could have some criticism, saying that speakers couldn't recognize their language, because without person marking, the abstract form is seen as inexistent. But what we must

[86] **-pa .** n. (reference term; identity for male and female Ego; consanguineous) 1. F father, *upa wï* 'my father came'; 2. FB father's brother (also **-pa nana**) 3. MH mother's husband; 4. MZH mother's sister's husband ~ **-pa i'uraha .**n. eFB eldest father's brother ~ **-pa iza.** n. yFB youngest father's brother (cf. **baba**)

[87] **baba**. n. [bába] (reference term and vocative term; identity for male and female Ego; consanguineous). 1. F pai , *baba wï* "father came"; *baba, atene wï* "daddy, come soon" ; 2. FB father's brother (**pã** – term into disuse, but considered more traditional) (cf. –pa)

think is that every dictionary deals with abstraction: we cannot find inflected verb forms and feminine nouns, in a Portuguese dictionary. Is it a problem to a consultant? Don't we know that these forms cannot be found? Surely, the lexicographic decisions must be known by the consultant, and this may happen with indigenous languages too, in different manners, of course, because languages are different.

Both terms have phonetic transcription and this is important, since it can mark prosodic features absent in written spelling: it is a tonal language, marked accent of intensity according to alternation of tones in the word. So, 'my father' and 'leaf', homographs by orthography, have different pronunciation, with the same stressed syllable, but with different alternating tones. In [u⍰pá] 'my father', low-high, and [ú⍰pá] 'leaf', high-high, the coincidence of stress is due to the application of default: the stress goes to the first syllable with high tone, from left to right, and if the tones are the same, the last syllable is stressed (cf . Fargetti, 2007). The orthography attempts to avoid excess of diacritics, which led to abolish any accent mark. The speaker of the language has the context to understand the pronunciation of the word (cf. Fargetti, 2012).

The linguistic and anthropological information that follow serve to show the distribution of the term, if it can function as reference or vocative; also whether there is identity for male speech and female speech (male Ego and female Ego), and the distinction "consanguineous x affine?". This information is important to better understand the system in question. In renderings, the English abbreviations are translated into Portuguese because it is important for the user who does not know it, including Juruna people. Examples were elaborated from possible occurrences between the speakers, and the verb 'come' is interesting even to present low tone and so allow us to know with certainty the tones of kinship term. However, such examples will be discussed further with the speakers of the language.

6. Conclusion

We have tried to discuss punctual issues, like the use of person marking and problems it could bring to a dictionary user; the knowledge needed to recollect and to re-analyze kinship data; some entries in Juruna language, trying to show a comprehensive discussion about some decisions. But, some questions are still open, such as: information about the kinship system itself, bringing the possibilities of relationship between kin and affines?, should be included in the entries? Accordingly, restrictions should be highlighted, even though sometimes rules are broken in some marriages? Anyway, what is enough information for the entries?

References

Araújo, C.; Storto, L. (2002) "Terminologia de parentesco Karitiana e Juruna: uma comparação de algumas equações entre categorias paralelas e gerações alternas". In Cabral, A.S.; Rodrigues, A. (eds.) *Atas do I Encontro Internacional do GTLI da ANPOLL*. Tomo II. Belém: EDUFPA, 430-442.

Fargetti, C. M. (2013) "Pesquisa de línguas indígenas – questões de método." Trilhas Linguísticas 23. São Paulo: Cultura Acadêmica. 115-135

Fargetti, C. M. (2012) "Dicionários de línguas indígenas e questões de prosódia." In Fargetti, C. M. (ed.). *Abordagens sobre o Léxico em Línguas Indígenas.* Campinas: Curt Nimuendaju. 67-82.

Fargetti, C. M. (2010) "Cultura material indígena: questões lexicográficas." In Cabral, A.S.; Rodrigues, A.; Duarte, F. B. (eds.) *Línguas e Culturas Tupí.* 1 ed., v. 2. Campinas: Curt Nimuendajú. 117-129.

Fargetti, C. M. (2007*) Estudo Fonológico e Morfossintático da Língua Juruna*. 1 ed., v. 1. Muenchen - Alemanha: LINCOM EUROPA.

Fargetti, C. M. (1992) *Análise Fonológica da Língua Juruna*. Campinas: UNICAMP.

Ghasarian, C. (1996) *Introdução ao Estudo do Parentesco.* Lisboa: Terramar.

Lévi-Strauss, C. (1967) *As estruturas elementares do parentesco*. (2003) [3rd ed.]. Petrópolis: Vozes.

Lima, T. S. (1995) *A parte do Cauim – etnografia juruna.* Rio de Janeiro: UFRJ

Mosel, U. (2011) "Lexicography in endangered language communities." In Austin, P.; Sallabank, J. (eds.) *The Cambridge Handbook of Endangered Languages.* New York: Cambridge University Press. 337-353

Oliveira, A. E. (1970) *Os índios juruna do Alto Xingu. Dédalo*, VI, 11, 12. São Paulo : USP.

Parkin, R. (1997) *Kinship:* an introduction to the basic concepts. Oxford: Blackwell.

Ramirez, H. (2001) *Dicionário Baniwa-Português.* Manaus: Editora da Universidade do Amazonas.

Real Instituto de Antropologia da Grã-Bretanha e Irlanda (1973) *Guia prático de antropologia.*[2ªed.] São Paulo: Cultrix.

Ribeiro, B. (1988) *Dicionário do artesanato indígena*. Belo Horizonte: Itatiaia; São Paulo: EDUSP.

Ribeiro, B. (1989) *Arte indígena, linguagem visual/ Indigenous art, visual language*. Belo Horizonte: Itatiaia; São Paulo: EDUSP.

Silva, M. (2009) *Romance de Primas e Primos – uma Etnografia do Parentesco Waimiri-Atroari*. Manaus: Valer/EDUA.

Silva, M. (1999) "Linguagem e parentesco." *Revista de Antropologia*, vol. 42. n 1-2. São Paulo.

Viveiros de Castro, E. (1995) *Antropologia do Parentesco – Estudos Ameríndios.* Rio de Janeiro: UFRJ.

A new pedagogical dictionary for DAF and ELE with an onomasiological focus

Maria Egido Vicente, Manuel Fernández Méndez & Mário Franco Barros

1. Diconale[88]: description and intended users

The majority of currently available bilingual dictionaries[89] for German and Spanish are aimed at a specific type of user, frequently as a result of the demands of publishers. Thus, for example, the most commonly used dictionaries conceived of as works of bilingual lexicography are designed primarily to suit the needs of users with German as a mother tongue (user 1). The German-Spanish part of the dictionary is thus active, facilitating production in the target language (ELE), whereas the Spanish-German part is essentially intended for the reception of the foreign language (ELE). For a learner of German with Spanish as a mother tongue (user 2), consulting such a dictionary involves exactly the opposite perspective, with the lexical information offered for German as a starting point and as the target language being insufficient.

Taking this as a point of departure, the DICONALE research project proposes the development of a bilingual and bidirectional dictionary of verbal lexemes for the German-Spanish language pair, with the aim of overcoming the shortcomings which tend to arise through not accounting for the different perspectives of access of users (production and reception in native and foreign languages) in the contexts of DaF (German as a foreign language) and ELE (Spanish as a foreign language).[90]

DICONALE is conceived fundamentally as a tool for *online* use for learners of German or Spanish as a foreign language (DaF / ELE) from an advanced level (B2),

[88] A conceptual dictionary of German and Spanish, the development of which has been framed within the following research projects, directed by Dr Meike Meliss, University of Santiago de Compostela: DICONALE-online: 'Elaboración de un diccionario conceptual bilingüe del alemán y del español: recurso online' (MINECO-FEDER FFI2012-32658) and DICONALE: 'Estudios para la elaboración de un diccionario conceptual de lexemas verbales del alemán y español' (Xunta de Galicia: 10PXIB 204 188 PR).

[89] Langenscheidt (2010), PONS (2007) , Slaby-Grossman (1994)

[90] For a more detailed discussion of this see Meliss (2011) and Meliss (2013).

according to the Common European Framework of Reference, and especially for the process of producing texts of a certain stylistic complexity in the foreign language. It offers a novel approach in the field of conceptual, onomasiological lexicography, and is also framed within the research area of corpus linguistics. The user accesses the dictionary in order to find the verbal lexemes needed to produce a text, and does so in the first instance by locating them within the conceptual fields offered in DICONALE. The search then continues through the accessing of lexical-semantic paradigms to which the lexemes corresponding the searched concept are attributed. The proposed dictionary is based on a modular and multilateral description[91] which allows the user to access information on verbal lexemes[92] in each language in a bidirectional way. The two languages under study – German and Spanish – receive the same treatment, manifested above all in different paradigmatic and syntagmatic relations, as well as in the correlation and reciprocity between content and form, which also favours a contrastive focus.

Methodologically, DICONALE begins with an empirical approximation to the object of study in both languages, this based on texts from print media covering the last 20 years, drawn from DEREKO[93], CREA[94] and WebCorp[95]. This allows us on the one hand to offer authentic examples in the dictionary, and on the other favours a degree of comparability of the two languages in question founded on contrastable, empirical data and frequencies.

Opting for an *online* version of the dictionary rather than *print* format reflects the real needs of users, given that for some time now the use of this type of dictionary has been the preference of the majority, as opposed to traditional format dictionaries (Meliss: 2013), especially among learners of foreign languages (see for example Pons[96] or Leo[97]). Recently some of these dictionaries (for example Pons) also offer the ability to download a free *smartphone* application. In addition to such economic factors, *online* dictionaries offer additional advantages, of which the following can be noted: faster searches, a dictionary which is always updated, interactivity and participation by users (for example, through sending translations of terms or meanings that currently do not figure in the dictionary), having the text in digital format, being able to copy and paste it in other documents, accessibility, immediacy etc.). In what follows we will concentrate on onomasiological issues and the structure of the dictionary, these being novel elements in this dictionary.

[91] See in the same volume the contribution of Meliss & Sánchez Hernández, which includes a detailed explanation of DICONALE's model of description.

[92] We refer to simple and affixed verbal lexemes as well as to plurilexemic ones, such as collocations or constructions with functional verbs.

[93] DeReKo: Das Deutsche Referenzkorpus, Institut für Deutsche Sprache.

[94] CREA: Corpus de Referencia del Español Actual, Real Academia Española.

[95]WebCorp created and managed by the Research and Development Unit for English Studies (RDUES) of the School of English, Birmingham City University.

[96] http://es.pons.eu/traducción .

[97] http://dict.leo.org/esde/ .

2. The onomasiological perspective and the structure of the dictionary

Unlike semasiological dictionaries, aimed above all at reception in the foreign language, onomasiological dictionaries more clearly offer assistance for the user in the expression in the second language of that which he or she has originated in the mother tongue, or in confirming hypotheses made during the verbalisation of concepts in the foreign language. However, existing onomasiological dictionaries[98] do not offer the assistance necessary for the user to adequately employ the lexeme found, in that they lack information on meaning and illustrative examples of usage which would make it possible to attend to the three basic issues which arise: 1) from the resources available, which form is best suited to the context, 2) what are the combinatory possibilities, and 3) what are the relevant contrasts between the two languages.

DICONALE aims to take address the challenge of offering the user the information required for linguistic production in the target language, featuring a "distinctive model" (cf. Haß-Zumkehr 2001: 269) in that it offers, in addition to the structuring of lexemes related to different concepts, information on the usage and meaning of the terms associated with them.

The lexical material is organized, in the first phase of the work, in ten conceptual fields (macrostructure), which the user accesses to carry out a search, and which in turn are deployed in conceptual subfields of different degrees. By associating the different possible lexemes, different lexico-semantic paradigms are created in both languages. For this, information is provided, at both a monolingual level and from a contrastive point of view, so that the correct choice of lexeme can be made, the one that best suits the situation in question (microstructure). These lexemes mainly differ from each other from an intra- and interlingual the point of view, based on the following parameters of description: (i) the existence of distinctive semantic features, (ii) differences in the morphosyntactic structure of some arguments, and (iii) semantic categories and descriptors that are associated with the realisation of each argument. All this makes it possible to create and describe the combinatorial profile of each lexeme, presenting it in a structured and detailed way and in a contrastive form.

There follows, as a means of illustrating what has been described above, a concrete example from the conceptual field CHANGE OF POSSESSION[99], this deriving from the subfield ROBAR (to rob). Related to this concept, the user might have in mind at least three distinct subconcepts corresponding to the following three conceptual subfields (SCC) represented by the images in *Table 1* (situations 4.3.1, 4.3.2 and 4.3.3), each with its respective argument structure:

SCC1: Robbery of someone with violence or threats[100]: "*Someone (A1) robs someone (A2)*".

SSC2: Wrongful, concealed theft from someone. "Someone (A1) steals something (A4) from someone (A2)".

[98] Among others Dornseiff (2004) for German and Casares (1981) for Spanish.

[99] One of the ten conceptual fields analysed in this phase of DICONALE.

[100] (A1) ladrón (thief), (A2) víctima (victim), (A3) lugar (place), (A4) objeto robado (stolen object).

SSC3: To break into a place using violence in order to steal. "*Someone (A1) breaks into a place (A3) to steal something (A2)*".

From these, the possible lexemes that exist in each language to verbalize each of these conceptual ideas emerge. The set of possible lexemes for a common concept constitutes a lexico-semantic paradigm (Pls). In this case there arise at least the following three paradigms for each language represented in Table 1; in the dictionary the user can find, linked to each of these, the following information corresponding to the four modules of description which complete the microstructure of the lexemes:

Module 1: General information on each meaning: map of the expression (morphology and phonetics), map of content, formation of words, etc.

Module 2: Meanings and variants relevant to the semantic field: explanation of the signified, paradigmatic relations, structural schema and register of use.

Module 3: Combinations of each meaning (argument structure, etc.) and correspondence in the target language.

Module 4: Other pertinent grammatical information (use of the passive, etc.).

Thus, from such a conceptual form of access, the possible utterances in both languages are offered, these detailed and contrasted in all their relevant aspects as a means of facilitating the choice and correct use of the lexeme corresponding to the initial idea.

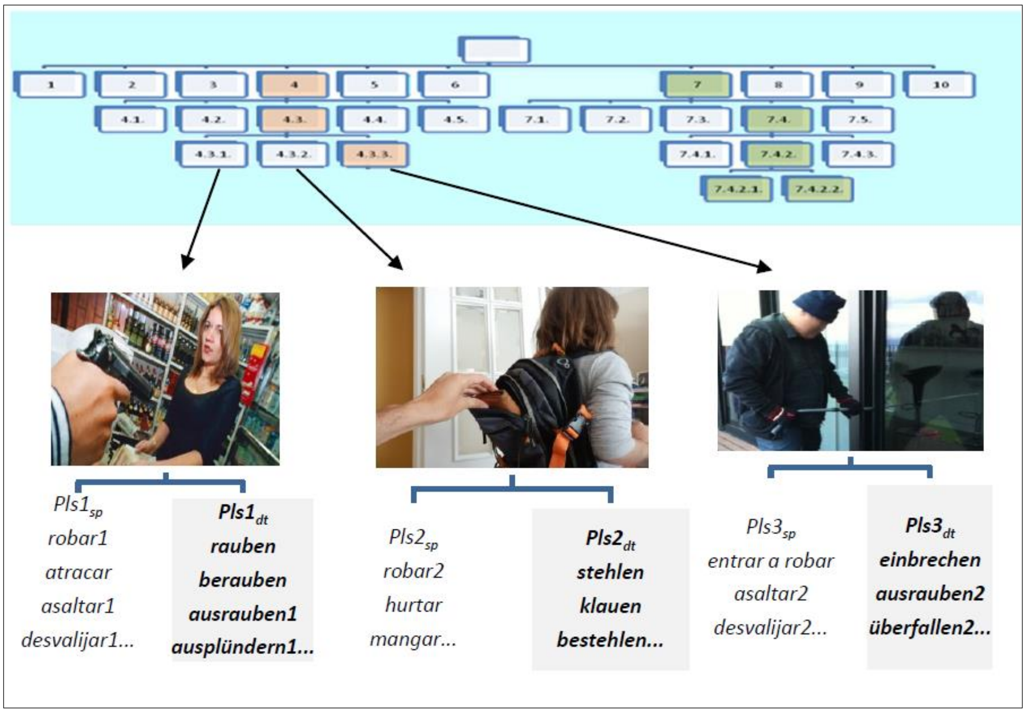


Table 1. General structure of the dictionary using the example of the conceptual subfield ROB in German and Spanish.

As an example of one of the most important contrastive aspects, shown here is the description of three lexemes in each language, corresponding to the lexico-semantic subparadigm (PLS3) "einbrechen / break in and steal". *Table 2* shows some of the parameters of Module 2 in the model of description, specifically the information about the combinations marking their differences (argument structures, morphosyntactic function and “semantische Füllung”)[101] and is illustrated with examples from the previously mentioned reference *corpora* (cf. *Table 2*):

Sub-paradigma 3 “einbrechen/ entrar a robar”[102]	**Ejemplos**[103]	**Rasgos semánticos distintivos**	**A1**[104] **Ladrón**	**A2**[105] **Víctima**	**A3 Lugar**	**A4 Objeto robado**
einbrechen	(1)	[+violencia]	s[106] [hum][107]	Ø	adv. [dir /sit]	-
ausrauben2	(2)	[+violencia] [+perfectivo]	s [hum]	Ø	cd [sit: casa, banco, negocio]	-
überfallen2	(9)	[+violencia]	s [hum]	Ø	cd [sit: banco, negocio]	-
entrar a robar	(6) (7)	[+/-violencia]	s [hum]	Ø	adv [sit: banco, negocio, casa etc.]	(cd) [mat: objeto: dinero, joyas etc.]
asaltar2	(4) (5) (3)	[+violencia] [-perfectivo]	s [hum]	Ø	cd [loc: casa, banco, negocio, medio de transporte]	-
desvalijar 2	(8)	[+violencia] [+perfectivo]	s [hum]	Ø	cd [loc: casa, banco, negocio, medio de transporte]	

Table 2. Comparison of some representative lexemes which verbalise the concept EINBRECHEN / ENTRAR A ROBAR (to break in and rob) corresponding to Pls3: Semantic and combinatory information.

The user gets the contrastive information which makes it possible to choose the appropriate lexeme for the specific needs of verbalisation, and at the same time finds morphosyntactic information relating to the argument structure, which allows for the correct use of the lexeme. For example, having detected a direct relation

[101] Others details have not been discussed here due to limitations of space.

[102] Spanish has been used as the metalanguage.

[103] See the numbered examples in the Annex.

[104] Of special note here is that in this subparadigm the argument referring to the victim is not realised, something which does indeed occur in the subparadigms 1 and 2 (Pls1 and Pls2).

[105] In this paradigm the information given in the arguments A2 y A4 is not relevant, but these arguments can be realised in the other subparadigms of this conceptual field.

[106] **s** = sujeto (subject), **cd** = complemento directo (direct complement), **adv** = complemento adverbial (adverbial complement).

[107] Features for the semantic-categorial description of the arguments are based on Engel (2004).

between *überfallen-asaltar*, between *ausrauben2* and *desvalijar 2*, and between *einbrechen* y *entrar a robar*, the user can compare the argument structure of two of the most used lexical units in both languages, *einbrechen* and *entrar a robar*, in order to understand their coincidences and above all their divergencies: *einbrechen* does not specify – in general – argument 1 (thief) of action [A1], and in this meaning is used almost exclusively in the passive without a personal subject, indicating either a direct movement to the place in question [A3], expressed as a direct complement adverbial (adv$_{dir}$), or the place in question [A3] of the action as a location through a situative adverbial complement (adv$_{sit}$). Nevertheless, *entrar a robar* (break in to steal) permits the expression of [A1] in the usual form in the active voice through a human subject and is not normally used in the passive to show impersonality. It only indicates the place in question [A3] through a situative adverbial complement (adv$_{sit}$), yet it does allow as an optional form the realising of the stolen object [A4]. If the communicative intention of the user is not accommodated in this verbal lexeme, he or she is invited to search the paradigm for another one which better suits the specific needs of expression, and to this end examples taken from the reference corpus are shown to the user.

The user of DICONALE can obtain the necessary information for production through searching for each lexeme (Spanish: *robar*, *atracar*, ... German: *rauben*, *berauben*...), and in this way can be assured of the right usage and combinations through the examples given. For each lexeme, the user receives information structured according to the four modules noted above, which can be accessed independently: in the information on each meaning and its correspondence in the language of production, examples from the reference corpus are given, and this in this way not only can the signified itself be contrasted, but also the argument structure of the lexeme in question, plus the specification of the arguments and correspondencies with their contexts.

In each example the syntactic arguments are displayed as hyperlinks in coded colours, so that after clicking on each one there appears information on its contrastivity, its paradigmatic relations, its usage log, plus information about the probability of it occurring in a given argument structure, and about other options for production (passive/active, deverbal nominalisations, etc.), all of which serve to improve its use in a given context. These examples of contrastive interest provide information about semantic relations of a paradigmatic type, with the aim that the user thus has several options. In this way the user can choose from among them the one that best suits the specific needs of production, such as the type of text, register or stylistic devices, while ensuring that this option allows for the combination with which the initial hypothesis was made, or finding other forms that do.

The dictionary allows the user to begin with a concept or idea, expressed verbally or conceptually. In the final search level, the perspective is semasiological, given that the initial idea has been delimited, and is searched for in lexematic or plurilexematic units arranged alphabetically within this level.

Although the dictionary is primarily conceptual-onomasiological, this semasiological search within the subfield can also provide valuable assistance. Also, thanks to the management and digital interface of the data, the possibility exists of offering a search for a lexeme from the initial input interface, so that it is redirected to different conceptual fields to which it may belong, according to its meanings. At that level, the user will be taken back to a conceptual-onomasiological reflection to check whether this was indeed the initial concept, and if not to search for another one more suited to his or her needs.

3. Conclusions

DICONALE, which employs a conceptual-onomasiological focus, offers an innovative and different form of lexical access in the field of ELE and DaF. With the *online* format, the dictionary adapts to the needs of accessibility and immediacy of a user in an increasingly digital environment, and in turn to the specific structural needs of a reference work. The modular approach of the dictionary[108], plus its bidirectional character, improve on the representation offered by traditional formats. From an informational point of view, DICONALE offers a wide range of syntagmatic and paradigmatic information of verbal lexemes for each language (argument and valency structure, sentence schema, etc.), thus furnishing the user with the necessary tools to successfully conclude an inquiry. This dictionary aims to overcome the limitations of some non-bidirectional bilingual dictionaries, especially in the context of a user actively producing text in the target language. Moreover, DICONALE must also be understood as a learning tool.

The empirical basis of the dictionary, beginning with the use of reference *corpora* in both languages, as well as frequency calculations carried out for the selection of the information given, allow a high degree of comparability of the languages in question and the verbal lexemes dealt with. This creates contrastive foundation, with which it is intended that the user reflects on the proper use of lexemes in both languages.

References

Casares, J. (1981^{2}) Diccionario ideológico de la lengua español. Barcelona et al.: Editorial Gustavo Gili.

DeReKo: Das Deutsche Referenzkorpus. Durch COSMAS II: Corpus Search, Managment and Analysis System. Mannheimer Institut für Deutsche Sprache [online]: DWDS: Digitales Wörterbuch der deutschen Sprache (Referenzkorpora) http://www.ids-mannheim.de/cosmas2/ (most recently consulted: 12.9.2013).

Dornseiff, J. (2004^{8}) Der deustche Wortschatz nach Sachgruppen. Berlin, New York: De Gruyter.

Engel, U. (2004). Deutsche Grammatik – Neubearbeitung. München: iudicium.

Hass-Zumkehr, U. (2001). Deutsche Wörterbücher. Berlin, New York: de Gruyter. 269.

Langenscheidt-redaktion (2010) Handwörterbuch Spanisch-Deutsch. Berlin und München.

LEO: http://dict.leo.org/esde/ (most recently consulted: 25.9.2013)

Meliss, M. (2011). "Wörterbücher von heute und morgen. Überlegungen zu lexikographischen Benutzersituationen im zweisprachigen Kontext: Spanisch-Deutsch". In Domínguez Vázquez, María José et al. (ed.). La Palabra en el Texto. Festschrift für

[108] For more information on the modular structuring of the dictionary, see the contribution of Meliss & Sánchez Hernández in this volume.

Carlos Buján. Santiago de Compostela: Universidade de Santiago de Compostela. 267-300.

Meliss, M. (2013). “Online-Lexikographie im DaF-Bereich: Eine erste kritische Annäherung: Bestandsaufnahme – Nutzen – Perspektiven”. In REAL (Revista de Estudos Alemães), Nr. 4. 176-199.

Real academia española: Banco de datos (CREA) [en línea]. Corpus de referencia del español actual: http://www.rae.es (most recently consulted: 12.9.2013).

PONS Wörterbuch fur Schule und Studium (2006). Deutsch-Spanisch. Ernst Klett, Stuttgart.

PONS. eu. Das Sprachenportal: http://www.pons.de/ (most recently consulted: 12.9.2013)

Slaby, Grossmann, I. (1994) Wörtebuch der Spanischen und deutschen Sprache. Herder, Barcelona

Annex

Illustrative examples of the lexemes in *Table 2*:

(1) Polizisten stellten fest: In das Haus war **eingebrochen** worden. Vermutlich hatten die Täter den Brand gelegt, um Spuren zu beseitigen. *Hamburger Morgenpost, 25.04.2013, S. 07.*

(2) Im feinen Blankenese kam es gestern Nachmittag zu einem blutigen Überfall. Ein etwa 20 Jahre alter Mann stürmte um 15.35 Uhr in die Schlecker-Filiale an der Blankeneser Bahnhofstraße. Er wollte den Laden **ausrauben**. *Hamburger Morgenpost, 28.04.2006, S. 14.*

(3) Dos encapuchados **asaltaron** y quemaron la sede socialista de Noreña. (...) *La Voz de Asturias, 21/08/2004.*

(4) Tres hombres **asaltaron** el martes en Madrid una oficina de una empresa de construcción y se apoderaron de más de cinco millones y medio de pesetas, (...). *El País, 02/06/1988.*

(5) Según una información de la Delegación del Gobierno, sobre las tres de la tarde varios musulmanes, integrantes de un piquete, **asaltaron** un automóvil en cuyo interior viajaban varios cristianos que se dirigían a la vivienda de uno de ellos. *El País, 02/02/1987.*

(6) La Policía ha detenido a tres atracadores que **entraron a robar** en el domicilio de una anciana de 80 años, donde maniataron a su nieto, de 14, (...). *El Mundo 13/06/1994.*

(7) Los Mossos d'Esquadra detuvieron a dos chicos de 20 y 22 años, vecinos de Bordils y Celrà, (...), y a un menor de edad que **entraron a robar** cromos de la Liga de Fútbol en la empresa Panini de esta última localidad,(...). *Vozpopuli online:* http://vozpopuli.com/deportes/ (most recently consulted: 23.1.2014).

(8) Según contó al programa *Lo que yo te diga*, de Radio El País, fue visitado por unos cacos que le **desvalijaron** la casa y él ni se enteró. *El País, 02.12.1986.*

(9) Ein 29-Jähriger aus dem Bezirk soll die Raiffeisenbank in Oggau **überfallen** haben. Ein Teil der Beute wurde sichergestellt. *BVZ13/JUN.01488 Burgenländische Volkszeitung, 13.06.2013.*

III

The e-lexicography challenge

Reverse search in electronic dictionaries

Álvaro Iriarte

1. Introduction

In traditional dictionaries (paper dictionaries, digitized dictionaries, human-readable dictionaries), we can look for information about a word covering the lemmas included in the nomenclature, usually listed alphabetically. After we find the searched word, we can get, in the microstructure, information on:

- meanings of the lemma and their definitions;
- phonetic transcriptions (or respelling);
- grammatical information (morphological, syntactic, semantic, lexical, etc.);
- encyclopaedic or cognitive information, specialised technical field, etc.;
- pragmatic or rhetorical information;
- combinatory possibilities (lexical or syntactic combinations);
- examples and citations;
- etymological information.

But what happens when we don't know the word we're looking for? How can a traditional dictionary help me when I want to know, for example:

- What is the name of 20 units box where are sold beer bottles?
- What euphemism or politically correct term I can use instead of a dysphemism?
- Where do I get information about the verb that combines with the word *passeio* ('walk') to express the sense "to go for a walk"?
- etc.

If the dictionary only presents that information under the entries *grade* ('crate'), the euphemism that you don't know, or the verb *dar* ('give'), I won't be able to find it, because I don't know that *dar* is the word I must use to express that sense ("take a walk" / "to go for a walk"). That's exactly the information I ignore. But that is what happens in most dictionaries: We found the combination *dar um passeio* under the entry ***dar***.

On the other hand, if we register, under the entry ***dar*** a collocation like "dar um passeio", we must also register many other possible combinations:

> **dar** a benção, **dar** a opinião, **dar** a palavra, **dar** a volta, **dar** acordo de si, **dar** alma a, **dar** ares de, **dar** asco, **dar** autorização, **dar** boleia, **dar** cabo de, **dar** carta branca, **dar** certo, **dar** como aberta a (conferência), **dar** conhecimento a, **dar** conta de, **dar** contas de, **dar** corda, **dar** entrada, **dar** entrada em, **dar** erros, **dar** faísca, **dar** faltas, **dar** fé de, **dar** feriado, **dar** forças, **dar** ganas, **dar** instruções, **dar** jeito, **dar** largas a, **dar** licença, **dar** medo, **dar** nas vistas, **dar** o nó, **dar** o sol, **dar** ordem para, **dar** os parabéns, **dar** ouvidos, **dar** patadas, **dar** pena, **dar** permissão, **dar** pontapés, **dar** preguiça, **dar** razão, **dar** saudades, **dar** um abraço, **dar** um beijo, **dar** um conselho, **dar** um golpe de coragem, **dar** um grito, **dar** um nó, **dar** um passo, **dar** um salto, **dar** um salto lá, **dar** um som, **dar** um suspiro, **dar** uma ajuda, **dar** uma bofetada, **dar** uma conferência, **dar** uma corrida, **dar** uma injecção, **dar** uma mão, **dar** uma negativa, **dar** uma ordem, **dar** uma palavrinha, **dar** uma queda, **dar** uma sugestão, **dar** uma vacina, **dar** vontade, **dar** (dois dedos de) conversa, **dar** um filme, não **dar** uma para a caixa, etc.

Which we can also join close to 570 set phrases beginning with the word *dar* (plus 46 initiated with "*não dar*") recorded by Simões (1994) in his *Dicionário de Expressões Populares Portuguesas,* or 570 set phrases recorded by Ramalho (1985) in his *Dicionário Estrutural, Estilístico e Sintáctico da Língua Portuguesa.*

These lexical combinations are unproblematic when we use the dictionary as a tool for decoding. It's not especially hard to discover the meaning of the combination *dar um passeio* if you look it up in a dictionary.

The problem arises in a particular way when we intend to proceed to codification or text production, because the combinatory possibilities vary from one language to another (and, in general, the lexical-syntactic and pragmatic uses).

What happens when I want to express the idea "take a walk" or "to go for a walk" in Portuguese? Let's just say that I do not know the construction *dar um passeio*, but I do know the Portuguese word *passeio* is "walk" in English. Will dictionaries help me?

It is in the entry **passeio** that the user should be able to find the information. A good lexicographic treatment of words implies collecting their different combinatory possibilities. These combinatory possibilities of words are determined as much co-textually (that is, by the linguistic context) as pragmatically and contextually. I mean, the meaning of a word or group of words can be defined or delimited by context, or by other lexical units, which, along with that word, compose the syntagma.

Idiomatic expressions (such as "it's raining cats and dogs") should be collected in the syntagmatic part of the lexicographic article[109], since it is not always possible to associate a multi-word expression to a concrete meaning of the words it is composed of. But, what happens in the case of lexical collocations? How should collocations be collected in dictionaries?

Take another example:

> *tirar uma conclusão* ('draw a conclusion')
> tirar uma fotografia ('take a picture')

[109] In the case of "it's raining cats and dogs", in encoding dictionaries, should be collected under the entry **rain** (but I will not talk about it in here).

The way most dictionaries are made, users may only find this information under the entry ***tirar*** (that's just exactly the information they ignore!):

> **tirar** [. . .] v. [. . .] **II.** Como verbo suporte de predicação, combina-se com nomes ... 1. Actos físicos ≃ FAZER. ... *Tirar uma fotografia* ... **2.** Actos morais ≃ OBTER. *Tirar conclusões* (Academia).
>
> **tirar** [. . .] v. **31.** captar (imagem), . . . fotografando, . . . : *t. uma foto de uma pessoa.* (Houaiss)
>
> **tirar** V. t. d. [. . .] **15.** Fazer (uma fotografia [2]); *Fique aí quieto, vou tirar a fotografia.* **16.** Fazer tirar, parar para tirar (uma fotografia [2]): *Aprontou-se toda para tirar o retrato.* [. . .] ([Aurélio]Ferreira).
>
> **TIRAR** v. tr. [. . .] Tirar (alguém) o retrato, fazer-se retratar: *Fui tirar o retrato para a carteira de identidade.* || Tirar o retrato a alguém, fazer-lhe o retrato: [. . .] ([Caldas] Aulete).
>
> **Tirar**, V. t. [. . .] Derivar: *tirar conclusões.* [. . .] ([Cândido] Figueiredo).

Only in the dictionaries *Houaiss* and *Aulete Digital* we have information under the entry ***fotografia*** (but nothing under the entry ***conclusão*****).** In the *Dicionário Houaiss*, in the grammatical information under the entry ***dar***, we can read:

> **a.3**) por sua importância, diversas acepções de *dar,* usado como verbo-suporte, estão registadas no corpo deste verbete; diversas outras devem ser procuradas pelo substantivo que faz parte do objeto direto, como de hábito no restante dicionário. (Houaiss, *s.v.* **dar**).

Another exception is the *Dicionário Básico da Língua Portuguesa,* which has a good resolution of the issue, because we have information about the collocation *tirar uma conclusão* under the entry **conclusão:**

> **tirar**... [...]
>
> **S. 8.** Tirar + nome, sentido VIII, equivale a um verbo simples: CONCLUIR (FRASE 1), [...] (Vilela).
>
> **conclusão**... [...]
>
> [...] // (pessoa) **tirar conclusões:** (5) — *Que conclusões podemos tirar da sua atitude?* • (6) — *Não quero tirar conclusões erradas do caso.* [...]
>
> **S.** • *Tirar conclusões* (frases 5, 6) tem como sins.: CONCLUIR, TIRAR ILACÕES OU DEDUÇÕES ... (Vilela).

2. Reverse search[110]: onomasiological dictionaries and encoding dictionaries

An onomasiological dictionary is the opposite of a traditional dictionary. When we use an ordinary dictionary, we have a word in our head and we want to look up its meaning. When we use an onomasiological dictionary, we have a concept in our head, and we want to look up which word, or words, best expresses it. In these

[110] A reverse dictionary is not the same as a reverse-order dictionary. In a reverse-order dictionary the alphabetical sorting is done from right to left. They let the user browse the dictionary searching by the end of the word, instead of its beginning.

dictionaries we can search an idea or concept in a descriptors structure or a structured list of concepts sorted by subjects (summary tables), together with a list of hypernyms or broader terms (categories, general ideas) that lead the reader to the searched word.

The major disadvantage of this sort of dictionaries is that the organization and classification of information varies from dictionary to dictionary because the knowledge's organisation varies from author to author (Béjoint, 2004: 15).

There is a long tradition of onomasiological dictionaries production for the European languages, especially in the nineteenth and twentieth centuries (see some examples in Sousa, 1995, s.v. diccionario ideológico).

Another type of reverse dictionary is the encoding dictionary. It is important to distinguish between encoding dictionaries and onomasiological dictionaries. In encoding dictionaries you don't look for an idea to find a word. Encoding dictionaries supply you with information about the co-text (that is, about the lexical, syntactic or semantic combinatory capacity) and the context (information of a pragmatic kind). So, encoding dictionaries supply you with information about the co-text and the context of certain lexical choices in order to transmit certain concepts.

In an onomasiological dictionary, you should look for the information in a kind of structure that is organized by subjects through hyperonyms or generics. These allow us to get to the word we are looking for. So, to find a word that means *forte* ('strong') applied to drinks, we should look under the entry **intensidade** ('intensity') or something like that. In onomasiological dictionaries the lemma is this synonym or hyperonym (**intensidade**):

> **intensidade:** forte [bebidas]

In encoding dictionaries the lemma is a word, or set of words, about certain information we are looking for. This is the opposite of what happens in onomasiological dictionaries, in which you look for an idea or a concept from its synonym or hyperonym. In an encoding dictionary, the user should look for the lexeme that expresses the idea of *forte* ('strong') applied, for example, to the word *café* ('coffee') under the entry **café**. In encoding dictionaries you search words, not concepts. You are using the "word" as a unit, as lemma, and you keep the traditional alphabetical order:

> **café** ... [muito intenso] > *forte*
> **ódio** ... [muito intenso] > *mortal, figadal*
> **passeio** ... [realizar] > *dar um passeio*

Onomasiological dictionaries are paradigmatic dictionaries while encoding dictionaries are paradigmatic and syntagmatic dictionaries. The main limitation of the onomasiological dictionaries, and in general of all the so-called paradigmatic dictionaries (onomasiological dictionaries, synonym and antonym dictionaries, and so on) is that they just enumerate synonyms, with no examples, no exact indication on contexts of usage, etc: "...en gran parte, son poco satisfactorios, ya que se limitan a dar una simple enumeración de sinónimos, sin indicaciones exactas sobre denotación, connotación, situaciones de uso, etc., y, sobre todo, sin ejemplos." (Haensch, 1982: 178).

3. Implementation of reverse search functionalities in electronic dictionaries

The reverse search capabilities transform electronic dictionaries not only in an ideological or conceptual dictionary but also in an encoding dictionary. It allows us to find not only the word that corresponds to an idea (like onomasiological dictionaries), but also which word we can associate to another to express an idea (like encoding dictionaries).

The reverse search capability in electronic dictionaries can help overcome some of the limitations present in onomasiological dictionaries and encoding dictionaries on paper. This feature can be more than a simple search tool. I can imagine the possibilities of reverse research into new lexicographical products, developed with more scientific rigor and with a more systematic microstructure. In this electronic dictionaries it should be possible to search anywhere on the macrostructure and the microstructure.

With a simple search tool in the electronic dictionary, we can find related lexical units (synonyms, quasi-synonyms, hyperonyms/hyponyms, meronyms/holonyms and other lexical-conceptual relations). But more importantly: this search functionality of the system can use these same relations to show results using not only the words entered by the user, but also these lexical-semantic relations just mentioned. We're talking about what we call "ontological research", as we'll see below (§ 4.3).

The present project aims to transform the *Dicionário Aberto* (DA)[111] into an advanced encoding dictionary. The intention is that, using the reverse search capability in the *DA*, users can search for related lexical units (synonyms, quasi-synonyms, hyperonyms, hyponyms, meronyms, holonyms, etc.) from one or more words.

To make this possible, it will be necessary that the search system available to users works not only with the terms input by the user (and their instances in the macrostructure and microstructure), but also with a semantic structure that provides lexical-conceptual relations between the terms introduced and other terms. These relations will be cross, and proximity measures will be calculated, thereby being possible presenting a set of results sorted by relevance.

4. The advanced search in the *Dicionário Aberto*

The *DA* application Web has evolved, incorporating various types of search: search by entry of article, search in parts of lemma (beginning, middle or end), or

[111] The *Dicionário Aberto* is available online for consultation and for automatic extraction of information in http://www.dicionario-aberto.net, but also for local use, open and free. The project began in June 2005, with the transcript of the 1913 edition of the two volumes of the *Novo Diccionário da Língua Portuguesa,* de Cândido de Figueiredo. For more information about the *Dicionário Aberto*, see Simões & Farinha (2011).

reverse search, among other capabilities. At this moment we are working in a semi-automatic system for extracting synonyms, hyperonyms/hyponyms, meronyms/holonyms that serves to explore the lexical-conceptual relations between terms entered in search (Simões, Iriarte & Almeida, 2012).

Like any electronic dictionary, the *DA* has a search tool. The simple search tool allows the user to find words related to the searched lemma (synonyms, hyperonyms or meronyms), as well as to obtain a set of words orthographically similar, useful for when you don't know the exact spelling of a lemma. The *DA* also has a feature that allows the user to "flip through the dictionary", that is, to browse the dictionary entries sequentially, watching the words near in alphabetical order.

More interesting are the advanced search capabilities, which can turn the dictionary into a real onomasiological dictionary and encoding dictionary at the same time. Using the reverse search and ontological search capabilities, users can search for related lexical units (synonyms, quasi-synonyms, hyperonyms, hyponyms, meronyms, holonyms, etc.) and co-occurring words from one or more words. These advanced features, available for registered users, can be very useful tools for linguists and researchers in Natural Language Processing.

4.1. Search in parts of lemma (beginning, middle or end)

The advanced search allows you to search lemma fragments (beginning, middle or end) by selecting, in the interface, the word "Prefix", "Infix" or "Suffix". These terms, that we hope will be substituted, are not the most appropriate, because the results may not correspond to these morphological categories. Initially, however, we found it clear enough for the general public, and, especially, short enough to be used comfortably in the interface.

We may find, for example, words ending in *–dade* (*abundidade; aceitabilidade; acessibilidade; aceitabilidade; acerbidade; acessibilidade; acetosidade; ...*), words beginning with *pre–* (*pre...; pre-romano; preá; preadamita; preadivinhar; preagónico; prealegar;...*) or words, in the original spelling of 1913, containing *–mm–* (*accommodação; accommodadamente; accommodadiço; accommodamento; accommodar; accommodatício; accommodável; ...*).

This search capability can be an important tool for morphology studies (Millán, 1999). We can imagine, for example, its usefulness for studying the affixes productivity, such as diminutive suffixes (*–inho, –ito –ino*, etc.); the productivity of certain suffixes in scientific terminology (*–ato, –eto –ito)*; the productivity and real synonymy of suffixes like *–dade/ –ção/ –são, –ança/ –ância*; etc.

The dictionary may prove to be an important resource for linguistic research, and for the development of grammars and other dictionaries, since it allows us to download the results of these searches.

Here's an example on the productivity of affixes: Can all the Portuguese adjectives ended with the suffix *–vel* (like *amável*) form adverbs ending in *–mente* (like *amavelmente*)?

From the *DA* you can download lists of adjectives formed with the suffix –vel and lists of adverbs ending in *–mente* (or, better than the suffix, the form *–velmente*):

…
agitável
aglutinável
agradável
agradecível
agricultável
ajuntável
alcançável
alcoolizável
alheável
aliável
alienável
alliável
alterável
amável
amigável
…

…
abominavelmente
admiravelmente
adoravelmente
afavelmente
affavelmente
agradavelmente
amavelmente
amigavelmente
amoravelmente
aprazivelmente
civelmente
comendavelmente
commendavelmente
compativelmente
compreensivelmente
…

The results can be easily aligned, thus answering the first question (Can all the Portuguese adjectives ended with the suffix –*vel* form adverbs ending in –*mente*?), which then can help us verify hypotheses that explain which can and which cannot:

agitável	…
aglutinável	…
agradável	agradavelmente
agradecível	…
agricultável	…
ajuntável	…
alcançável	…
alcoolizável	…
alheável	…
aliável	…
alienável	…
alliável	…
alterável	…
amável	amavelmente
amigável	amigavelmente
…	…

4.2. The reverse search

As we said, the reverse search capability turns the DA into a real onomasiological dictionary and encoding dictionary. The DA can thus help answer questions like, for instance:

- *"O que acontece à água com o frio?"* ('What happens to the water with the cold?'). Selecting reverse search and typing *água* ('water') and *frio* ('cold') we have the following results: *desnevado, fresco, gelo* e *neve;*

- "Quem é o médico dos olhos?" ('Who is the eye doctor?'). Selecting reverse search and typing *médico* ('doctor') and *olhos* ('eyes') we have the following results: *oculista, oftalmiatra, oftalmologista*;

- "What is the word used in Portuguese to mean 'improve the hardness of metal'? Selecting reverse search and typing *endurecer* ('to harden') and *metal* we obtain as a result the entry **temperar** ('to temper').

Work is being done to ensure that research can be done using not only word forms occurring in the microstructure but also the lemmas corresponding to these word forms, using a morphological analyser.

4.3. The ontological search

I believe that what we call "ontological research" is much more interesting and promising. This research is based on an ontology constructed dynamically (and, therefore, fully automatically).

Unlike what happens with the reverse research, in ontological research, some results are entries that don't contain any of the terms entered by the user in the search window. Take, for example, the following results after introducing the term *mamífero* ('mammal') in the ontological search window:

otária[1] — *f.* ; Género de plantas asclepiadáceas. Espécie de foca, de orelhas bem visíveis....

lontra[1] — *f.* ; Pequeno quadrúpede carnívoro, da fam. das martas.*M. Prov. trasm.*; Pescador...

golfinho[1] — *m.* ; Grande peixe marítimo, carnívoro, da fam. dos cetáceos. *Heráld.*; Represen...

macoco[1] — *m.* ; Animal do Congo, talvez espécie de antílope.

capiguará[1] — *m. Bras*; Espécie de lontra. (Do guar.)

...

We don't find the term *mamífero* ('mammal') in these definitions.

In extraction of lexical-conceptual relations (Simões, Iriarte & Almeida, 2012), some structures present in the definitions were used. Here are some exemples:

– *o mesmo (ou melhor) que ...* ['the same (or better) than'] > Synonymy (SYN);
– *que não é ...* ['that is not'] > Antonymy (ANT);
– *espécie de ...* ['kind of'] > Hyponymy (HIPO);
– *que tem por tipo...* ['which are the type'] > Hyperonymy (HIPER);
– *cada uma das partes que formam ...* ['each of the parties that form'] > Meronymy (MERO);
– *composto por...* ['compound of']> Holonymy (HOLO).

Thanks to the regularity of the structure of the definitions, we can establish a set of rules or patterns (Hearst 1992) using sequences of words that can be found in the definitions. There are high chances of the word that follows the pattern having a lexical-conceptual relation with the lemma of the respective definition.

To the ontological search capability we also use calculated relations (for example, using the transitivity of the relation of hypernym). Thus, mathematical rules were defined for completion of the ontology rules, inferring new relations from the initial relations:

- a SYN b ⇒ b SYN a (symmetry property between synonyms: if *a* is a synonym for *b*, then *b* is also synonymous for *a*). This relationship is very productive because, in many situations of true synonyms, lexicographers collected the relation of synonymy only in one of the entries of the words involved. Thus, this information can be retrieved for other entry.
- a HIPO b ∧ c HIPO b ⇒ a COHIPO c (If two words, *a* and *c,* are hyponyms of the same word, *b*, then they, *a* and *c,* are co-hyponyms). Relation of co-hyponymy can be calculated from relations of hyponymy.
- a HIPO b ∧ b HIPO c ⇒ a HIPO c. Transitivity of the hierarchical relationships, such as the hypernym or hyponymy, allows one to search for generic terms using a very specific term, or search for specific terms using very generic terms. Think, for example, in a search using the term *animal* ('animal'). It is unlikely that you will find entries for animals. But, by transitivity, we can find other classes as *mamífero* ('mammal') or *peixe* ('fish'), which are hyponyms of "animal". We can find hyponyms terms of *mamífero* ('mammal') or *peixe* ('fish') that are also hyponyms of *animal* ('animal').

We are aware that some rules can be problematic (for example, the case of synonyms and quasi-synonyms). In any case, we believe that a set of possible false synonyms is preferable to no results or drastically reduce the number of ontological relations resulting. Ontology becomes more useful when there is great diversity in the existing relations.

5. Conclusions

Running experiments and implementing algorithms on the DA has been very rewarding. We are convinced the DA will be an excellent tool that cannot only be used as a traditional dictionary, but also as a resource for tasks of natural language processing, as a tool to assist in hypothesis testing for linguistic research and assist in the development of grammars and other dictionaries.

References

Aulete, F. J. Caldas (1987). *Dicionário da Língua Portuguesa Caldas Aulete.* Rio de Janeiro: Editora Delta. [5a edição brasileira, revista, actualizada e aumentada por Hamílcar de Garcia e Antenor Nascentes].

Béjoint, H. (2004). *Modern lexicography: an introduction.* Oxford: University Press.

Casteleiro, J. Malaca (coord.) (2001). *Dicionário da Língua Portuguesa Contemporânea da Academia das Ciências de Lisboa.* Lisboa: Academia das Ciências de Lisboa/Editorial Verbo.

Ferreira, A. Buarque de Holanda (1999). *Novo Aurélio Século XXI: O Dicionário da Língua Portuguesa.* Rio de Janeiro: Nova Fronteira.

Figueiredo, C. de (1982). *Grande Dicionário da Língua Portuguesa.* Lisboa: Livraria Bertrand.

Haensch, G. (1982). "Tipología de las obras lexicográficas". In Haensch, G.; Wolf, L.; Ettinger, S.; Werner, R. (eds.). *La lexicografía. De la linguística teórica a la lexicografía práctica.* Madrid: Gredos. 95-187.

Hearst, M. (1992). “Automatic acquisition of hyponyms from large text corpora”. *Proceedings of the Fourteenth International Conference on Computational Linguistics,* Nantes, France. Volume 2. 539–545.

Houaiss, A. (2001). *Dicionário Houaiss da Língua Portuguesa.* Rio de Janeiro: Objectiva.

Martínez de Sousa, J. (1995). *Diccionario de lexicografía práctica.* Barcelona: Bibliograf.

Millán, J. A. (1999). “Zigzag, gong, ping-pong, iceberg. donde se descubre que hay diccionarios inversos, y su utilidad manifiesta para el progreso de la humanidade” [on-line]. http://jamillan.com/inverso.htm [Access date: 1 July. 2013]

Ramalho, E. (1985) *Dicionário Estrutural, Estilístico e Sintáctico da Língua Portuguesa.* Porto: Lello & Irmão Editores.

Simões, A.; Farinha, R. (2011). “Dicionário Aberto: Um novo recurso para PLN”, *Vice-versa* 16, 159–171.

Simões, A.; Iriarte, Á; Almeida, J. J. (2012). “Dicionário-aberto – a source of resources for the portuguese language processing”. In Caseli, H.; Villavicencio, A.; Teixeira, A.; Perdigão, F. (eds.) *Computational Processing of the Portuguese Language, Lecture Notes for Artificial Intelligence,* Berlim: Springer, 7243, 121–127.

Simões, G. A. (1994). *Dicionário de Expressões Populares Portuguesas.* Lisboa: Dom Quixote.

Vilela, M. (1991). *Dicionário do Português Básico.* Porto: Edições Asa.

Theoretical and methodological foundations of the DICONALE project: a conceptual dictionary of German and Spanish

Meike Meliss & Paloma Sánchez Hernández

1. Introduction

The research project DICONALE-online[112] concerns the development of an *online* dictionary of verbal lexemes, onomasiological and bilingual-bidirectional in nature, and covering the German and Spanish languages. The dictionary is intended for those who are learning German or Spanish as a foreign language at an advanced level (B2),[113] and as a pedagogical dictionary it is conceived especially as a resource for those actively producing texts.[114] The project is a response to studies which have shown that conventional dictionaries (both monolingual and bilingual), in both *print* and *online* format, do not satisfy the specific needs of users involved in the production of texts (cf. Haß 2005, Fuentes Morán 1997: 84, Meliss 2013a, 2014c, 2014d). Thus, it arises from the need to fill the current gap in German-Spanish bilingual lexicography, and is intended to create a dictionary with a conceptual-onomasiological MACROSTRUCTURE which offers a more appropriate kind of help here, with the possibility of searching for forms of expression according to context, and hence differing from a traditional, alphabetic-semantic orientation from the

[112] This study forms part of the following DICONALE-online research projects: 'Development of a conceptual bilingual dictionary of German and Spanish: an online resource' (MINECO-FEDER FFI2012-32658), and DICONALE: 'Studies towards the development of a conceptual dictionary of verbal lexemes in German and Spanish' (Xunta de Galicia: 10PXIB 204 188 PR), led by Meike Meliss of the University of Santiago de Compostela, and also to the research group GI-1920 and the "Rede de Lexicografía" (Relex) (Consellería de Cultura, Educación e Ordenación Universitaria da Xunta de Galicia CN2012/290).

[113] According to the common European framework of references.

[114] In the area of DaF and ELe the following learners' dictionaries exist, among others: For German: Kempcke 1999, Pons-DaF 2004 (print+digital), Duden-DaF ^{2}2010, Wahrig-DaF 2008, Götz, D. et al.: Langenscheidt-DaF ^{3}2010; For Spanish: Diccionario de español para extranjeros de SM 2002, Diccionario de Alcalá 1995 and El Diccionario Salamanca 2007.

very outset. This perspective represents a challenge in the context of German-Spanish, in that currently no lexicographic works of this type exist.[115]

The project is based on a **modular-integrative** and **bidirectional** model of description, with special interest in **paradigmatic** and **syntagmatic** aspects, from both interlingual and intralingual perspectives. In this way, the special focus on the systematic presentation of syntagmatic structures is of assistance to the user in the correct use of a lexeme. We must note here the absence of dictionaries, above all of Spanish as a foreign language, offering sufficient syntagmatic information for use in situations of production.[116]

In this study we will present in detail the most relevant innovations of the DICONALE model, highlighting various specific elements, including the type of user (DaF and ELe), its onomasiological organisation, the empirical basis of the data, online access, the descriptive model which focuses the description of paradigmatic and syntagmatic relations in the dictionary, and the contrastive perspective.

2. Innovative aspects of DICONALE

2.1. For whom is DICONALE? Typology

DICONALE is intended to be a **pedagogical bilingual dictionary** which addresses in particular the needs of **production in the foreign language from level B2 onwards**. Hence, the following questions by users might motivate the use of the dictionary:

Selection: Which of the possible lexical resources of the target language is best suit to the communicative situation?

Production: What information is especially relevant from a contrastive point of view: divergences between the user's language and the target language.

These needs can be summarised through taking a closer look at the information offered to the user for the **selection** of one lexeme or another, and at a detailed description of the combinatory potential for a lexeme's appropriate use. In studies of learner dictionaries, both German and Spanish monolingual dictionaries and bilingual ones of these two languages, it has been shown that little information required in situations of production is given. Thus, for example, in DAF monolingual dictionaries there is in general a need to attend more closely to the parameters that define the combinatorial potential of a lexeme, as well as to provide more information on the distinguishing parameters between semantically close

[115] Currently, for German and Spanish the only works available with this perspective are the visual, onomasiological dictionary of Álvar Esquerra: Duden/Oxford 1993 and the multilingual Pons 2008.

[116] In recent years there has been a great deal of interest in syntagmatic dictionaries, given the scant attention paid to this kind of information in traditional dictionaries. Of note are those dictionaries aimed at offering information on aspects of syntax (verbal complementation, semantic and syntactic valency, collocations etc.), such as, for Spanish, Cuervo 1953/1998, Seco et al. ^{3}1999, Bosque 2004, and for German, Helbig/Schenkel 1969, Engel/Schumacher 1978, Schumacher, H. et al.: Valbu 2004, Duden 2 ^{9}2010, Quasthoff 2011, and for English Herbst et al. 2004.

lexemes, in order to facilitate the selection of one or the other, according to the context (Meliss 2014c). Spanish monolingual dictionaries for educational purposes largely lack sufficient information for situations of production (Meliss 2014d). Moreover, bilingual dictionaries of German and Spanish[117] have also shown deficiencies in syntagmatic information (Morán Fuentes 1997, Meliss 2011: 279 ff., Meliss 2013a, Model 2010), effectively bringing into question their true bilingual-bidirectional nature (Engelberg/Lemnitzer ^{4}2009: 129ff. and 210f.).[118] Lacking the necessary information, our "user" increasingly has recourse to *online* lexicographic resources, affording access both to other monolingual dictionaries of German and Spanish, and to works of a specific nature (paradigmatic and syntagmatic dictionaries, construction dictionaries etc.). Despite the many advantages which these *online* resources offer, they are not always adequate for the type of user that interests us here, given that they are not conceived of specifically for users of an L2, and thus the information they contain is in general too complex for contexts of DaF and/or ELe (Meliss 2013b). In order that the user of the dictionary finds the right information for his or her purpose, DICONALE aims to offer, in an explicit way, contrastive information that can provide the relevant information, with the aim of making it possible to select, from the variety of possibilities, one or another equivalent form in the target language.

2.2 Conceptual-onomasiological focus and paradigmatic structuring: background and new proposal

Unlike semasiological dictionaries, onomasiological dictionaries are oriented towards production (Reichmann 1989, Martín Mingorance 1994). The classic onomasiological dictionaries in German (Wehrle/Eggers 1961, Dornseiff 1965, Dornseiff/Quasthoff ^{8}2004), like Casares' (^{2}2007) ideological dictionary of Spanish, while being the most significant works of onomasiological lexicography in their respective languages, nevertheless exhibit a lack of transparency in their structuring and contain insufficiently detailed information for their use in the production of text. Although they offer various possible expressions when one is looking for a specific *signifiant* for a *signifié*, they do not offer the user enough scope in the the selection of a term, and thus make it necessary to consult other dictionaries in order to verify a meaning and its combinations (Meliss 2005: 65ff., 2011: 293ff.). From the user's point of view, these drawbacks mean that consulting such dictionaries is not very productive. Such deficiencies justify the need to find new lexical avenues from the onomasiological perspective. In this context, we might note, among others, proposals for organising parts of the lexicon in line with the theoretical assumptions of structural semantics and the theory of lexical-semantic fields. Of special mention here are the studies of Coseriu (1977 and especially ^{2}1986), Geckeler (1971/^{3}1982) and Trujillo (1970) which, along with renewed interest in the 90s (Dupuy-Engelhardt 1990, Geckeler 1993, Lutzeier 1993, Wotjak 1992), have made possible, since the 70s,

[117] For general aspects of German and Spanish lexicography: Fuentes Morán (1997), Werner (1998), Haensch/Omeñaca (^{2}2004) and Hausmann (1991).

[118] For information on the user profile and a description of the situations of use, see Egido/Fernández/ Franco, this volume.

numerous lexicographic studies, both monolingual and bilingual. The starting point for these studies is the paradigmatic structuring of a series of lexemes, mutually related through shared semantic features, that lexicalise related concepts and that, in part, combine the paradigmatic structuring with syntagmatic information.[119] Since the 80s, some onomasiological works have appeared which present a systematic description of the lexicon and combine paradigmatic structuring with syntagmatic information based on the theory of valence (Schumacher et al 1986, Harras et al. 2004, 2007) and which are especially of use in the area of DaF (Schreiber et al. ^{2}1990). Recent lexicographic studies also exist in Spanish, in *online* format, that offer paradigmatic and syntagmatic information (ADESSE, DICE).

For contrastive studies, the onomasiological perspective provides the *tertium comparationis* through conceptual units. This approach reflects work in the field of cognitive linguistics (Blank/Koch 2003). Also, it finds its application in the needs of foreign language learning and the practice of translation.

The DICONALE model is, according to the classification of Haß-Zumkehr (2001: 269), a distinctive onomasiological dictionary, given that it provides, in addition to the structuring of concepts, information on the use and meaning of terms associated with these concepts. With greater specification still, conceptual fields are themselves divided into subfields of different degrees. Lexemes associated with these are differentiated through the existence of distinctive semantic features and different argument structures. In this first phase, we are interested in simple and affixed verbal lexemes, and also –although to a lesser extent – plurilexematic forms (Sánchez Hernández 2013a). These conceptual fields are analysed using the same parameters and with the same descriptive model. Hence, the broad differences and similarities between fields can be observed. In this sense, the user conducts searches using concepts rather than lemas and their meanings, in that the user only has access to the meanings once he or she has selected the possible lexicalisations for the concept in question. As Haß-Zumkehr notes (2001: 264), the onomasiological structuring of the lexicon is completed with the structuring of the elements, attending to their semantic relations, that is, synonymy, antonymy, hyponymy, hyperonymy, etc. and this constitutes an important parameter of the model of description.

The development of an onomasiological dictionary along the lines proposed for DICONALE represents a major innovation in German and Spanish lexicography, and predetermines the resulting macrostructure of the work, given that it makes use of certain specific concepts and permits a structuring in fields and subfields with different degrees. Onomasiological access will be complemented by semasiological access for those who opt for an alphabetically ordered search.

[119] Regarding studies from an onomasiological perspective, and with a starting point of the lexical structuring of German and Spanish contrastively through paradigmatic principles in combination with syntagmatic information, one might mention, among others, Hernández Eduardo (1993), González Ribao/Proost (2014), Meliss (2005, 2006, 2014a) and Sánchez Hernández (2010, 2012).

2.3. Methodology: Empirical basis

The need to use linguistic corpora and data on frequency of use as a means of selecting information[120] can be seen in a pilot study, based on the corpus of DEREKO, which looked at the syntagmatic information relating to different meanings of the German verb *abhören*[121] (Meliss 2014c).[122] In order to follow a consistent approach in the selection of the information which DICONALE is intended to provide, it is necessary to work with an empirically valid method. For this, we compiled our own corpus, using journalistic texts drawn from DEREKO, CREA and Web-Corp. In this way, the corpora for our two languages were guaranteed to be of maximum comparability (cf. González Ribao 2014).

2.4. Online access

The specific characteristics of DICONALE, especially the onomasiological focus combined with a semasiological focus, the bidirectionality of this bilingual dictionary, together with a complex, modular descriptive model for both languages in contrast, clearly exceeds the possibilities of representation of a dictionary in *print* format. Like Engelberg/Lemnitzer (^{4}2009: 220), we believe that the future of dictionaries lies in electronic access, and indeed a shift of habits in the use of reference tools has recently been observed here, with some recent questionnaire-based studies on the use of dictionaries (cf. Domínguez Vázquez et al. 2013) reporting on the growing use of all manner of *online* reference works. The advantages of this type of access are diverse, and above all, in the area of bilingual lexicography it appears that reference works in *print* format are destined to become relics from another era, at the same time as specific works, such as paradigmatic dictionaries of synonymy and antonymy, plus syntagmatic dictionaries (cf. Meliss 2013b) are available in ever greater numbers on the internet. However, it is evident that not all dictionaries and other *online* resources are appropriate and adequate for

[120] Cf. Bubenhofer et al. (ed.) (2010), Lemnitzer/Zinsmeister (2006), Renouf (ed.) (2009), Schmidt (ed.) (2012).

[121] Some correspondences in Spanish are: auscultar, escuchar, examinar, controlar, interceptar, intervenir, etc.

[122] The analysis of the corpus here, compared to the information found in five learner dictionaries of German as a foreign language revealed that not only was there considerable difference in the information provided by the five dictionaries, but that there was a discrepancy with respect to the data from the corpus analysis. In the analysis, for example, structures relating to one meaning and its argument structure (21% of documented cases in the corpus) and another relating to the realisation of the verb in nominal form (23%) were almost entirely absent from the dictionaries, not even seen implicitly through examples; on the other hand, some meanings for which no examples were registered in the corpus were nevertheless present in all five dictionaries. The following pedagogical dictionaries of German as a foreign language have been used: Götz et al.: Langenscheidt (^{3}2010), Kempcke (1999), Pons print (2004) and online, Duden (^{2}2010), Wahrig (2008).

all users and for all possible circumstances and situations.[123] This has been shown, for example, in a study of *online* lexicography and its use in the area of DaF by Meliss (2013b).[124] A central element of attention with *online* dictionaries tends to be related to the added value of the experience compared with the traditional format, such as their multimedia, interactive, modular and hypertextual character (Storrer 2010: 155; Tarp 2012: 253; Haß/Schmitz 2010: 6ff.). For the bilingual area in general, and DaF/ELe in particular, it is necessary to create resources adapted to users and their specific needs, given that those lexicographic portals currently available are not in general conceived of for the use of non-natives, which often makes access to information in L2 contexts difficult (cf. Meliss 2013b, Müller-Spitzer/Engelberg 2013). DICONALE seeks to embrace these enormous technological possibilities and create a reference work adapted to the needs of our specific users.

2.5. The structure of the dictionary: macro-, micro- and mediostructure and the modules of description

The descriptive model through which we aim to codify and interpret the information for DICONALE at the inter- and intralingual level encompasses five **levels. Levels 1** and **3** provide conceptual structuring and the organisation of lexical material in lexico-semantic fields (→ macrostructure). Through **level 2** lexicological information is codified in detail (→ microstructure) and this will then form the basis for the different types of analysis in **levels 4** and **5** (→ mediostructure). Taking the lexicological information from the five levels of description as a basis, an effective system of access to the dictionary's information is achieved, one which provides the user with pertinent modular information through different search options. Some relevant aspects of this will now be illustrated, using examples from the conceptual fields AUDITION and COGNITION. **Level 1** contains lexemes corresponding to the ten conceptual fields (CC) and their different (sub)fields (SCC).

This level, along with **level 3** into which it leads, both to a conceptual subclassification of other degrees and to the formulation of the different lexico-semantic (sub)paradigms (SP), will form the conceptual MACROSTRUCTURE of the dictionary (cf. *table 1*),[125]

[123] On new search techniques and options, and the need for better lexicographic training in the classroom to optimise resources and avoid the risk of a loss of orientation ("lost in hyperspace") see Haß/Schmitz (2010: 4); also Engelberg/Lemnitzer (^{4}2009: 111).

[124] See the criteria for evaluation, for differentiation, and the criteria for the users' guide in: Engelberg/Lemnitzer (^{4}2009: 73 ff., 220ff.), Storrer (2010), Kemmer (2010) and Klosa et al. (2008).

[125] The metalanguage of the information provided in the dictionary and, therefore, also the metalanguage of all the tables which are presented to the user, can be German, or Spanish interchangeably. Therefore, in this article we have decided not to translate into English the information contained in these tables.

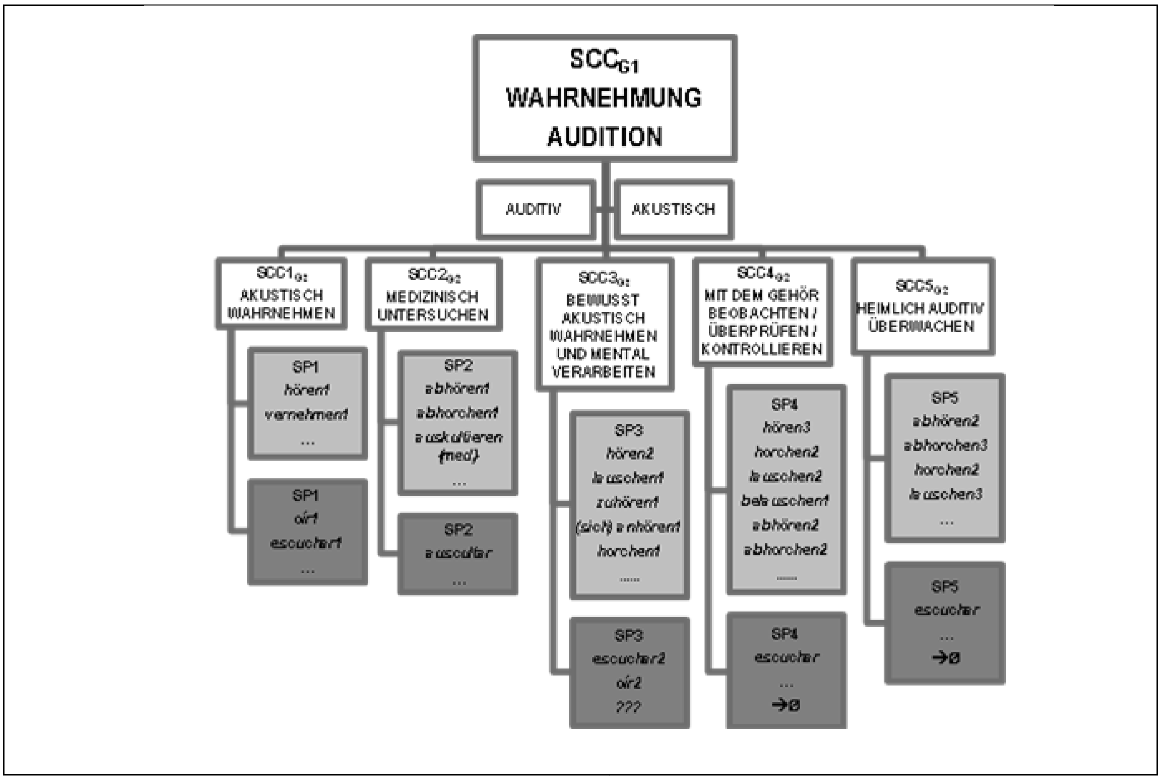

SCC$_{G1}$
WAHRNEHMUNG
AUDITION
AUDITIV
AKUSTISCH
SCC1$_{G2}$: AKUSTISCH WAHRNEHMEN
SCC2$_{G2}$: MEDIZINISCH UNTERSUCHEN
SCC3$_{G2}$: BEWUSST AKUSTISCH WAHRNEHMEN UND MENTAL VERARBEITEN
SCC4$_{G2}$: MIT DEM GEHÖR BEOBACHTEN / ÜBERPRÜFEN / KONTROLLIEREN
SCC5$_{G2}$: HEIMLICH AUDITIV ÜBERWACHEN
SP1 hören1 vernehmen1 ...
SP1 oír1 escuchar1 ...
SP2 abhören1 abhorchen1 auskultieren (med) ...
SP2 auscultar ...
SP3 hören2 lauschen1 zuhören1 (sich) anhören1 horchen1 ......
SP3 escuchar2 oír2 ???
SP4 hören3 horchen2 lauschen2 belauschen1 abhören2 abhorchen2 ......
SP4 escuchar ... →Ø
SP5 abhören2 abhorchen3 horchen2 lauschen3 ...
SP5 escuchar ... →Ø

Table 1: **Level 3:** Conceptual subfields 2nd degree (SCC$_{G2}$), creation of lexico-semantic subparadigms (SP) with a list of corresponding lexemes illustrated through the conceptual subfield (SCC$_{G1}$) AUDITION (sequence).

In **level 2** a codification is made, separately for each language through **four modules**, of the detailed lexicological data of each lema and its different meanings associated with one of 10 conceptual fields (level 1) and subfields (level 3). **Module 1** contains, amongst other data, general information covering formal elements (type of conjugation, suprasegmental features, etc.) and content such as the semantic features of the lema, whereas **module 2** deals with the different meanings of each lema through a semantic description that includes, amongst others, distinctive semantic features and the different paradigmatic relations of sense, together with argument structures and pragmatic information. In **module 3** syntagmatic information is specified through data based on the empirical frequency of each argument and its morphosyntactic and semantic characteristics[126] (cf. Engelberg et al. 2012, Engelberg 2014a, 2014b, Meliss 2014b), as well as offering an equivalent in the contact language for each meaning. In the DICONALE model special importance is given to information relating to the argument structure, given that it constitutes, along with the componential semantic structure, the *tertium comparationis* between the languages in contact.[127] The detailed semantic and morphosyntactic information related to each argument supplies the specific nuances in which many of the lexical divergences between the languages are based,

[126] For example, sentence structure, syntactic and semantic valency, descriptores, collocations etc.

[127] Cf. some contrastive studies based on argument structures in German and Rumanian: Cosma/Engelberg (2014).

and to which special attention is given[128]. **Level 4** relates, through different foci, the results of the analysis relating to the different parameters of description of the central modules 2 and 3 of level 2, and presents different types of lexical paradigms that will configure the mediostructure of the dictionary. We differentiate the following interrelations between interlingual data: **level 4.1.1** (for German) and **4.2.1** (for Spanish) interrelate and contrast the different meanings of a lema associated with the same conceptual (sub)field, not only through semantic information based on distinctive semantic features (level 2: module 2) but also in respect of specific argument structures, to which morphosyntactic information is associated (level 2: module 3). Normally, differences in the semantic structure and in the argument structure make it possible to attribute the different meanings to different lexico-semantic (sub)paradigms, as is the case with *aprender1* and *aprender2* (cf. *table 2*) and *lauschen1*, *lauschen 2* and *lauschen3*. Each meaning differs in its argument and semantic structure and this allows for its structuring in the lexico-semantic subparadigms SP1 and SP2 corresponding to the SCC LERNEN/APRENDER (cf. *table 2*) and SP3, SP4 and SP5 corresponding to the SCC AUDITORY PERCEPTION. Other data of interest, such as frequency of use, illustrative examples, and information on semantically close lexemes, is also included in this type of table, which is available for the user of DICONALE. Despite the fact that at this level the perspective is semasiological, we take it that a dictionary which is fundamentally onomasiological should make available to its users this type of information, given that in certain situations, a semasiological perspective might provide relevant information.

Campo conceptual: KOGNITION / COGNICIÓN								
Subcampo conceptual LERNEN / APRENDER								
aprender		**Rasgos semánticos**			**Argumentos**			
[+adquirir conocimientos]	(Sub) Paradig-ma	S 1	S 2	S 3	A 1	A 2	A 3	A 4
aprender1 ⌂ *aprender, adquirir conocimientos*	SP1	+	-	+/-		d		
►Alguien (A1) *aprende* algo (A2) **ES**[129] **<s (cd)>**					s	(cd)		
►Alguien (A1) *aprende* algo (A2) de algo/ alguien (A3) **ES <s (cd) (cp$_{de}$)>**					s	(cd)	cp$_{de}$	
►Alguien (A1) *aprende* algo (A2) en un lugar concreto (A4) **ES <s (cd) (adv$_{en}$)>**					s	(cd)		adv$_{en}$

[128] Cf. some pilot studies on German and Spanish which consider a conceptual perspective in combination with argument structures in contrast: González Ribao/Proost (2014), Meliss (2014a) and Engelberg et al. (eds.) (2014).

[129] ES = *esquema sintáctico* (= syntactic schema).

aprender2 ⌂ *aprender de memoria, memorizar*	SP2	+	+	+				
►Alguien (A1) *aprende* algo (A2) **ES <s cd>**					**s**	**cd**		
Rasgos semánticos (distintivos):	**S1**: [cognitivol], **S2**: [memoria], **S3**: [iterativo];							
Complementos:[130]	**s** = sujeto, **cd** = complemento directo, **cp** = complemento preposicional, **adv** = complemento/suplemento adverbial, (...) = facultativo;							
Argumentos: Con la descripción semántico-categorial[131] y los descriptores necesarios para realizar una especificación detallada	**A1**: APRENDIZ [+anim]; **A2:** LO APRENDIDO: [+intell]: vocabulario, disciplina: matemáticas, etc.; **A3:** ORIGEN DEL ESTÍMULO, [+hum] [+zool]; **A4:** LUGAR DE APRENDIZAJE:[+loc]: escuela, etc.							

Table 2: **Level 4.1.1.** Meanings of *aprender* and information relating to modules 2 and 3 of level 2 of the descriptive model (sequence).

Level **4.1.2.** (for German) and **4.2.2.** (for Spanish) relate the lexemes of the same conceptual field in order to establish different lexico-semantic paradigms for both languages based on distinctive semantic features. Using this structuring it is possible to analyse diverse paradigmatic relations[132] of the signified from the elements of the same lexico-semantic field or from outside this field. Levels **4.1.2.** and **4.2.2.** pay special attention, on the one hand, to existing paradigmatic relations both between lexemes that form a lexico-semantic (sub)paradigm and between the different (sub)paradigms that are all linked to a common concept, and, on the other hand, they also contemplate some lexicalisations of opposite concepts. Offering the user a series of lexical elements along with information on semantic relations of a paradigmatic kind thus addresses the issue of users having at their disposal an array of possibilities to express themselves. From these possibilities the user can choose that which best suits his or her purpose, according to a variety of requisites, such as the type of text, stylistic recourses etc., so that the most fitting linguistic element can be inserted into the text (Sánchez Hernández 2013b). Paradigmatic relations can be seen in light of the semantic definition, or may form part of the lexical entry in an independent way. Forming the paradigmatic information of an entry in an independent way is known as *intentionelle Paradigmatik*, which constitutes a series of advantages in the dictionary (Hausmann 1991: 2794). This kind of information aids the processes of production and expands vocabulary, if and when it is possible to relate to the combinatory information corresponding to **level 2, module 3**. *Table*

[130] The subclassification of the types of complements is based on Engel (2004), with the metalanguage adapted for our own contrastive needs.

[131] Features for the semantic-categorial description of the arguments are based on Engel (2004).

[132] The paradigmatic semantic relations have to do with the integration of vertical lexical relations, as with synonyms, antonyms, hyponyms, hyperonyms etc.

3 gives an example concerning the CC: COGNITION: SCC 1º degree: DAS LERNEN/APRENDER.[133] One of the principle difficulties in collating this type of information involves the distribution of the different conceptual subfields in the corresponding lexico-semantic paradigms[134]:

Campo Conceptual CC	**Subcampo conceptual SCC**	**Lexemas alemanes**	**Relaciones paradigmáticas en alemán**	**Lexemas españoles**	**Relaciones paradigmáticas en español**
COGNICIÓN	**LERNEN GEISTIGE WAHRNEHMUNG**	*lernen1* 'sich ein spezielles Wissen aneignen'	Synonym: •*s. beibringen* Hyperonym: •*s. ausbilden* Hyponym: *memorieren* Antonym: •*verlernen*	*aprender1* 'adquirir conocimientos o el conocimiento de cierta cosa'	Sinónimo: •*instruirse* Hiperónimo*:* •*cultivarse* Hipónimo: •*familiarizarse* Antónimo: •*desaprender*
COGNICIÓN	**LERNEN MEMORIEREN**	*lernen2*[135] 'auswendig lernen'	Synonym*: •pauken* (umg.) Hyponym: •*s. anlesen* Antonym: •*verlernen*	*aprender2* 'Fijar algo en la memoria'	Sinónimo: •*memorizar* Hiperónimo: •*adquirir* conocimiento Hipónimo*: •repetir*
COGNICIÓN	**LERNEN SICH BILDEN**	*studieren1* 'eine Hochschule besuchen'	Hyperonym: •*s. ausbilden* Hyponym: •*lesen*	*estudiar1* 'recibir enseñanza en cierto centro o de cierto profesor'	Hiperónimo: •*Ilustrarse* •*formarse* Hipónimo: •*cursar* Antónimo: •*trabajar*
COGNICIÓN	**LERNEN MEMORIEREN**			*estudiar3* 'aplicar la inteligencia a aprender o comprender algo'	Hiperónimo: •*cultivarse* Hipónimo: •*releer*

Table 3: Some paradigmatic relations: Conceptual field COGNITION.

Level 4.1.3. (for German) and **4.2.3.** (for Spanish) relate the different lexemes of the same lexico-semantic SP (cf. level 3) in terms of their semantic and combinatory characteristics (cf. level 2: modules 2 and 3) and thus offers a first approach to possible equivalences in the contact language (cf. *table 4*).

[133] It is very interesting how the paradigmatic-lexical information takes shape in *elexico*. In terms of the division of the semantic relations: synomomy, antonomy, hyperonomy, hyponymy etc., the attached structure has taken as a base that which is proposed in *elexiko*. This information is drawn from the onomasiological dictionaries: Dornseiff 1965, Wehrle Eggers 1961, Casares ^{2}2007, and also Schumacher - VALBU (2004) in German and M. Moliner ^{2}2002 in Spanish.

[134] In the following table only a few examples of the conceptual field are given. The lexemes *studieren2/estudiar2* are not included, in that the meaning of these is "analyse something in detail", and we consider that these lexemes, from a contrastive point of view, and within the subfield LERNEN, are less relevant than those which appear in the table.

[135] The equivalences of the lexem *lernen2* are *aprender2* and *estudiar3* in Spanish, within the same conceptual subfield, that is, MEMORIEREN.

Konzeptuelles Feld: WAHRNEHMUNG / PERCEPCIÓN						
Konzeptuelles Subfeld: AUDITIV / AUDITIVA **[+Wahrnehmung] [+akustisch]**						
Lexikalisch-semantisches Subparadigma: AKUSTISCH WAHRNEHMEN UND BEWUSST MENTAL VERARBEITEN SP3: "zuhören - Paradigma" **[+bewusst] [+mental verarbeitend]**						
	Modul 2	**Modul 3**				
	dist. sem. Merkmale	**A1**	**A2**	**A3**	**A ....**	**Supplemente**
zuhören1	**Ø**					
►ASTM1 Jemand (A1) *hört* jemandem (A2) *zu* **SBP <s d>** → **escuchar2**		**s** [+hum] .... **wer?**	**d** [+hum] ... **wem?**			...
►ASTM2 Jemand (A1) *hört* etwas (A3) *zu* **SBP <s d>** → **escuchar2**		**s** [+hum]		**d** [+intell] Predigt, Ausführungen, Rede, Konzert, ... **was?**		...
(sich) anhören1	[+genau]					
►ASTM2 Jemand (A1) *hört* (sich) etwas (A3) *an* **SBP <s a>** → **escuchar2** **seguir**		**s** [+hum] ... **wer?**		**a** [+intell] Probleme, Musik, ... **was?**		...
lauschen1	[+konzentriert]					
►ASTM1 Jemand (A1) *lauscht* jemandem (A2) **SBP <s d>** → **escuchar2** **??**		**s** [+hum] ... **wer?**	**d** [+hum] ... **wem?**			...
►ASTM2 Jemand (A1) *lauscht* etwas (A3) **SBP <s d>** → **escuchar2** **seguir?**		**s** [+hum] ... **wer?**		**d** [+intell] Krimis, Erklärungen, Gesang, ... **was?**		...
[...]						

Table 4: Level 4.1.3.: Lexemes corresponding to **SP3** with partial information corresponding to level 2 (modules 2 and 3) (sequence).

Level 5 of the descriptive model is concerned with intralingual information. Two points of focus are of interest here. First, a contrast is made between the the lexico-semantic subparadigms of both languages that correspond to the same (sub)field-concept. For this, of greatest use to us are the interrelations at the level of each lexico-semantic SP (cf. level 3) and we contrast these with the contact language. It involves comparing the different lexicological data at the level of the lexico-semantic fields, such as the degree of lexicalisation, as well as others (**level 5.1.**). The second point of interest is a contrastive study of the lexemes that lexicalise the same concepts and their specific lexicological characteristics (**level 5.2.**). It is hoped that the divergencies between the two languages are to be found above all in the semantic configuration (K1) and in the different morphosyntactic and semantic specifications of the different arguments (K2$_{1\text{-}xyz¿}$).[136]

[136] K1 = contrast at the level of the semantic configuration; K2 = contrast at the argument level; the specification of the K2 contrasts is indicated with superscript.

However, also of particular interest are the data on frequency and use in reference to other parameters of modules 2 and 3 of level 2 of the DICONALE model. *Table 5* shows in level 5.2. the possible contrasts at the semantic (K1) and morphosyntactic levels (K2$_1$) between lexemes of SP3 "zuhören-Paradigma" linked to the SCC AUDITION, from the point of view of Spanish. (i) Contrasts are evident in the semantic specification (K1): *sich anhören1* and *lauschen1* are characterised by the distinctive features [genau] and [konzentriert]. (ii) The lexemes *zuhören1* and *lauschen1* possess the same argument structure as *escuchar2*, but the morphosyntactic realisation of the argument A3 differs (K2). Whereas in Spanish it is realised through a cd (direct object), in German we have recourse to a dative. In the CC COGNITION - SCC DAS LERNEN/APRENDER certain contrastive peculiarities can also be observed, for example, those linked especially to semantic differentiation through the lexico-semantic structure and to categorial features linked to verbal arguments (Sánchez Hernández 2014). The presentation from different contrastive perspectives allows the user to find the most appropriate lexical resource for each particular expressive need in the foreign language and the correct use of these, given that special focus, in a structured and systematic way, is put on the various divergencies between the two languages.

Beschreibungsstufe 5.2.: einzellex. Vergleich				***tertium comparationis***
AKUSTISCH WAHRNEHMEN UND BEWUSST MENTAL VERARBEITEN: SP3: "zuhören – Paradigma"				**[+Wahrnehmung] [+akustisch] [+bewusst] [+mental verarbeitend] • ASTM | A1 A3 |**
escuchar 2		**SBP/ES**	**sem. dist. Merkmale / rasgos sem. dist.**	**Belege / ejemplos:**
	Alguien (A1) escucha algo (A3)	**<s (cd)>**	**≠K2** 2A2sp.	(1a) La prueba estuvo ayer en la sala pequeña del teatro Dramaten, donde el público se rió de lo lindo y ***escuchó*** atentamente la lectura de dos pasajes de Infancia [...] (CREA) (1b) Cuando ***escucho*** la música de mis colegas, me gusta, lo paso bien oyéndola. (CREA) (1c) Durante la visita a la fábrica, el Rey ***escuchó*** las explicaciones de los hermanos Puig sobre el proceso de producción de colonias y perfumes [...] (CREA)
→	***zuhören1***			
	Jemand (A1) *hört* etwas (A3) *zu*	**<s d>**	**≠K1**	(2) Mit großem Interesse hatten die Pflegekräfte [...], dem Vortrag ***zugehört*** [...]. M03/FEB.11701 Mannh. Morgen, 22.02.2003
→	***(sich) anhören1***		[+genau]	
	Jemand (A1) *hört* (sich) etwas (A3) *an*	**<s a>**	**≠K1**	(3) Sie ***hören sich*** die Probleme ***an***, die den Kindern auf den Nägeln brennen [...]. (R97/SEP.72386 Frankf. Rundschau, 15.09.1997, S. 4)
→	***lauschen1***		[+konzentriert]	
	Jemand (A1) *lauscht* etwas (A3)	**<s d>**	**≠K2**	(4) In dem bis auf den letzten Platz gefüllten Gotteshaus ***lauschten*** mehr als 2000 Besucher andächtig den Chorgesängen [...]. (L98/NOV.16306 Berliner Morgenpost, 01.11.1998, S. 9)

Table 5: Level 5.2.: Some possible divergences between lexemes of the **SP3** linked to the SCC PERCEPCIÓN: AUDITION (sequence).

3. Conclusion

In this article we have tried to present the innovative and most relevant aspects of the DICONALE model. In particular, we have sought to focus on two fundamental points in the conception of the dictionary, the information on paradigmatic and syntagmatic relations, and conceptual-onomasiological access to the dictionary.

The dictionary offers information of a lexical nature which facilitates the interlingual and intralingual use of terms pertaining to a specific conceptual field, hence the relevance of the information on paradigmatic relations in the dictionary. Onomasiological access, which supposes the initial ordering, is completed with a semasiological arrangement allowing for a detailed description of the characteristics of the syntagmatic structure. The conceptual-onomasiological ordering, making online access a very useful tool, is an innovation in this kind of **pedagogical bilingual dictionary.**

In this context, one of the novelties of the dictionary is that it is developed on an empirical basis, with data on frequency of use from a number of linguistic corpora containing authentic texts. Due to its online format, the user can access in a modular way the exact information required according to the specific needs of production in the foreign language at a given moment. DICONALE is conceived of as a means of filling a gap in bilingual German-Spanish lexicography, and has been realised with the aim of integrating diverse kinds of information into a single dictionary and adapting it to the needs of specific users in specific situations.

References

Dictionaries and corpora:

Adesse: Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español. URL: <http://adesse.uvigo.es/data/verbos.php>.

Alvar Ezquerra, M.: Duden/Oxford Bildwörterbuch Deutsch und Spanisch. Mannheim: Duden, 1993.

Bosque, I.: Redes: Diccionario combinatorio del español contemporáneo. SM Madrid, 2004.

Casares, J.: Diccionario ideológico de la lengua española. [1. Ed. 1942]. Barcelona: Gustavo Gili, ^{2}2007.

CREA: Real Academia Española: Banco de datos. Corpus de referencia del español actual. URL: <http://www.rae.es>.

Cuervo, R. J.: Diccionario de construcción y régimen de la lengua castellana. Barcelona: Herder, 1953/1998.

DEREKO: Deutsche Referenzkorpus. URL: `<http://www.ids-mannheim.de/kl/projekte/korpora/> über COSMAS II: Corpus Search Management and Analysis System. URL: <http://www.ids-mannheim.de/cosmas2/>.

Diccionario de Alcalá: Alvar Ezquerra, M. (dir.): Diccionario para la enseñanza de la lengua española. Español para extranjeros, Barcelona, Vox y Universidad de Alcalá, (1995/^{2}2000).

Diccionario Salamanca: Gutiérrez Cuadrado, J. (dir.): Diccionario Salamanca de la lengua española, Madrid, Santillana y Universidad de Salamanca,1996/2007.

Diccionario SM: Maldonado, C. (dir.) : Diccionario de español para extranjeros. Madrid: SM, 2002.

DiCE: Diccionario de Colocaciones del Español. URL: <http://www.dicesp.com/paginas>.

Dornseiff, F./Quasthoff, U.: Der deutsche Wortschatz nach Sachgruppen. Berlin: de Gruyter, ^{8}2004.

Dornseiff, F.: Der deutsche Wortschatz nach Sachgruppen. [1. Auflage 1934]. Berlin: de Gruyter, 1965.

Duden 2: Das Stilwörterbuch: Mannheim, Dudenverlag, ^{9}2010.

Duden Deutsch als Fremdsprache – Standardwörterbuch. Mannheim: Duden, ^{2}2010.

ELexiko: Online-Wörterbuch der deutschen Gegenwartssprache. IDS-Mannheim URL: <http://www.owid.de/wb/elexiko/start.html>.

Engel, U./Schumacher, H.: Kleines Valenzlexikon deutscher Verben. Tübingen: Gunter Narr, 1978.

Götz, D./Haensch, G./Wellmann, H.: Langenscheidts Großwörterbuch Deutsch als Fremdsprache. Neubearbeitung. Berlin, München: Langenscheidt, ^{3}2010.

Harras, G. et al.: Handwörterbuch deutscher Kommunikationsverben. Teil 1: Wörterbuch. Teil 2: Lexikalische Strukturen Berlin: de Gruyter, 2004&2007. Online: OWID – IDS: Kommunikationsverben: URL: <http://www.owid.de/docs/komvb/start.jsp>.

Helbig, G./Schenkel, W.: Wörterbuch zur Valenz und Distribution deutscher Verben. Leipzig: VEB Bibliographisches Institut Leipzig, 1969.

Herbst, T. et al.: A Valency Dictionary of English. A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives. Berlin: de Gruyter, 2004.

Kempcke, G.: Wörterbuch Deutsch als Fremdsprache. Berlin/NewYork: de Gruyter, 1999.

Moliner, M.: Diccionario de uso del español. Madrid: Gredos, ^{2}2002.

Pons: Das große Bildwörterbuch. Stuttgart: Ernst Klett Sprachen, 2008.

Pons: Großwörterbuch DaF. Stuttgart: Pons, 2004. Online: URL: < http://de.pons.eu/>.

Quasthoff, U.: Wörterbuch der Kollokationen im Deutschen. Berlin: de Gruyter +versión electrónica, 2011.

Schumacher, H. et al.: Valbu-Valenzwörterbuch deutscher Verben. Tübingen: Gunter Narr, 2004. E-Valbu: URL: <http://hypermedia.ids-mannheim.de/evalbu/index.html>.

Schumacher, H. et. al.: Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben. Berlin: de Gruyter, 1986.

Seco, M./De Andrés, O./Ramos, G.: Diccionario del español actual. Madrid: Aguilar, ^{2}1999.

Wahrig (2008): Großwörterbuch Deutsch als Fremdsprache. Berlin: Cornelsen, 2008.

Web-Corp: URL: <http://www.webcorp.org.uk/live/>.

Wehrle, H./Eggers, H.: Deutscher Wortschatz. Stuttgart: Klett, 1961.

Articles and monographs:

Blank, A./Koch, P. (ed.) (2003). Kognitive romanische Onomasiologie und Semasiologie. Tübingen: Niemeyer.

Bubenhofer, N./Ptashnyk, St. (ed.) (2010): Korpora, Web und Datenbanken: Computergestützte Methoden in der modernen Phraseologie und Lexikographie. Nürnberg: Schneider Verlag Hohengehren (Phraseologie und Parömiologie, 25).

Coseriu, Eugéne (1977/v. esp. ^{2}1986): Principios de semántica estructural. Madrid: Gredos.

Cosma, R.,/Engelberg, St. (2014): „Subjektsätze als alternative Valenzen im Deutschen und Rumänischen“. In*:* Cosma, R. et al. (eds.): Komplexe Prädikationen als Argumente. Kontrastive Untersuchungen zum Deutschen, Rumänischen und Englischen. Berlin: Akademie-Verlag. (in press).

Domínguez Vázquez, M. (ed.) (2013). Trends in der deutsch-spanischen Lexikographie. Frankfurt: P. Lang Edition.

Domínguez Vázquez, Mª J. et al. (2013): “Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen“. In: Domínguez Vázquez, Mª J. (ed.), 135-172.

Dupuy-Engelhardt, H. (1990): La saisie de l’audible. Étude lexématique de l’allemand. Tübingen: G. Narr.

Egido Vicente, M./Fernández Méndez, M./Franco Barros, M. (2014): in this volume (to appear).

Engel, U. (2004). Deutsche Grammatik – Neubearbeitung. München: iudicium.

Engelberg, S./Lemnitzer, L. (2001), (^{4}2009). Lexikographie und Wörterbuchbenutzung. Tübingen: Stauffenburg.

Engelberg, St. (2014a): “The argument structure of psych-verbs: A quantitative corpus study on cognitive entrenchment”. In: Boas, H./Ziem, A. (eds.): Constructional approaches to argument structure in German. Boston, Berlin: De Gruyter Mouton. (in press).

Engelberg, St. (2014b): „Gespaltene Stimulus-Argumente bei Psych-Verben. Quantitative Verteilungsdaten als Indikator für die Dynamik sprachlichen Wissens über Argumentstrukturen“. In: Engelberg, St. et al. (eds.): Argumentstruktur – Valenz – Konstruktionen. Tübingen: Narr. (to appear)

Engelberg, St./Koplenig, A./Proost, K./Winkler, E. (2012): “Argument structure and text genre: cross-corpus evaluation of the distributional characteristics of argument structure realizations”. In: Lexicographica 28, 13-48.

Engelberg, St./Meliss, M./Prosst, K./Winkler, E. (eds.) (2014): Argumentstruktur – Valenz – Konstruktionen. Tübingen: Narr. (to appear)

Fuentes Morán, Mª T. (1997). Gramática en la lexicografía bilingüe. Morfología y sintaxis en diccionarios español-alemán desde el punto de vista del germanohablante. Tübingen: Niemeyer.

Geckeler, H. (1971/^{3}1982): Strukturelle Semantik und Wortfeldtheorie. München: W. Fink.

Geckeler, H. (1993): „Strukturelle Wortfeldforschung heute“. In: Lutzeier, P.R. (ed.), 11-21.

Geckeler, H. (1996): „Einzelsprachliche Analysen von Teilbereichen der lexikalischen Semantik“. In: Wiegand, E.F./Hundsnurscher, F. (eds.): Lexical Structures and Language Use. Tübingen: Niemeyer, 17-28.

González Ribao, V./Proost, K. (2014): “El campo léxico al servicio de la lexicografía: Un análisis contrastivo en torno a algunos subcampos de los verbos de comunicación en alemán y español”. In: Domínguez Vázquez, Mª J. et al. (eds.): Lexicografía de las lenguas románicas. Aproximaciones a la lexicografía moderna y contrastiva. (Coord.: Sánchez Palomino, Mª. D./Domínguez Vázquez, Mª J. (coords.). (vol. 2). Berlin: de Gruyter. (to appear)

González Ribao, Vanessa (2014): “Sobre algunos conflictos en la 'pre'-lexicografía: la selección de corpus para la elaboración de un diccionario contrastivo alemán-español. In: Domínguez Vázquez, Mª J. / Gómez Guinovart, Xavier / Valcárcel Riveiro, Carlos (eds.) (2014): Lexicografía de las lenguas románicas II. Aproximaciones a la lexicografía contemporánea y contrastiva (Coord.: Sánchez Palomino, Mª. D./Domínguez Vázquez, Mª J. (coords.). (vol. 2). Berlin: de Gruyter. (to appear).

Haensch, G./Omeñaca, C. (1997, ^{2}2004). Los diccionarios del español en el siglo XXI. Salamanca: Ediciones Universidad.

Hass, U. (ed.) (2005). Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Berlin: de Gruyter.

Haß-Zumkehr, U. (2001): Deutsche Wörterbücher. Berlin, New York: de Gruyter.

Hausmann, F. J. (1991). “Die zweisprachige Lexikographie Spanisch-Deutsch, Deutsch-Spanisch”. In: Steger H./Wiegand, H.E. (eds.). Wörterbücher: Ein Internationales Handbuch zur Lexikographie. Berlin/NewYork: de Gruyter. 2987-2991.

Hausmann, F. J. et al. (ed.) (1989-1991). Dictionaries. An International Encyclopedia of Lexicography. HSK 5.1-5.3, Berlin, New York: de Gruyter.

Haβ, U./Schmitz, U. (2010). “Lexikographie im Internet 2010 – Einleitung”. In: Gouws, R. H. et al. (ed.). Lexicographica. Internationales Jahrbuch für Lexikographie. 26. Berlin: de Gruyter, 1-18.

Hernández Eduardo (1993): Verba dicendi. Kontrastive Untersuchungen Deutsch-Spanisch. Frankfurt/Berlin: P. Lang.

Kemmer, K. (2010). “Onlinewörterbücher in der Wörterbuchkritik. Ein Evaluationsraster mit 39 Beurteilungskriterien”. OPAL2/2010, 1-33. <http://pub.ids-mannheim.de/laufend/opal/pdf/opal2010-2.pdf>.

Klosa, A. (ed.) (2008). Lexikographische Portale im Internet. *OPAL-Sonderheft* 1/2008, . <http://pub.ids-mannheim.de/laufend/opal/pdf/opal2008-1.pdf>.

Lemnitzer, L./Zinsmeister, H. (2006): Korpuslinguistik. Eine Einführung. Tübingen: Narr.

Lutzeier, P. R. (ed.) (1993): Studien zur Wortfeldtheorie. Tübingen: Niemeyer.

Martín Mingorance, L. (1994). “La lexicografía onomasiológica”. In: Hernández, H./Mederos, H. (Coord.). Aspectos de lexicografía contemporánea. Barcelona: Biblograf, 15-27.

Meliss, M. (2005): Recursos lingüísticos alemanes relativos a “GERÄUSCH” y sus posibles correspondencias en español. Un estudio lexicológico modular-integrativo. Frankfurt: P. Lang.

Meliss, M. (2006): „Kontrastive Wortfeldforschung für das Sprachenpaar Deutsch-Spanisch am Beispiel der Verben für GERÄUSCH“. In: Wolf, D. et al (eds.): Lexikalische Semantik und Korpuslinguistik. Tübingen: G. Narr, 141-167.

Meliss, M. (2011). “Wörterbücher von heute und morgen. Überlegungen zu lexikographischen Benutzersituationen im zweisprachigen Kontext: Spanisch-Deutsch”. In: Domínguez Vázquez, Mª J. et al. (ed). La palabra en el texto. Santiago de Compostela: Servicio de publicaciones USC. 267-300.

Meliss, M. (2013a). “Das zweisprachige Wörterbuch im bilateralen deutsch-spanischen Kontext. Alte und neue Wege”. In: Domínguez Vázquez, Mª J. (ed.), 61-87.

Meliss, M. (2013b): “Online-Lexikographie im DaF-Bereich: Eine erste kritische Annäherung: Bestandsaufnahme – Nutzen – Perspektiven”. Real Revista de Estudos Alemães, 4. 176-199. <http://real.fl.ul.pt/textos.page/pag/2>.

Meliss, M. (2014a): “Los verbos de percepción en DICONALE: propuestas para un diccionario conceptual bilingüe para ELE y DaF”. In: Domínguez Vázquez, Mª J. et al. (eds.): Lexicografía románica. Aproximaciones a la lexicografía moderna y contrastiva. (Coord.: Sánchez Palomino, Mª. D./Domínguez Vázquez, Mª J. (coords.). (vol. 2). Berlin: de Gruyter. (to appear)

Meliss, M. (2014b): „Argumentstrukturen, Valenz und Konstruktionen im Umfeld von Verben der WAHRNEHMUNG im Deutschen und Spanischen“. In: Engelberg, St. et al. (eds.): Argumentstruktur – Valenz – Konstruktionen. Tübingen: Narr. (to appear)

Meliss, M. (2014c): „Das verbale Kombinationspotenzial in einsprachigen DaF-Lernerwörterbüchern: Kritische Bestandsaufnahme – Neue Anforderungen“. En: ZDaF (to appear).

Meliss, M. (2014d). “(Vor)überlegungen zu einem zweisprachigen Produktionslernerwörterbuch für das Sprachenpaar DaF und ELE”. In: Reimann, D. (ed.): Kontrastive Linguistik und Fremdsprachendidaktik Iberoromanisch – Deutsch. Studien zu Morphosyntax, nonverbaler Kommunikation, Mediensprache, Lexikographie und

Mehrsprachigkeitsdidaktik (Spanisch/Portugiesisch/Deutsch) (Reihe: Romanistische Fremdsprachenforschung und Unterrichtsentwicklung). Tübingen: Narr. (to appear)

Model, B. (2010). Syntagmatik im zweisprachigen Wörterbuch. Berlin: de Gruyter.

Müller-Spitzer, C./Engelberg, St. (2013): "Dictionary Portals". In: Rufus H. Gouws et al. (eds.): Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent developments with special focus on computational lexicography. Berlin, New York: de Gruyter (to appear).

Proost, K. (2007). Conceptual structure in lexical items: The lexicalisation of communication concepts in English, German and Dutch. Pragmatics & Beyond New Series; 168. Amsterdam/Philadelphia: Benjamins.

Reichmann, O. (1989). "Das onomasiologische Wörterbuch: Ein Überblick". In: Steger H./Wiegand, H.E. (ed.). Wörterbücher: Ein Internationales Handbuch zur Lexikographie. Berlin/New York: de Gruyter. 1057-1067.

Renouf, A. (ed.) (2009): Corpus linguistics: refinements and reassessments. Amsterdam [u.a.]: Rodopi (Language and computers, 69).

Sánchez Hernández, P. (2010). "Análisis contrastivo alemán español de los verbos fragen-antworten/ lehren-lernen". In: Revista de Filología Alemana 18. 261-283.

Sánchez Hernández, P. (2012).). "*Lernen-aprender*: una aproximación contrastiva dentro del campo semántico Kognition en torno a diversas peculiaridades semánticas y sintácticas". In: Revista de Filología alemana 20. 139-158.

Sánchez Hernández, P. (2013a). "La lematización de las unidades fraseológicas en diccionarios generales semasiológicos y onomasiológicos". In: Mellado, C. (coord.) et al. (ed.). La fraseología del alemán y el español: lexicografía y traducción. [Reihe: Linguistische Studien]. München: peniope. 129-143.

Sánchez Hernández, P. (2013b). "Zur Konzipierung eines deutsch-spanischen kombiniert onomasiologisch-semasiologisch ausgerichteten Verbwörterbuchs mit online-Zugriff – ausgewählte Aspekte". Aussiger Beiträge 7. (in press)

Sánchez Hernández, P. (2014): Syntagmatische Information zu einigen Verben der Kognition im Rahmen des Projekts DICONALE. En: Engelberg, St. et al. (eds.): Argumentstruktur – Valenz – Konstruktionen. Tübingen: Narr. (to appear)

Schmidt, Th. (ed.) (2012): Multilingual corpora and multilingual corpus analysis. Amsterdam, Philadelphia: Benjamins (Hamburg studies on multilingualism, 14).

Storrer, A. (2010). "Deutsche Internet-Wörterbücher: Ein Überblick". In: Gouws, R. H. et al. (ed.). Lexicographica. Internationales Jahrbuch für Lexikographie 26. Berlin: de Gruyter, 154-164.

Tarp, S. (2012): "Online dictionaries: today and tomorrow". In Heid, U. (ed.): Thematic Part: Corpora and Lexicography. Lexicographica (International Annual for Lexicography) 28/2012. Berlin, 253-267.

Trujillo, R. (1970): El campo semántico de la valoración intelectual en español. Universidad de La Laguna.

Werner, R. (1998). "La selección de lemas en los diccionarios español-alemán y alemán-español o ¿un diccionario de qué lengua es un diccionario de las lenguas española y alemána? " In: Fuentes Morán, Mª T./Werner, R. (eds.). Lexicografías iberrománicas: problemas, propuestas y proyectos. Frankfurt a.M.: Vervuert, 139-156.

Wotjak, G. (1992): „Semantische Makrostrukturbeschreibung (lexikalischer-semantischer Felder) und (enzyklopädische) Wissensrepräsentationen". In: Lutzeier P.R. (ed.), 121-136.

Wotjak, G. (2013): "Inwieweit kann das kommunikative Potenzial lexikalischer Einheiten als Bezugsbasis fúr kontrastive Lexikographie im Sprachenpaar Spanisch-deutsch dienen? In: Domínguez Vázquez, M.J., 109-133.

A survey of *Dicionário Houaiss da Língua Portuguesa* lexicographic corpora

Mauro Villar

In the 1980s, the Portuguese language utterly lacked any broad specific corpora designed for the making of dictionaries. England itself, the power behind the revolutionary contextual approach to lexicography, was barely starting to use such a tool, and, for that matter, mostly thanks to the database developed years earlier by John Sinclair and Sue Atkins for the *Collins Cobuild English Dictionary*.

The shaping of the British National Corpus by these two brilliant British lexicographers dates back no further than 1988, at that time already counting 100 million words of text and open to public use against a small fee.

In other words, in the 1980s, the world of dictionaries still relied on extensive reading and painstakingly surveying and gathering illustrative citations in order to prove the actual use of a sentence, word or sub-sense; this was how anyone undertaking lexicographic work anywhere in the world could collect data about a language.

On the other hand, it should be noted that the significance of *corpora* was only acknowledged by the computational linguistics community in 1989, following an ACL (Association for Computational Linguistics) congress in Vancouver, Canada.

Therefore, the group that started working on *Dicionário Houaiss da Língua Portuguesa* back in the 1920s did not have many tools at hand. Whatever had been and was being gathered by project NURC (Projeto de Norma Urbana Oral Culta / Educated Urban Speech Norm) from its start in the 1970s, through the 80s and into the 1990s; plus the survey "A Frequency Dictionary of Portuguese Words", which amounted to a PhD dissertation defended by John Clifton Duncan before Stanford University's Linguistics Committee in 1971. It goes without saying that nothing existed online.

Therefore, *Dicionário Houaiss da Língua Portuguesa* is basically a work built upon the gathering of manually collected citations. And the early moves in its making rested on some copious materials gathered in file cards belonging to Antônio Houaiss, the fruit of many years of obstinate readings of documents and literature, and collected with the aim of eventually being used in a future lexicographic work.

Adding to such materials, other files were provided by friends and collaborators moving around the Brazilian philological and lexicographical scene. Among others, I wish to give an honorable mention to Carlos Francisco de Freitas Casanovas, author of published dictionaries who worked at Instituto Nacional do Livro (National Book Institute), and whose compilations eventually mingled together with various other collections to provide the foundations for our early work.

Another one who granted us illustrative citation files from his readings and notes was Manuel da Cunha Pereira, a Portuguese who had settled in Brazil and author of an orthographic vocabulary, in 1954, in collaboration with Luiz Peixoto Gomes Filho. Later on, this vocabulary was authored by Aurélio Buarque de Holanda Ferreira, first with Manuel as co-author, then by himself. Manuel was part of our team through the first stage of the construction of our dictionary, but passed away before the second stage started.

We also had access to a fine collection of words, phrases and senses, gathered by philologist Olavo Aníbal Nascentes – etymologist Antenor Nascentes' son –, and backed by illustrative citations.

The readings and notes of Antônio Geraldo da Cunha, a well-known author of, a vocabulary index of medieval Portuguese and an etymology dictionary, among other works, were another spring of words. As to the index, it is worth noting that its first volume (letter A) was published by Fundação Casa de Rui Barbosa in 1986, followed by the second volume (letters B & C), in 1988, and a third volume (letter D and a supplement to A-C) only in 1994. Another file (referred to in our dictionary as FichIVPM and deposited with Fundação Casa de Rui Barbosa), containing around 170 thousand typed cards transcribing the excerpts which document the words, was added later on and became, in 2007, the Vocabulário Histórico-Cronológico do Português Medieval (Historical and Chronological Vocabulary of Medieval Portuguese), published with a CD-ROM by Edições Casa de Rui Barbosa. With Antônio Geraldo da Cunha in our dating team, we had the chance to review all this material, and to use part of it even before its publication.

The *Dicionário Houaiss da Língua Portuguesa* was designed based on a personal perspective of the lexicographers in the team. We thought about what pleased us, and what displeased us in the dictionaries which existed in the language. We thought for ways to do something different. About ways of improvement and what new additions to make while retaining the project's feasibility in terms of the time required for its elaboration. We thought about groups of lexicographer and collaborators, both generalists and specialists, which might be trusted to carry on the task, without incurring excessive production costs, or losing control of the creation flow, and ways to structure such a team for such a purpose. We also wondered about technical innovations from contemporary foreign dictionaries to consider for final incorporation. Finally, we thought about which methodological contributions, as compared with existing dictionaries, we could try to achieve.

Assumptions of the sort, among many others considered before and in the early stages of development of *Dicionário Houaiss*, bounded the establishment of the methodological plan to be adopted by the team.

Four fundamental principles were then established for the work ahead:

First, the compilation, by means of various gathering techniques, of a nomenclature as comprehensive as possible – but consistent with the scope of a general-purpose dictionary – that would also incorporate Portuguese linguistic performances outside the Portugal/Brazil axis. The register of lemmas and sub-senses needed the backing of reliable sources, and their microstructural development had to be based on the proven use of collected citations – to the extent allowed by the sources we were using.

Second, a full-fledged effort of dating, that would act as illustrative citations for the lexical units and convey reliability to the way in which sub-senses appear arranged under the headwords.

Though ours is not a historical dictionary, our aim has been to classify in chronological order the sequence in which the sub-senses appear within the microstructure of our entries by dating the first written occurrence of each headword, as well as each sub-sense that we managed to track from written documents. We could imply a historical path for the gradual appearance of new senses, as well as hypothesize or clarify the linguistic phenomenon of semantic derivation behind the chain of alterations. To this day, we have over 100 thousand words dated for their first written occurrence in the language, and, within the entries, around seven thousand senses and related idiomatic expressions, also dated. The work of dating was based on extensive literature and allowed us, for the first time in the Portuguese language, to attempt a non-random structural organization of each entry, from the earliest sub-sense onwards.

The third rule was to invest in a more consistent etymological investigation than we chanced upon in dictionaries, often advancing more than one single hint about the word's possible origins, if these pose any doubts, and stating the hypothesis' authorship.

Fourthly, a survey, as exhaustive as possible, of the language's formative elements, in order to assemble large families of words with common origins; this would shorten the etymology presented for each word and enlighten users about semantic relationships among the elements within such families. What led us to plunge deep into this issue was its instant usefulness for clarifying a number of questions raised by dictionary users – for instance, the reason for a certain form, or meaning, the date when the element slipped into the language, whence did it come from, which family of units did it give rise to, why is the spelling of a given form endorsed as preferable over another, and so on. The dictionary endeavoured to register what was already common knowledge about the subject and, whenever possible, add something further; the result was a specific collection of 13,295 units, converted into headwords of a general word-list that encompass collections of examples with dozen, even hundreds of interconnected words, and disclose the broad net of families in the language.

This expedient freed etymology to plunge into each word's immediate etymon; then, it was just a matter of stating the morphic element which served as the remote etymon, without having to repeat, in each entry, common data about the family to which the defined word belongs.

The first stage of work on the dictionary primarily involved the gathering of data to give shape to these four orders, but not only this, as it was also necessary to start off with its writing. The documental base upon which the dictionary was built came from surveys carried out in many sources, some that I have just mentioned, as well as literary, scientific, technical and pedagogical works, general information and entertainment magazines and so on. Most of them quoted in the dictionary's bibliographies, which in turn generated extensive new files.

In the making of our dictionary, we enlisted the work of collaborators from Portugal, São Tomé and Príncipe, Guinea-Bissau, Cape Verde, Angola, Mozambique, as well as someone familiar with Macao. The list of headwords and sub-senses mirrors a diachronic survey of standard, literary, and educated Portuguese, but also includes and defines informal words and dialect registers, both Brazilian and Portuguese; words and phrases from Eastern and African Portuguese creoles; and several foreign words – from languages like Chinese and some African languages – that have slipped into our lexicon, since they appear in literary works set down in Portuguese.

All major dictionaries of the Portuguese language, and foreign lexicons too, were also meticulously reviewed, so as to avoid any significant gaps, and to figure out the overall universe we would be exploring.

The *Dicionário Houaiss* did not privilege any given chronological or geographical segment of the language. Rather, it portrayed a diachronic cut across phenomena from contemporary Portuguese, both in Portugal and Brazil, and chosen words from the archaic and old language, whose inclusion, as stated in the technical explanations which open the dictionary, is warranted by the frequency of their occurrence in the history of Portuguese literature.

It took us 15 years to complete the project, which engaged the work of 200 people, among whom 28 generalist and 15 specialist editors, plus 42 external collaborators, mostly from the fields of science and technology.

The dictionary's first edition had over 220 thousand word entries and provided, in addition to a semantic investigation as thorough as possible, information on orthoepy, phonetic transcription (only foreign words), verbal regency, and also, though in a tentative and cautious manner, as such information requires, diachronic variation and regional location of word senses. The closing fields of the microstructure bring plurals with their own sense, information about phrases, phraseology, grammar and usage, variants, synonyms, antonyms, homonyms, paronyms, collectives, plus historic spellings of the word, and even a final, onomasiological section, informing the user as to which morphic elements of the language correspond to the headword's notion, so as to allow the composition of perfectly correct neologisms.

Completed in December 2000, the dictionary was published in Brazil in September of the following year, both in print and on CD-ROM. An adapted version of this edition, conforming to the spelling and other peculiarities of European Portuguese, was published in Portugal in 2002, and its four portuguese reprints in would sell over 70 thousand copies altogether.

Though concatenation of our lexical units rested essentially upon the citations collected, we chose not to print those *ipsis verbis*, but to strip them to their basic significant elements, simplified as formular examples. Otherwise, the dictionary would almost double in size. In spite of our efforts to streamline the texts, the first Brazilian edition weighed over four kilos. We hoped to print a second volume, just to document the citations collected, including datings. This would have been better, but was not affordable. The cost of the first edition had amounted to 6,5 million reais (6,5 million dollars, at the time). Another 1,1 million reais were required after the spelling agreement was signed and came into effect in Brazil, to review and fit the whole thing to the new spelling. Taking this in consideration, we were not able print the second volume.

However, specific readings for lexicographic citation remain effective as a way to highlight the occurrence of language changes, even after Google. As stated in *The Oxford Guide to Practical Lexicography* (2008), computers can easily spot neologisms while "reading" texts (let's say, words like *desaposentadoria*, or *denisovan*, a new hominid discovered by paleonthologists). Nevertheless, to a large extent, the "new" vocabulary consists of compounds, phrases and uses of already existing words. This is where the problem lies, and this is where human reading excels. The traditional method allows, for example, the distinguish between new genuine uses and *ad hoc* coinages (op. cit., 2008, p. 51).

The web's arrival at the scene of our *métier* opened up new perspectives for the search of quotations through human reading: illustrative citations gathered by the traditional process may now be verified and compared at the plentiful sources available on the web. Also promising and comforting is the fact that a large number of written texts, even rare and ancient ones, are becoming more and more available on the web.

The advent of both corpora and the Internet brought about deep transformations: while the first changed the lexicographic *modus operandi*, the latter broadened its possibilities, changing the whole picture once over (op. cit., p. 53). Nowadays, Instituto Houaiss de Lexicografia employs this means of information and communication for various specific tasks, such as semantic confirmations, reading texts, gathering quotations to infer senses or improve definitions, antedating, collecting textual evidence from the current language, and even storing our databases in the cloud.

It is not a very good idea to rely only, or essentially, on corpora drawn from newspaper or magazine texts, a task which is now considerably facilitated by the media corporations' initiative to supply us with daily screenings of their own editions. This is largely useful for smaller, synchronic dictionaries. The problem for us lies in the fact that such media cover only a few segments of the language, while there are so many others that we must take into account to fulfill our task. This is why reading books and other related materials remains so important, as these encompass much more varied and extensive language categories. Many books are, yet, not available on the web, unlike most major newspapers and magazines, which explains why thorough reading remains indispensable for lexicographic quotation.

For some time now, Instituto Houaiss de Lexicografia has been using a number of web sites, among which the following 25 are worth mentioning: corpus do português.org; Núcleo Interinstitucional de Linguística Computacional (NILC); Brasiliana USP (online library at Universidade de São Paulo); Projecto AC/DC (Acesso a Corpos/Disponibilização de Corpos); CINTIL-Corpus Internacional do Português (developed at Universidade de Lisboa); Corpus de Extractos de Textos Electrónicos MCT/Público (CETEMPúblico); Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo (CETENFolha); clp.dlc.ua.pt (Corpus Lexicográfico do Português da Universidade de Aveiro e do Centro de Linguística da Universidade de Lisboa); sistemas BNP (Biblioteca Nacional de Portugal); the browsing systems of Academia Brasileira de Letras; internet archive (free online bookstore based in San Francisco, California); purls (Persistent Uniform Resource

Locators); Wikisource (free online library of public domain works); pt.scribd.com (sharing platform); ibiblio.org/ml (online database); almamater (online library at Universidade de Coimbra); books.google.com; carlamaryoliveira.pro.br (personal site of a professor at Universidade Federal da Paraíba, Brazil); objdigital.bn.br and docvirt.com (online collection, Biblioteca Nacional, Brazil); The University of Florida Digital Collections (UFDC); bibliotecavirtual.sp.gov.br (online library of state government of São Paulo, Brazil); ihgb.org.b (online library of Instituto Histórico e Geográfico Brasileiro); acervo.estadao.com.br (online collection of Estado de São Paulo, a Brazilian newspaper); biblio.etnolinguistica.org (Curt Nimuendaju online library), etc.

Any user can have access to corpora for searching through one or more databases, either online (e.g, the web), or offline (e.g., a computer's local files); research requires the use of keywords, which are supplied by the user himself/herself. This amounts to having millions and millions of different texts literally at hand.

As for neologisms, new words and senses are captured, stored, and eventually discussed as regards their inclusion in our macrostructure – the practice is nowadays more like a game of trial, since past protocols (for instance, the number of years elapsed before the word is entered in a dictionary) are no longer the sole indicators accounted for in lexicography. Factors like usage extent and occurrence in different kinds of sources, for instance, are now relevant for a decision. "Technology and the speed of communication have greatly increased the pace of establishment of new uses and words", states Angus Stevenson, co-author of the latest edition of the *Shorter Oxford English Dictionary* (2002). At least in Brazil, words like *funkeiro*, though very recent, have slipped fast into the wordlist of current dictionaries. The point is that these words are widely used, and are expected to last fairly long in the language.

New means of written communication around since the 1990s have produced more corpora which certainly must be accounted for: I refer here not only to the ubiquitous e-mails, blogs, chat rooms, and those short messages exchanged through cell phones by pressing repeatedly the number key corresponding to each letter, but also exchanges of correspondence, confessions, and, not seldom, insults, on social networks. The amount of data generated by these sources about recent means of communications is undoubtedly considerable: according to data published by VEJA magazine (May 15th 2013), 2,5 exabytes of information (1 exabyte means 1 followed by 18 zeros) are produced daily, worldwide, of which 24 petabytes (or 10^{15}, i.e. one thousand billion flops, the calculation unit for computer processing speed) run through Google, 43 petabytes run through cell phones and tablets connected to the web, and 10 petabytes run through sent e-mails.

A large and daunting universe, for sure, but so far excluded from our line of work, as we cannot take into account the information it provides because it does not conform to the pattern of text categories established in the olden days (op. cit., p. 53).

We do not have a regular service for channeling words, phrases and new senses suggested by the public, but we count on a fine group of collaborators to scan a wide variety of reading sources and suggest new additions.

New words and senses are entered on the basis of their frequency of employment, as verified by ourselves on the web, and also, as previously noted, on the variety of texts in which they occur. This input is stored in special files and may be recovered, and eventually used, either by word, or by entry date, or by subject etc. New editions of our dictionary usually incorporate the most recent science and technology terminology through the contribution of experts in each field.

In short, then, *Grande Dicionário Houaiss da Língua Portuguesa* is the sort of work originally built on the manual collection of illustrative quotations, having been started and developed at a time when there were no other methods available either in Portugal, Brazil, or most of the planet. However, as early as possible, we engaged with corpora and web possibilities, which, though more comfortable and practical to work with, are, so far, not so rich in historical information about the language – and ours is a diachronic dictionary. We resorted to the new media as a way of improving parts of our previous work, updating the dictionary, documenting new uses and senses, reading rare works that had been out of our reach, adding to what was already gathered and organized. And thus we kept going, set on progressively adapting and developing our gathering structure, so as to take advantage of the best that the newly available electronic resources can provide.

Refereces

Atkins, B.T. S.; Rundell, M. (2008). *The Oxford guide to practical lexicography*. New York: Oxford University Press.

Collins Cobuild English dictionary. [3. ed.] Glasgow: HarperCollins Publishers, 2001.

Cunha, A. G. da (2008). *Vocabulário histórico-cronológico do português medieval*. Rio de Janeiro: Edições Casa de Rui Barbosa. 1 CD-ROM.

Duncan, J. C. (1970). *A frequency dictionary of Portuguese words*. Stanford: Stanford University, Committee on Linguistics. 2. vol.

Fontenelle, T. (dir.). (2008). *Practical lexicography*. New York: Oxford University Press.

Houaiss, A.; Villar, M. de S. (2001). *Dicionário Houaiss da língua portuguesa*. Rio de Janeiro: Objetiva.

Shorter Oxford English dictionary on historical principles. [5. ed.] New York: Oxford University Press, 2002. 2. vol.