	Candidates	**Underspecified**	**Noun**
N	516	206	51
Max	0.964	0.951	0.958
Min	−0.216	−0.170	−0.026
Median	0.056	0.095	0.471
Ave	0.136	0.198	0.477
StDev	0.208	0.256	0.342

Table 2: Descriptive statistics of headword candidates, underspecified headwords without PoS tags, and the ambiforms tagged as case forms of nouns by EstNLTK morphological analysis

The data in Table 3 reveals that the maximum levels of DI are similar in all three sets, indicating that there are good candidates for decategorisation in each set, regardless of the current lexicographic status of the ambiforms. The average and median are considerably lower in the "Underspecified" group, the ambiforms in headword status without PoS tags, and the lowest in the case of "Candidates". This indicates that the lexicographic status, on average, follows the trend characterised by the relative salience of the word forms.

In relation to the "Noun" sample, the "Candidates" and "Underspecified" groups stand out for showing similar tendencies. These two sets have more tightly grouped DI values: the median results of these sets (0.056 and 0.095) are considerably lower than that of the comparison basis of "Noun" (0.471). Moreover, the average DI of the two analysed groups is 3.5 and 2.4 times lower than that of "Noun". The range of variation outside the box of 50% of the data, however, is much wider in the "Candidates" and "Underspecified" groups than in "Noun"; the extreme outliers over the upper quartiles show "abnormal" cases in these two groups.

The DI values of the headword candidates with no CombiDic headword tags are displayed in a dot chart in Figure 5:

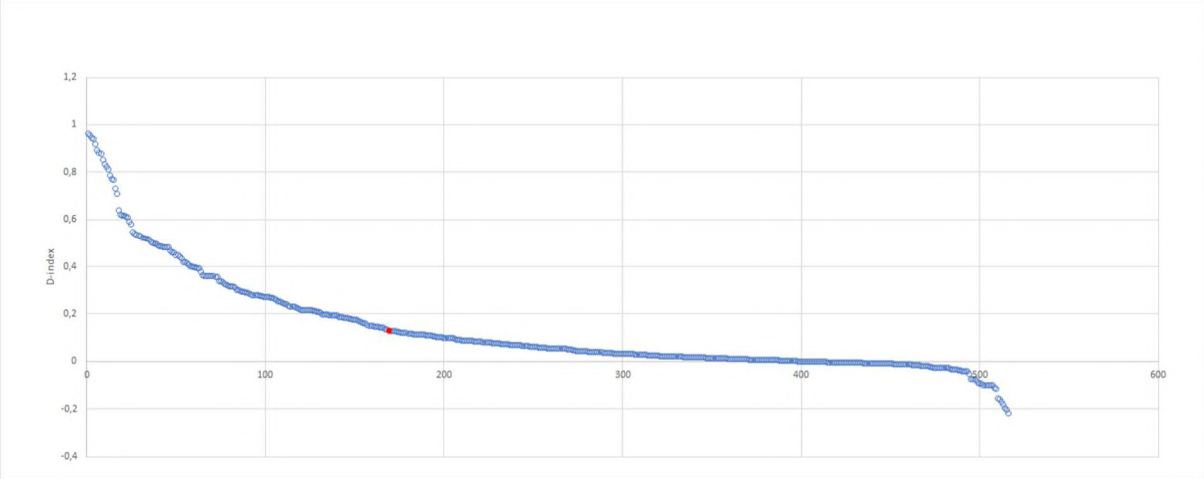

Figure 5: Descending values of the "Candidates" for dictionary headwords

This is a large set of ambiforms (N = 516). The value closest to the threshold (0.129 for the word form *keskmesse* [midpoint-ILL] 'to the centre') is highlighted. Only 33% of the ambiforms in this selection exceed the threshold (0.130) and truly qualify as candidates for headwords based on their morphological distribution statistics. Overall, this group shows particularly broad variation, from extremely high DI values (0.964) to negative values down to −2.16, indicating underrepresentation in relation to the expected frequency. At the top of the list are several compound ambiforms (see 8), but there are also non-compound words with exceptionally high DI (9):

(8)	*tikutulega*	(DI 0.96)	[match.light-COM]	'scrupulously'
	ajajooksul	(DI 0.94)	[time.run-ADE]	'over time'
	äravahetamiseni	(DI 0.89)	[away.exchange-TER]	'interchangeable'
	reaalajas	(DI 0.88)	[real.time-INE]	'in real time'
	vastutasuks	(DI 0.87)	[for.pay-TRA]	'in return'
(9)	*alustuseks*	(DI 0.95)	[commencement-TRA]	'for a start'
	nõrkemiseni	(DI 0.92)	[exhaustion-TER]	'to exhaustion'
	maksvusele	(DI 0.82)	[validity-ALL]	'validated'

The "Underspecified" ambiforms show a smoother decline in Figure 6. The value closest to the tentative threshold (0.129 for the ambiform *võtmes* [key-INE] 'à la') is highlighted. Compared to the "Candidates", this group has more ambiforms over the threshold: 45% of the calculated DI values. These 93 case forms are good candidates for decategorisation as indeclinable words.

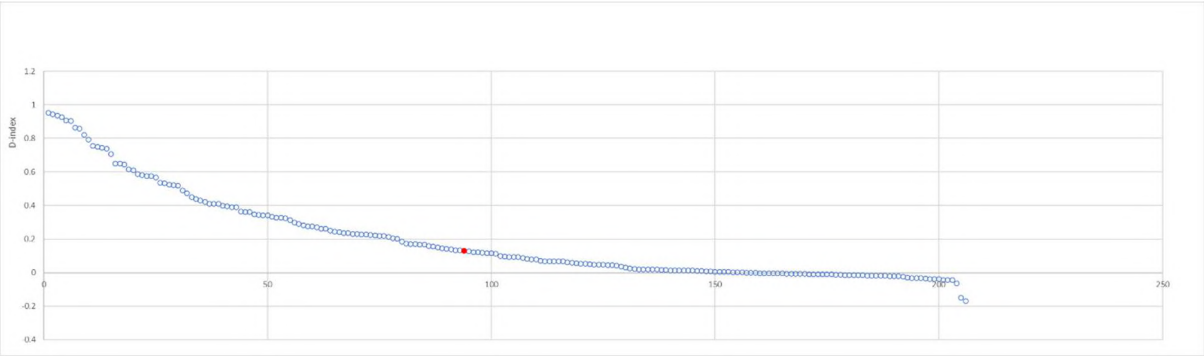

Figure 6: The Descending DI values of the "Underspecified" CombiDic headwords without PoS tags

Similarly to the previous group, "Candidates", the ambiforms with the highest DI are mostly compounds (see (10) and (11)). The ambiforms with DI levels indicating abnormal distributions in the form of underrepresentation (see (12)) are all provided with the comment "used only in negations" in the CombiDic. The reason for that is the emphatic suffix *-gi/-ki* after the case endings, often adding a sense of negation to the stem.

(10)	*üldjoontes*	(DI 0.95)	[common.feature-PL-INE]	'generally'
	esmapilgul	(DI 0.94)	[first.glance-ADE]	'at first glance'
	täismahus	(DI 0.93)	[full.capacity-INE]	'in full'
	lõppkokkuvõttes	(DI 0.9)	[end.conclusion-INE]	'in conclusion'
	eestvedamisel	(DI 0.86)	[front.leading-ADE]	'led by'
	tavamõistes	(DI 0.86)	[ordinary.sense-INE]	'colloquially'
	imeväel	(DI 0.78)	[miracle.power-ADE]	'miraculously'
	noaotsaga	(DI 0.76)	[knife.edge-COM]	'in a pinch'
(11)	*kamaluga*	(DI 0.93)	[cupped hands-COM]	'abundantly'
	mahitusel	(DI 0.9)	[encouragement-ADE]	'with the connivance of sb.'
	kuhjaga	(DI 0.61)	[pile-INE]	'heaped'
	kuubis	(DI 0.57)	[cube-INE]	'cubed'
	moel	(DI 0.57)	[way-ADE]	'in a way'
	sõnul	(DI 0.53)	[word-ADE]	'according to'
(12)	*varjugi*	(DI –0.15)	[shadow-PART-EMPH]	'(not) in the slightest'
	viluvarjugi	(DI –0.17)	[shade.shadow-PART-EMPH]	'(not) in the slightest'
	piiskagi	(DI –0.06)	[drop-PART-EMPH]	'not a drop'

4.3 Implications of morphological and lexicographic PoS tagging status on DI values

An examination of the impact of the morphological analyser on the DI results in Section 4.1 suggests that the most relevant and reliable results of the DI derive from the analysis of ambiforms that are processed as case forms of nouns without splitting the PoS interpretations into noun and additional categories. This suggests that for a realistic outline of the distributional analysis of an ambiform, all of its PoS-readings should be reverted to the noun if possible.

The influence of the headword-labelling situation of ambiforms on their DI levels examined in Section 4.2 raises the question of the relation of lexicographic treatment and ambiforms. We can ask if the DI exposes the lexicographic status of ambiforms, i.e. can the DI predict which word forms are headwords in the combined dictionary? According to our results, the answer is no: the DI variation of ambiforms that are headword candidates (not headwords in the CombiDic) and underspecified ambiforms (headwords without PoS tags) does not show significant differences.

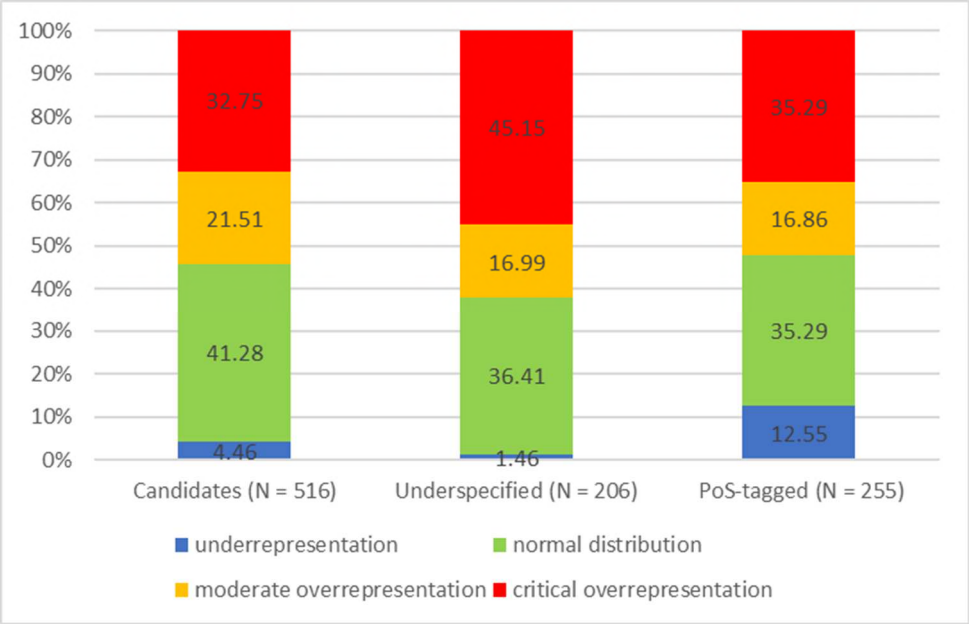

Figure 7: The division of DI results in three data sets: headword candidates, underspecified headwords and PoS-tagged headwords in the CombiDic

The results of the analysis in Sections 4.1–4.2 are summarised in Figure 7. The diagram visualises the division of DI results according to the four degrees of DI values in four data proportions: underrepresentation, normal distribution, moderate overrepresentation, and critical overrepresentation. The three columns represent the examined data from the perspective of their lexicographic status:

- "Candidates" – the ambiforms without headword status in the CombiDic
- "Underspecified" – the ambiforms with headword status but no PoS tags in the

CombiDic

- "PoS-tagged" – the ambiforms with PoS tags other than noun in the CombiDic (this column unites the data analysed in Section 4.1: the case forms of nouns in the EstNLTK morphological analysis ("Noun") and the ambiforms with split PoS analyses ("Noun+")

The proportion of critical and moderate overrepresentation is the highest and the underrepresentation the lowest in the group of underspecified ambiforms, which might indicate why these ambiforms have been given headword status in the CombiDic, although not PoS yet. The headword candidate group has a slightly smaller proportion of critical overrepresentation forms, but the highest proportion of moderate overrepresentation. The group with the expected highest proportion of critical and moderate overrepresentation, the PoS-tagged ambiforms, do not stand out in this respect; surprisingly, this group shows the largest underrepresentation level. It should be noted here that the headword inclusion in the CombiDic has not been related to the statistical distribution of the form so far. For further discussion about the reasons for including word forms with lower-than-normal distribution levels, see Vainik et al. (2021).

After the examination of the ambiform groups with different statuses in morphological analysis and lexicographic practice, we can ask if it is possible to specify any further thresholds in the relatively large area of the critical overrepresentation between the DI values 0.13−1.0. The analysis of the four groups of ambiforms (cf. Figures 3−6) reveals a gap in the line graphs around the value 0.62−0.63. This makes it possible to establish an indicative level of DI of the stage near the indeclinable words. The threshold for ambiforms approaching the characteristics of uninflected words can thus be assigned a provisional value of 0.63.

5. Conclusions

This study aimed to examine the effect of the distributional character of case forms of nouns that have already been or may be decategorised into other parts of speech. We tested the D-index developed a part of this study to detect the deviating frequency of case forms in different settings. PoS-tagging discrepancies between the morphological analyser and the combined dictionary enabled us to study the effect of "inured" and absent decategorisation on the D-index score. The results suggest that for the outcome to be most authentic, the noun-based ambiforms should be analysed without the decategorisation influence, i.e. the D-index analysis should be applied in the pre-PoS-disambiguation stage.

The threshold levels of DI posited in the previous study seemed to function relatively well as indicators of the underrepresentation, normal and moderate and critical overrepresentation of forms. The threshold value of 0.13, the marker of heightened frequency, appears to hold. The analyses of different groups of ambiforms suggest that

the upper part of the critical overrepresentation (0.63), as a quite broad stage, could be preserved for the stage of "approaching the characteristics of uninflected words". A closer study of the ambiforms in this upper area is recommended for future research.

In our opinion, the D-index contributes statistical corpus post-processing information in certain stages of the lexicographic workflow: the specification of a lexeme's status as a headword and its PoS affiliation. For easy and fast access to a form's D-index, we have developed the Distribution Index Calculator for Estonian. It is a web-based application that retrieves the frequency data of word forms and lemmas from an annotated corpus and retrieves DI statistics on a lexicographer's workbench (see Vainik et al., 2021).

Since the results of the D-index (and the PoS-tagger) analysis depend on the outcome of morphological dissection, the future development of the natural language processing tasks is also relevant for our purposes. In this article, we have tested one morphological disambiguator available for the Estonian language; the other possibilities are currently the Universal Dependencies PoS Tagger[14] and the TreeTagger[15]. The development of a pre-trained language model, such as Bert, has shown promising results in PoS and morphological tagging of Estonian (see Kittask et al., 2020), which has the potential to also improve the results of the D-index calculus.

In the process of examining the D-index in use, we have determined that "dry" statistical analysis has the potential to give us new knowledge about language. The qualitative study of the groups selected for the analysis in this study and possibly the adjustment of the threshold values of the D-index form an interesting prospect for future research. There are also broader questions arising from this study, for instance: Could the D-index help improve corpus tagging systems? Can it be used in other languages? As an answer to the first question, we suggest that the D-index could help to choose the PoS that is more likely correct in disambiguation processes. The D-index itself is quite readily applicable to other morphologically rich languages, given that the norms of the forms are established.

6. Acknowledgements

We are thankful to the three anonymous referees for their valuable comments on this article. This work was supported by Estonian Research Council grant PSG227.

7. References

Blensenius, K. & von Martens, M. (2019). Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek &

[14] https://cloud.gate.ac.uk/shopfront/displayItem/tagger-pos-et-maxent1

[15] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

C. Tiberius (eds.). *Proceedings of eLex 2019 conference. 1–3 October 2019. Sintra, Portugal*. Brno: Lexical Computing CZ, s.r.o., pp. 660–675.

Brinton, L. J. & Traugott E. C. (2005). *Lexicalization and language change*. Cambridge: CUP. DOI: 10.1017/CBO9780511615962.

CombiDic = *The EKI Combined Dictionary*. (2020). Hein, I., Kallas, J., Kiisla, O., Koppel, K., Langemets, M., Leemets T., Melts, M., Mäearu, S., Paet, T., Päll, P., Raadik, M., Tiits, M., Tsepelina, K., Tuulik, M., Uibo, U., Valdre, T., Viks, Ü. & Voll, P. Institute of the Estonian Language. Accessed at: Sõnaveeb 2020. https://sonaveeb.ee. (5 March 2021)

The Estonian Collocations Dictionary = *Eesti keele naabersõnad*. (2019). Kallas, J., Koppel, K., Paulsen G. & Tuulik, M., Institute of the Estonian Language. Accessed at: http://www.sonaveeb.ee. (14 February 2020)

Ekilex. Accessed at: https://ekilex.eki.ee/ (20 March 2021)

The Explanatory Dictionary of Estonian = *Eesti keele seletav sõnaraamat* I–VI. (2009). M. Langemets, M. Tiits, T. Valdre, L. Veskis, Ü. Viks, P. Voll (eds.). Institute of the Estonian Language. Tallinn: Eesti Keele Sihtasutus. Accessed at: http://www.eki.ee/dict/ekss/. (5 April 2021)

Grünthal, R. (2003). *Finnic Adpositions and Cases in Change*. Suomalais-Ugrilaisen Seuran toimituksia 244. Helsinki: Finno-Ugrian Society.

Habicht, K., Penjam, P. & Prillop, K. (2011). Sõnaliik kui rakenduslik ja lingvistiline probleem: sõnaliikide märgendamine vana kirjakeele korpuses. *Estonian Papers in Applied Linguistics* 7, pp. 19–41.

Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics*, 39(6), pp. 1041–1070.

Heine, B. & Kuteva, T. (2007). *The genesis of grammar. A reconstruction*. Oxford: Oxford University Press.

Jakubíček, M. (2021). Morphology is an open problem of NLP. Talk given at the Workshop on Parts of Speech. Tallinn: Institute of the Estonian Language. Available at: https://portaal.eki.ee/component/content/article/101-projektid/3414-workshop-on-the-role-of-parts-of-speech-in-language-technology.html.

Kaalep, H-J. & Vaino, T. 2001. Complete Morphological Analysis in the Linguist's Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pp. 9—16, Tartu. Available at: http://www.cl.ut.ee/yllitised/smugri_toolbox_2001.pdf.

Kasik, R. (2015). *Sõnamoodustus* [Word formation]. Tartu: Tartu University Press.

Karelson, R. (2005). Taas probleemidest sõnaliigi määramisel [Once again on the problems of assigning the PoS]. *Estonian Papers in Applied Linguistics* 1, 53—70.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Information Technology*, 105, pp. 116–127.

Kittask, C., Milintsevich, K. & Sirts, K. (2020). Evaluating Multilingual Bert for Estonian. In A. Utka, J. Vaičenonienė, J. Kovalevskaitė & D. Kalinauskaitė (eds.). *Human Language Technologies – The Baltic Perspective*. IOS Press, pp.

19−26. (Frontiers in Artificial Intelligence and Applications). DOI: 10.3233/FAIA200597.

Koppel, K. (2020). *Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele* [Corpus-Based Automatic Detection of Example Sentences for Dictionaries for Estonian Learners]. PhD thesis. Tartu: Tartu University Press.

Koppel, K., Tavast, A., Langemets, M. & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Proceedings of the eLex 2019 conference. 1–3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 434−452.

Langemets, M., Kallas, J., Norak, K. & Hein, I. (2020). New Estonian Words and Senses: Detection and Description. *Journal of the Dictionary Society of North America* 41 (1), pp. 69–82.

Laur, S., Orasmaa, S., Särg, D. & Tammo, P. (2020). EstNLTK 1.6: Remastered Estonian NLP Pipeline. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 7152–7160.

Orasmaa, S., Petmanson, T., Tkatšenko, A., Laur, S. & Kaalep, H-J. (2016). EstNLTK – NLP Toolkit for Estonian. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & P. Stelios (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, Slovenia: ELRA, pp. 2460−2466. http://www.lrec-conf.org/proceedings/lrec2016/pdf/332_Paper.pdf

Paulsen, G., Vainik, E., Tuulik, M. & Lohk, A. (2019). The lexicographer's voice: word classes in the digital era. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Proceedings of eLex 2019 conference. 1−3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 319–337.

Paulsen, G.; Vainik, E.; Tuulik, M. (2020). Sõnaliik leksikograafi töölaual: sõnaliikide roll tänapäeva leksikograafias [On word classes in contemporary lexicography: The lexicographers' view]. *Estonian papers in applied linguistics*, 16, pp. 177−202. DOI: 10.5128/ERYa16.11.

Sahkai, H. (2008). Konstruktsioonipõhine keelemudel ja sõnaraamatumudel [A construction-based model of language and dictionary]. *Estonian Papers in Applied Linguistics*, 4, pp. 177−186.

Tavast A., Koppel K., Langemets M. & Kallas J. (2020). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, Vol. 1., Greece: Democritus University of Thrace, pp. 215−223.

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & Simon Krek (eds.) *Proceedings of the XVIII EURALEX International*

Congress: EURALEX: Lexicography in Global Contexts. Ljubljana, Slovenia.

Tkachenko, A. & Sirts, K. (2018). Neural Morphological Tagging for Estonian. In Muischnek, K. & Müürisepp K. (eds.). *Human Language Technologies – The Baltic Perspective*. IOS Press. (Frontiers in Artificial Intelligence and Applications), pp. 166–174. DOI: 10.3233/978-1-61499-912-6-166.

Vainik, E., Paulsen, G. & Lohk, A. (2020). A typology of lexical ambiforms in Estonian. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, Vol. 1. Alexandroupolis, Greece: Democritus University of Thrace, pp. 119–130.

Vainik, E.; Paulsen, G. & Lohk, A. (2021). Käändevormist sõnaks: mida näitab sagedus? [From inflected form to a word: the role of frequency]. Accepted by *Estonian Papers in Applied Linguistics*, 17.

Vainik, E.; Lohk, A. & Paulsen, G. (2021). The Distribution Index Calculator for Estonian. *Proceedings of eLex 2021 conference*. 5–7 July 2021, Brno, Czechia. Brno: Lexical Computing CZ, s.r.o.

Veskis, K.; Liba, E. (2010). Automatic Tagger Evaluation. Syntax assignment report. NGSLT (Nordic graduate school on language technology) NLP course 2008. Available at: http://teataja.ee/veskis-liba-syntax-assignment-modified.pdf

Viitso, T-R. (2003). Structure of the Estonian language: Phonology, morphology, and word formation. In M. Erelt (ed.) *Estonian language*. Tallinn: Estonian Academy Publishers, pp. 9–92.

Viks, Ü. (1992). *Väike vormisõnastik. I: Sissejuhatus & grammatika; II: Sõnastik & lisad* [A Concise Morphological Dictionary of Estonian. I: Introduction & Grammar; II Dictionary and Appendices]. Tallinn.

MOR*Digital*:

The Advent of a New Lexicographic Portuguese Project

Rute Costa[1], Ana Salgado[2], Anas Fahad Khan[3], Sara Carvalho[1,4],

Laurent Romary[5], Bruno Almeida[1,6], Margarida Ramos[1],

Mohamed Khemakhem[7], Raquel Silva[1], Toma Tasovac[8]

[1] NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal
[2] Academia das Ciências de Lisboa, Portugal
[3] Istituto Di Linguistica Computazionale 'A. Zampolli', Italy
[4] CLLC, Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro, Portugal
[5] Inria, team ALMAnaCH, France
[6] ROSSIO Infrastructure, Portugal
[7] Arcascience, France
[8] BCDH – Belgrade Center for Digital Humanities
E-mail: rute.costa@fcsh.unl.pt, anasalgado@campus.fcsh.unl.pt, fahad.khan@ilc.cnr.it, sara.carvalho@ua.pt, laurent.romary@inria.fr, brunoalmeida@fcsh.unl.pt, mvramos@fcsh.unl.pt, medkhemakhemfsegs@gmail.com, raq.asilva@gmail.com, ttasovac@humanistika.org

Abstract

MOR*Digital* is a newly funded Portuguese lexicographic project that aims to produce high-quality and searchable digital versions of the first three editions (1789; 1813; 1823) of the *Diccionario da Lingua Portugueza* by António de Morais Silva, preserving and making accessible this important work of European heritage. This paper will describe the current state of the art, the project, its objectives and the methodology proposed, the latter of which is based on a rigorous linguistic analysis and will also include steps necessary for the ontologisation of knowledge contained in and relating to the text. A section will be dedicated to the various investigation domains of the project description. The output of the project will be made available via a dedicated platform.

Keywords: digital humanities; GROBID-Dictionaries; legacy dictionary; lexicography; ontologies; standards

1. Introduction

The *Diccionario da Lingua Portugueza* by António de Morais Silva, hereafter referred to as Morais, constitutes a considerable piece of cultural heritage since it marks the beginning of modern Portuguese lexicography, serving also as a model for all subsequent lexicographic production throughout the 19th and 20th centuries. In this paper, we present MOR*Digital*, a newly funded Portuguese lexicographic project, which was successfully submitted to the IC&DT 2020 Projects Call under the scientific area of 'information sciences computing', which falls under 'languages and literatures –

linguistics, subarea computer sciences and information sciences'. The project will be funded over the next three years (2021–2024).

The MOR*Digital* project aims to produce high-quality and searchable digital versions of the first three editions (1789; 1813; 1823) of Morais in order to preserve this important European heritage work while also making it accessible. These digital versions will be converted into structured data and made publicly available with the purpose of guaranteeing the preservation of this legacy resource. After an introduction to the dictionary itself, we provide a general outline of the project and detail its main objectives, focusing on the importance of using standards and formats for interoperability purposes. We then explore the research methodology adopted. This methodology for the creation of an open-access Portuguese language dictionary is based on a comprehensive understanding of lexical units and the privileging of a strictly linguistic analysis to create future ontologies that adequately represent the lexical data in the study, in addition to making them accessible and reusable.

This project aims to make a substantial contribution to the scientific community and aspires to apply innovative computational methodologies to digitise lexicographic texts and coding based on a comprehensive analysis of lexicographic articles and their components.

This paper is organised as follows: the first (and current) section introduces and outlines the article. Section 2 reviews the theoretical framework and existing standards. In Section 3, we historically frame our object of study. Section 4 introduces the Morais dictionary. Section 5 describes the MOR*Digital* project, the methodology, as well as tools and formats. Finally, in Section 6, we highlight our future work and present concluding remarks.

2. Theoretical Framework

European lexicography can boast a long tradition of theoretical and descriptive work on dictionaries and especially in the case of historical dictionaries, as is discussed in several works, amongst which Zgusta (1971), Wiegand (1984), Quemada (1987), Atkins and Rundell (2008), Tarp (2008), Durkin (2019) and Considine (2019). These authors have approached lexicography from either a theoretical or methodological perspective, helping to bring to light the paradigm shift we witness in the convergence between lexicography, computational linguistics, digital humanities, and ontologies.

In Portugal, this scientific activity around lexicography work is present in Villalva & Williams (2019), Salgado et al. (2019), Salgado & Costa (2019), Lino (2018), Silvestre (2016), Gonçalves & Banza (2013), Correia (2009) and Verdelho (2003), among others. The *European Dictionary Portal*[1] points to the existence of four online Portuguese dictionaries and a portal. Despite being electronic, most of these resources are

[1] http://www.dictionaryportal.eu/en/

structured and formalised according to a paper-based methodology, and therefore do not fully explore their digital potential. In turn, the *Dicionário Aberto*, one of the dictionaries available on the portal, differs from our objectives, even though it is based on a historical dictionary. This is because the researchers' primary focus (Simões & Farinha, 2009) was not so much preserving the original source but mainly modernizing the dictionary. Thus, and according to the available data, there are no dynamic, open-access resources based on Portuguese heritage dictionaries, so efforts must be made to provide this accessibility to recognised heritage value sources in the form of searchable, dynamic resources.

Lexicography has undergone a radical change in the past two decades, especially with technological advances, the fall of many publishers, as well as the changes introduced into their business models (Rundell, 2010: 170). This paradigm shift is also directly related to the advancement of digital humanities, which quickly became an aggregator of several scientific disciplines. Although the first definitions of the term 'digital humanities' were limited to humanities computing (Terras & Vahouette, 2013), today, these definitions are far from being universally accepted (Gold & Klein, 2016). Instead, the term now covers a variety of lines of research belonging to a number of different disciplines, and is characterised by the use of tools, computational methods and standards, implying, above all, a new general perspective of the humanities in response to the epistemological challenges that these changes impose.

The perspective underpinning the construction of lexical resources that we propose in this project presupposes rethinking the methodologies of the Portuguese lexicographic tradition, perceiving lexicography, terminology, ontologies and computational linguistics as an integral part of the digital humanities, which will imply a paradigm shift in the construction of dictionary resources. In this new paradigm, ontologies will play a key role in organising and representing linguistic and metalinguistic knowledge, bringing added value by providing greater logical consistency in the representation of data (Carvalho et al., 2018; Almeida et al., 2019), as well as supporting its operationalisation and, therefore, its preservation in the long term.

The European lexicographic scenario is currently quite heterogeneous, both in what concerns the types of existing lexicographic resources and their particular structural component, which relates to how the data are represented, the adopted models, as well as the respective applied formats. Each format has its own syntax and vocabulary, defined according to certain parameters to enable the reusability of the lexicographic content. The diversity of incompatible formats creates severe problems in the digital landscape, making it impossible to interconnect resources and their respective metadata and lexical data. Herein lies the importance of following compatible standards and formats such as LMF (ISO 24613: 2008), TEI Lex-0[2] (Tasovac and Romary et al., 2018) and Ontolex-Lemon (McCrae et al., 2017).

[2] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

3. Historical Background

Diccionario da Lingua Portugueza by António de Morais Silva was elaborated during the Age of Enlightenment. This century brought a renewal in several fields of knowledge, namely those concerning the description of living languages, at a time when Latin was still the language of instruction. Dictionaries were perceived as metalinguistic instruments. The 17th century marked a very prolific period in terms of lexicographic production, especially with regard to the French dictionary production (for example, *Dictionnaire françois, contenant les mots et les choses, plusieurs nouvelles remarques sur la langue françoise* (1680) by Father Richelet or *Dictionnaire universel* (1690) by Antoine Furetière), which served as a model for all subsequent lexicographic works.

Portuguese lexicography benefited from this moment, especially with the Morais dictionary's publication in 1789, which inaugurated modern Portuguese lexicography. This dictionary followed the publication of the third edition of the *Vocabolario degli Accademici della Crusca* (1691), the *Dictionnaire de l'Académie Française* (1694) and the *Vocabulario Portuguez and Latino* (1712–1728) by Father Rafael Bluteau. The latter marked the transition between the Latin-Portuguese dictionary and the first Portuguese monolingual dictionary [Morais] (Silvestre, 2008, p. 7), thus paving the way for the emergence of a new way of working in lexicography that would influence subsequent publications, such as the *Diccionario da lingoa portugueza* (1793), published by the Lisbon Science Academy and the *Elucidário das Palavras, Termos e Frases* by Joaquim de Santa Rosa de Viterbo (1798). As Verdelho (2003: 473) mentions, Morais 'laid the foundation to all the lexicographic genealogy developed over the last 200 years' and, according to Biderman (1984: 5), referring to the second edition, 'constitutes a milestone in Portuguese-language lexicography'.

Despite all this, lexicographic production arises late in Portugal when compared with that of other countries. The publication of dictionaries in vernacular languages was already proliferating throughout Europe, as can be seen from the publishing timelines of other monolingual dictionaries.[3]

4. Morais Dictionary

The first edition of the known Morais dictionary is entitled in its main edition (1789) *Diccionario da Lingua Portugueza composto pelo Padre D. Rafael Bluteau Diccionario da Lingua Portugueza composto pelo Padre D. Rafael Bluteau, reformado, e accrescentado por Antonio de Moraes Silva, natural do Rio de Janeiro* [Diccionario da Lingua Portugueza composed by Father D. Rafael Bluteau, retired, and accredited by

[3] Such as the *Tesoro de la lengua castellana*, española by Sebastián de Covarrubias in 1611, which, in addition to being the first Spanish monolingual dictionary, is the first European one. Other examples include the first edition of the *Vocabolario degli Accademici della Crusca*, which was compiled in Florence and printed in Venice in 1612, as well as the French dictionaries mentioned before.

Antonio de Morais Silva, born in Rio de Janeiro], as seen in Figure 1.

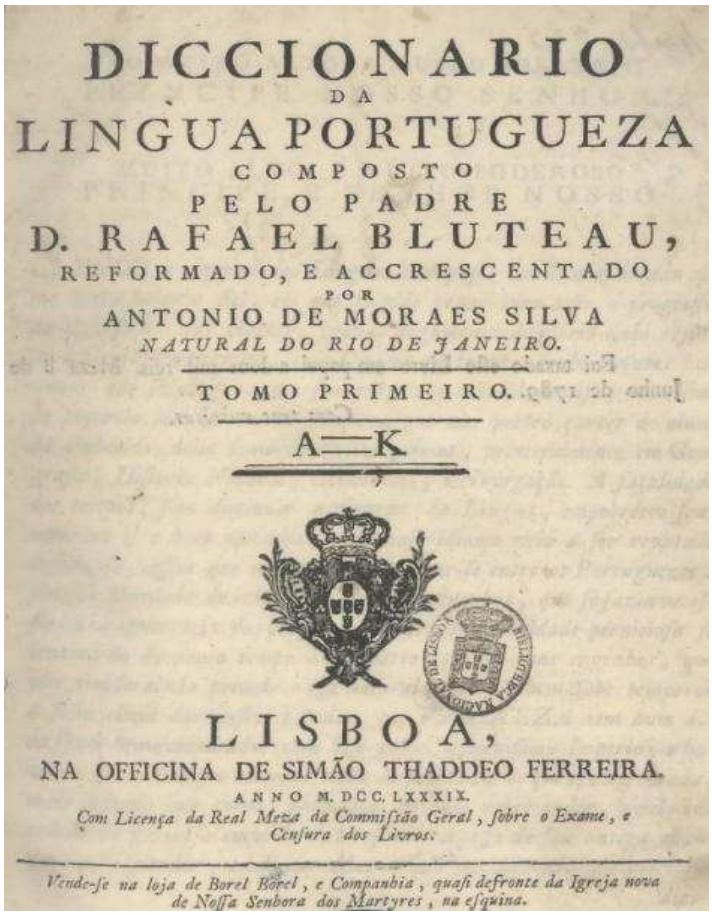
DICCIONARIO
DA
LINGUA PORTUGUEZA
COMPOSTO
PELO PADRE
D. RAFAEL BLUTEAU,
REFORMADO, E ACCRESCENTADO
POR
ANTONIO DE MORAES SILVA
NATURAL DO RIO DE JANEIRO.
TOMO PRIMEIRO.
A—K
LISBOA,
NA OFFICINA DE SIMÃO THADDEO FERREIRA.
ANNO M. DCC. LXXXIX.
Com Licença da Real Meza da Commissão Geral, sobre o Exame, e Censura dos Livros.
Vende-se na loja de Borel Borel, e Companhia, quasi defronte da Igreja nova de Nossa Senhora dos Martyres, na esquina.

Figure 1: Frontispiece of Morais (1789), first volume

The information that immediately stands out concerns the authorship attribution, since Morais does not claim to be the author, assigning this condition to Bluteau, author of the *Vocabulario Portuguez and Latino*. However, Morais recognises in the '*Prólogo ao Leitor*' [Prologue to the Reader] that the additions he brought to the dictionary are quite relevant. Morais further developed Bluteau's work and systematically took into account most of the entries and definitions. Verdelho (2003) considers this attitude inevitable, which, in reality, reflects, '*o que todos os dicionaristas não podem deixar de fazer ao retomar e renovar a nomenclatura dos seus predecessores, uma espécie inevitável de 'plágio por ordem alfabética*' [what all dictionary-makers cannot fail to do when resuming and renewing the nomenclature of their predecessors, an inevitable kind of 'plagiarism in alphabetical order'].

As mentioned above, Morais represents the first modern work to systematise the lexicon of the Portuguese language, a model and example for all the ones that followed. It was also, for almost two centuries, a work of mandatory consultation for Portuguese language, both in Portugal and in Brazil. As Correia (2009) observes, the Morais dictionary '*tornou-se uma referência incontornável para o estudo da evolução do léxico*

do Português, tendo constituído, simultaneamente, um elemento de normalização e mesmo de padronização da língua' [has become an essential reference for the study of the evolution of the Portuguese lexicon, having simultaneously constituted an element of normalisation and even of language standardisation].

The first edition was first published in two volumes: first, from the letters A to K, in a total of 752 pages, and then, from the letters L to Z, with 541 pages. The work was printed at Simão Thaddeo Ferreira's publishing house, in Lisbon.

The following two editions (1813; 1823) are considered new dictionaries, due to both their enrichment and the updating. The second edition, corrected and enlarged in two volumes (A–E; F–Z), was also published in Lisbon, in Typographia Lacerdina. Morais claims the authorship of the dictionary on the title page, where the work is presented as the *Diccionario da Lingua Portugueza, recopilado dos vocabularios impressos ate agora, e nesta segunda edição novamente emendado, e muito accrescentado, por Antonio de Moraes Silva natural do Rio de Janeiro* [*Diccionario da Lingua Portugueza*, compiled from the vocabularies printed so far, and in this second edition, again amended and incredibly enriched by Antonio de Moraes Silva]. The same happened to the third edition, coordinated by Pedro José de Figueiredo, who expanded it from five to six thousand articles, as stated in the title.

The author died the following year, in 1824. The work continued to be published and enhanced over the years until 1949. From then to 1959, in 12 volumes, the tenth edition was prepared, under the coordination of Augusto Moreno, Cardoso Júnior and José Pedro Machado, but maintaining Morais as the author.

Even though the Morais dictionary is available on some web pages (e.g. CEPESE[4]), it is provided as a PDF document, resulting from the digitisation of the work on paper. This format does not take great advantage of the digital environment and its potential, since it does not allow advanced searches. It is this issue that we intend to explore in our project.

5. MOR*Digital*

5.1 The Project

As stated in the introduction, the main goal of MOR*Digital*[5] is to encode the selected editions of *Diccionario de Lingua Portugueza* by António de Morais Silva. MOR*Digital* aims to promote accessibility to cultural heritage while fostering reusability and contributing towards a greater presence of lexicographic digital content in Portuguese through open tools and standards. MOR*Digital* follows a new paradigm in lexicography,

[4] https://www.cepese.pt/portal/pt/bases-de-dados/dicionario/apresentacao

[5] MORDigital – Digitalização do *Diccionario da Lingua Portugueza* de António de Morais Silva [PTDC/LLT-LIN/6841/2020]

which results from the convergence of lexicography, terminology, computational linguistics, and ontologies as an integral part of digital humanities and Linked (Open) Data.

In this project, we connect data and metadata within the same lexicographic resource and between different resources, through the Web of Data, which is based on principles structured around the use of RDF, URIs and SPARQL, a language for querying and retrieving information. Underlying the formalisation and application of the standards is the linguistic and lexicographic knowledge that permeates the entire project and contributes to the necessary systematisation of data and metadata. Being a project dedicated to Portuguese, it has the added value of bringing a historical resource into the LLOD cloud in a language that is still underrepresented.

Retrodigitising historical dictionaries into machine-readable dictionaries poses several challenges that the scientific community has tried to resolve by creating tools, different formats, and establishing standards, following the FAIR[6] principles for modelling lexical resources and making them available.

Our starting point will be the Morais digitisations available as PDF at the Portuguese National Library and the Brasiliana Library[7]. However, the lack of quality of the available PDF may lead us to undertake a new digitisation process of Morais. High-quality digitisation is required to use GROBID-Dictionaries (Khemakhem, Foppiano, Romary, 2017, Khemakhem et al., 2019), a machine learning system for converting PDF into the TEI/XML format and structuring the content of the digitised versions of the dictionaries.

Following current open data best practices, the main goal is to put forward a methodology that can be replicated in other legacy paper dictionaries, using tools that allow the automatic extraction of lexicographic content, as well as the modernisation of the spelling in an automated way.

5.2 Methodology

MOR*Digital* proposes to: (i) analyse all components that comprise the dictionary's macro- and microstructure; ii) identify, organise and describe the different levels of linguistic knowledge to apply the aforementioned standards systematically; (iii) develop methodologies that can be replicated for other applications and test the alignment of the different encodings of Morais; (iv) participate in reviewing the corresponding standards as members of the standard bodies and scientific forums; (v) propose best practices for harmonising the encoding of lexicographic resources; (vi) make Morais available via an open-access platform.

[6] Findable, Accessible, Interoperable, Reusable; cf. Wilkinson et al. (2016).
[7] http://dicionarios.bbm.usp.br/pt-br/dicionario/edicao/2

Our methodology is based on 5 central axes:

> (1) high-quality retrodigitisation of Morais and automatic structuring of the lexical content for the creation of a computer-readable resource;
>
> (2) lexicographically-oriented language description;
>
> (3) Morais encoding, using the TEI Lex-0 specifications mapped to the LMF standard and their respective serialisations, as well as to OntoLex-Lemon;
>
> (4) creation of an ontology for alignment purposes;
>
> (5) and conception of a platform for Morais, enriched with both lexicographic and ontological modules.

All defined tasks will be accomplished successively and managed through subtask assignments, which will be carried out either simultaneously or sequentially, depending on their nature.

We will initiate by surveying the dictionary sources and by a prior evaluation of the quality of digitised versions of these sources (paper to text), for the extraction of lexical information (text to structure). Firstly, this involves transforming the native encoding format into a TEI/XML compliant one (the encoding will be based on TEI standards according to the TEI Lex-0 specification) and LMF metamodels into advanced techniques for semi-structured text acquisition.

The result will be a model of a historical dictionary whose entries are structured in a standard format, namely TEI Lex-0. We plan to adapt the system's cascading architecture to allow the extraction of the different TEI constructs corresponding to the lexicographic structures and conventions. The outcome is a chain of cascading machine learning models, trained and evaluated against manually annotated data. Once the source is digitised, further corrected and marked-up, it will be compared to precedent and subsequent versions, and a series of queries will be conducted to extract all available information about labels. We will then convert TEI Lex-0 datasets into RDF by means of the W3C recommendation for publishing lexicons as Linked Data, namely OntoLex-Lemon. More specifically, we intend to test the implementation of the lexicography module of the Lexicon Model for Ontologies (lexicog)[8], which was recently specified by the Ontology Lexicon community group of the W3C. This will allow for the publication of the Morais datasets as LOD graphs, enabling further NLP applications.

A further step will be the creation of an ontology of all the previously identified and systematised labels (e.g. domain, register, grammar, among others). This will be

[8] https://www.w3.org/2019/09/lexicog/

implemented by resorting to Protégé[9], a free, open-source ontology editor. The ontology will be represented in OWL.

The next step is the alignment of the dictionary versions, which will be carried out in stages: i) alignment of the entries; ii) alignment of the senses; iii) alignment of other lexicographic content.

During the testing phase, formally controlled tests will be carried out to discover errors and bugs that need to be resolved. Finally, we will build a platform that integrates all Morais versions while also mapping the different heterogeneous annotation models, in order to provide access to high-quality digital lexicographic content enhanced by ontologies.

Thus, the search functionalities will include basic and advanced queries, namely searches by lexical relations. A specialised team will be hired to build and develop the interface. Its robustness will be tested according to the types of functionalities defined on validation tests. The alignment between the various editions will be searchable, and the scanned pages made available. In another module, where there will be considerable investment by the team, it is intended that the lexicographic content can be deconstructed and organised in the form of an ontology. We will develop advanced search engines (search for entries by different labels or lexical relations). As part of the aforementioned platform, we will include a section to promote training for the sustainable development of lexicographic resources. This will foster both the qualification of Portuguese lexicographers as well as the users' linguistic knowledge. Moreover, this will provide quality data for researchers.

We aim for our lexical resources to maintain the original spelling. However, making a resource available to the public today, and considering the prevalence of search engines, requires the modernisation of the spellings, especially at the lemma level. The original spelling of the lemma will have to be aligned with more current spellings. To this end, the original forms will be noted as a lemma, but we will first match them with the most current spellings and simultaneously work on their encoding in the XML annotation file. This topic represents the added value of enabling reduplication in other related works, since the correspondences between the lexical units and their respective coding can be reused. We will subsequently create a correspondence between the MOR spellings and the spellings in accordance with the 1945 Luso-Brazilian Convention[10] and the 1990 Portuguese Spelling Agreement[11], taking advantage of work previously developed by one of the team members on the *Vocabulário Ortográfico da Língua Portuguesa* (VOLP-ACL) [Portuguese Language Spelling Vocabulary] of the Lisbon Science Academy[12]. The result will allow the end-user to search the current spellings,

[9] https://protege.stanford.edu

[10] http://www.portaldalinguaportuguesa.org/?action=acordo&version=1945

[11] https://dre.pt/application/file/a/403254

[12] Available at https://www.volp-acl.pt/

with which he/she is familiar, and find the entry corresponding to the old spelling, which will thus remain faithful to the original.

The way we look at Morais transcends the traditional concept of dictionary and is in line with the evolution of e-lexicography itself. We will take advantage of standard formats and linked data technologies for encoding dictionaries, which will allow us to abandon, once and for all, the editorial perspective that is still present in most digital resources. To achieve our goal, we also believe it is necessary to put forward methodologies for improving the quality of lexicographic descriptions.

At the end of the project, we expect to have encoded a vital heritage dictionary, compliant with the most advanced standards for scholarly digital editions and made available via an open licence. The versions will be accessible and searchable through an advanced interface, which will enable the selective querying of text by lemma and type of lexicographic content. The source data will be made available separately from the querying interface, both for research and long-term preservation. Thus, the project will have significantly contributed towards the analysis and annotation of dictionaries through computer-assisted processes.

6. Concluding remarks

This project will represent a substantial contribution to the scientific community, aiming to create innovative and data-driven computational methods for text digitisation and encoding, based on a comprehensive analysis of lexicographic articles and their respective components. Tests on automatic text capture will refine processes and techniques, advancing the state of the art regarding semantic annotation of semi-structured documents. A rigorous linguistic treatment will make it possible to organise and structure the lexicographic components, and to elicit lexical relationships between various elements. The linking mechanisms of the resulting structured dictionary to other resources will constitute a prototype that can be replicated in other works, namely in the Portuguese-speaking world.

MOR*Digital* will be a user-friendly, open-access web interface, equipped with a robust research system that will not only facilitate the search on a more traditional lexicographic perspective but will also allow undertaking research on various types of structured lexicographic and terminological information (Costa et al., 2020). Combining semasiological and onomasiological approaches applied to the three editions of Morais will be possible via the inclusion of ontologies (e.g. diasystematic marking, namely domain labels, registers and part of speech categories). This method will make a new type of dictionary emerge which will contribute to creating a digital linguistic resource that is central to digital humanities. End-users will be predominantly scholars dealing with language and historical issues.

7. Acknowledgements

This paper is supported by (1) the MOR*Digital – Digitalização do Diccionario da Lingua Portugueza de António de Morais Silva* [PTDC/LLT-LIN/6841/2020] project financed by the Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia (2) Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020 and (3) the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS) (European Lexicographic Infrastructure).

8. References

Almeida, B., Costa, R. & Roche, C. (2019). The names of lighting artefacts: extraction and representation of Portuguese and Spanish terms in the archaeology of al-Andalus. *Revue TAL, 60*(3), pp. 113–137.

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

Biderman, M. T. C. (1984). A Ciência da Lexicografia. *Alfa*, n. 28, Brasil: São Paulo, pp. 1-26.

Carvalho, S., Costa, R. & Roche, C. (2018). The Role of Conceptual Relations in the Drafting of Natural Language Definitions: an Example from the Biomedical Domain. In I. Kernerman & S. Krek (eds.), *Proceedings of the LREC 2018 Workshop 'Globalex 2018 – Lexicography & WordNets'.* Miyazaki: European Language Resources Association (ELRA), pp. 10–16. ISBN 979-10-95546-28-3.

Considine, J. (2019). *The Cambridge World History of Lexicography.* Cambridge: Cambridge University Press.

Correia, M. (2009). *Os Dicionários Portugueses.* Coleção: O Essencial Sobre Língua Portuguesa. Lisboa: Editorial Caminho.

Costa, R., Carvalho, S., Salgado, A., Simões, A. & Tasovac, T. (2020). Ontologie des marques de domaines appliquée aux dictionnaires de langue générale. In Xavier Blanco (ed.), La lexicographie en tant que méthodologie de recherche en linguistique *Revue de Philologie Française et Romane - Langue(s) & Parole*, n. 55. Mons: Edition du CIPA, pp. 201-230.

Durkin, P. (ed.) (2019). *The Oxford Handbook of Lexicography.* ISBN: 9780199691630. DOI: 10.1093/oxfordhb/9780199691630.001.0001.

Gold, K. M. & Klein L. F. (eds.) (2016). *Debates in the Digital Humanities.* Mineápolis: University of Minnesota Press.

Gonçalves, M. F. & Banza, A. P. (2013). Fontes de metalinguísticas para a história do português clássico – O caso das Reflexões sobre a Lingua Portugueza. In M. F. Gonçalves e A. P. Banza (coord.), *Património Textual e Humanidades Digitais: da antiga à Nova Filologia*, pp. 73–111. Col. Biblioteca – Estudos & Colóquios, Série ebook, n. 1. Évora: CIDEHUS.

ISO 24613. 2008. *Language resource management – Lexical markup framework (LMF).* Geneva: ISO.

Khemakhem, M., Galleron, I., Williams, G. Romary, L. & Suárez, P. J. O. (2019). How OCR Performance Can Impact on the Automatic Extraction of Dictionary Content Structures. In *19th Annual Conference and Members' Meeting of the Text Encoding Initiative Consortium*. Austria: Graz. https://hal.archives-ouvertes.fr/hal-02263276.

Khemakhem, M., Foppiano, L. & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources Using Conditional Random Fields. In *Proceedings of eLex 2017 Conference: Electronic lexicography in the 21st century: Lexicography from Scratch*. Netherlands: Leiden, pp. 598–613.

Lino, T. (2018). Portuguese lexicography in the internet era. In P. Fuertes-Oliveira (ed.), *The Routledge Handbook of Lexicography*. Abingdon: Routledge, [n.a.]. ISBN 9781138941601.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 Conference: Electronic lexicography in the 21st century: Lexicography from Scratch*. Netherlands: Leiden, pp. 587–597.

Morais: Silva, António de Morais (1789). *Diccionario da lingua portugueza composto pelo padre D. Rafael Bluteau, reformado, e accrescentado por Antonio de Moraes Silva, natural do Rio de Janeiro*, 2 vols. Lisboa: Officina de Simão Thaddeo Ferreira. [For the purpose of this project, other editions will be consulted.]

Quemada, B. (1987). Notes sur lexicographie et dictionnairique. *Cahiers de Lexicologie*, v. 51, n. 2, pp. 229–242. Paris.

Rundell, M. (2010). What future for the learner's dictionary? I. J. Kernerman & P. Bogaards (eds.), *English Learners' Dictionaries at the DSNA 2009*. Jerusalem: Kdictionaries, pp. 169–175.

Salgado, A. Costa, R. & Tasovac, T. (2019). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. In *Lexicography: Journal of ASIALEX 6* (2), pp. 133–156. DOI: https://doi.org/10.1007/s40607-019-00061-x.

Salgado, A. & Costa, R. (2019). Marcas temáticas en los diccionarios académicos ibéricos: estudio comparativo. *RILEX. Revista sobre investigaciones léxicas 2 (2)*, pp. 37–63. DOI: http://dx.doi.org/10.17561/rilex.v2.n2.2.

Silvestre, J. P. (2008). *Bluteau e as Origens da Lexicografia Moderna*. Lisboa: INCM.

Silvestre, J. P. (2016). Lexicografia. In A. M. Martins & E. Carrilho (eds.). *Manual de Linguística Portuguesa*. Berlin: De Gruyter Mouton, pp. 200–223.

Simões, A. & Farinha, R. (2009). Dicionário Aberto: um recurso para processamento de linguagem natural. In *Viceversa: Revista Galega de Traducción*, v. 16. Spain: Vigo, pp. 159–171.

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge.* Tübingen: Niemeyer.

Tasovac, T. & Romary, L., et al. (2018). *TEI Lex-0: A baseline encoding for lexicographic data*. Version 0.8.6. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Terras, M., Nyhan, J. & Vahouette, E. (eds.) (2013). *Defining Digital Humanities: A*

Reader. London: Ashgate.

Verdelho, T. (2003). O Dicionário de Morais Silva e o Início da Lexicografia Moderna. *História Da Língua e História Da Gramática – Actas do Encontro*: 473–490. Braga: ILCH, Universidade do Minho.

Villalva, A. & Williams, G. (2019). *The Landscape of Lexicography*. Lisboa–Aveiro: Centro de Linguística da Universidade de Lisboa–Universidade de Aveiro.

Wiegand, H. E. (1984). On the Structure and Contents of a General Theory of Lexicography. In R. R. K. Hartmann (ed.), *LEXeter'83 Proceedings*, pp. 13–30.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data3*:160018. DOI: 10.1038/sdata.2016.18.

Zgusta, L. (1971). *Manual of Lexicography*. Prague: Academia/The Hague: Mouton.