EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Book of Abstracts

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

The XIX EURALEX International Congress: Lexicography for inclusion

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100
e-edition

Publication is free of charge

Mapping domain labels of dictionaries

Ana Salgado[1,2], Rute Costa[1], Toma Tasovac[3]

[1] *NOVA CLUNL Universidade NOVA de Lisboa, Lisbon, Portugal*
[2] *Academia das Ciências de Lisboa, Lisbon, Portugal*
[3] *Belgrade Center for Digital Humanities, Belgrade, Serbia*

Abstract

The purpose of this paper is to compare and analyse the use of domain labels in three large scholarly dictionaries – *Dicionário da Língua Portuguesa Contemporânea* (DLPC), published in 2001 by the Academia das Ciências de Lisboa (ACL); the 23th online edition of *Diccionario de la lengua española* (DLE), published by Real Academia Española (RAE); and the 9th online edition of *Dictionnaire de l'Académie Française* (DAF), a work in progress – in order to a) highlight the commonalities and differences in their editorial practices and approaches to knowledge organisation; b) report on a mapping exercise for a particular domain (GEOLOGY) which can serve as a test case for establishing procedural rules for the alignment of domain labels in general language dictionaries. We show how "meta-labels" can be used to optimise the alignment of specialised senses in lexicographic works.

General dictionaries register, describe and define the specialised senses of lexical items, or terms, specific to different areas of knowledge. As a result of technological changes, the evolution of society, and globalisation, the number of terms found in dictionaries entry has increased (Wiegand, 1984, Boulanger and L'Homme, 1991 and Ahumada, 2002). The labels assigned to these specialised senses are called "domain labels". As markers which identify the specialised field of knowledge in which a lexical unit is mainly used (Salgado et al., 2019), domain labels can serve multiple functions: aiding lexicographers by providing specific information and by identifying specialised lexica in general language dictionaries that can serve as terminology control mechanisms; facilitating user searches by grouping lexical items according to a field so that the user can determine beforehand if the complete lexicographic article is relevant for them; facilitating end-user word sense disambiguation tasks; facilitating terminology extraction in diverse languages; enhancing machine translation and NLP projects.

A domain can be the name of a field in which a specific knowledge area is developed (GEOLOGY) or the specific object of the knowledge area (SHOEMAKING). Lexicographers often make subjective assignments according to a certain tradition they subscribe to (Ptaszynski, 2010, p. 413). For example, the dictionaries we analysed contain labels for domains such as CYNEGETICS (DLPC, DLE) and HUNTING (DAF) but not for MANAGEMENT or TOURISM.

All three Academy dictionaries lack explicit explanatory information regarding their labelling practices (Salgado et al., 2019). Our previous work on DLPC (Salgado and Costa, 2019) has already detected the problematic use of: i) domains with multiple labels, for example, football terms were found to be classified under the SPORT and FOOTBALL labels in DLPC (e.g. *líbero* [sweeper] in SPORT and *lateral* [back] in FOOTBALL); ii) unlabeled equivalent headwords, for example, *paleozóico* [palaeozoic] *adj.* is unlabeled and *primário* [primary] *adj.*, a synonym, appears with a GEOLOGY label; iii) combinations of labels referring to closely related domains, such as *antracite* [anthracite] being associated with both MINERALOGY) and GEOLOGY or *glaciar* [glacier] being associated with both the GEOLOGY and GEOGRAPHY domains. Such inconsistencies can lead to numerous issues that complicate the sharing, aligning, and linking data.

Atkins and Rundell (2008) argue that instead of conceiving "a totally 'flat' (non-hierarchical list of domains)", "it is more practicable to try to build a domain list with a certain hierarchical structure" (p. 184). Applying previously organised hierarchical structure is advantageous both when composing and

when editing a lexicographic resource because it helps the lexicographer control the terminology. The geology domain was reorganised to illustrate examples of existing frameworks (WordNet Domains Hierarchy[1]; Dewey Decimal Classification[2]).

In this paper, we will present the theoretical framework, a threefold methodology and the analysis of the chosen domain:

1) Theoretical framework: The theoretical framework upon which this research is based is summarised to provide related background information (e. g. assumptions about domain labelling by Atkins and Rundell (2008); labelling classifications by Hausmann (1989), followed by Svensén (2009); works on WordNet domains by Magnini and Cavaglià (2000), Bentivogli et al. (2004), Gella et al. (2014)), and to argue for a conceptual modelling based on ISO standards (704:2009; 1087:2019) for terminology.
2) The methodology applied in this research:
   i) Monolingual dictionaries were chosen due to their highly discursive properties. Academy dictionaries were selected for study due to their authoritativeness.
   ii) Datasets were compiled manually from dictionary abbreviation lists. Three hundred eighty-seven multilingual domain labels were collected. There were 184, 74, and 237 domain labels in DLPC, DLE, and DAF, respectively. Generic domains and subdomains coexisted. We noted the case of MATHEMATICS and its sub-domains ALGEBRA (DLPC, DAF), ARITHMETIC (DLPC, DAF), GEOMETRY (DLPC, DLE, DAF) and TRIGONOMETRY (DLPC) or STATISTICS (DLE, DAF). In our comparison, a flagrant imbalance in the number of domains was found: the DLE contains generic domains alone, whereas the DLPC and DAF register multiple subdomains and even multiple labels for the same or very similar domains (e.g. COURSES DE CHEVAUX and COURSES HIPPIQUES [horse races] in DAF).
   iii) In order to systematise the labels and to detect overlapping, the compiled domain label lists were compared. The DLPC list was set as baseline, against which the DLE and DAF counterparts were compared. DLE and DAF were also separately compared. Domain labels were manually mapped using semantic properties such as "exact" and "related" (to a generic domain) and "none". The equivalent English term was assigned as the "meta-label" of the corresponding domain (Table 1 and Appendices).

DLPC	RELATION	DLE	RELATION	DAF	METALABEL
Acústica	EXACT	acústica	EXACT	Acoustique	acoustics
Aeronáutica	EXACT	aeronáutica	EXACT	Aéronautique	aeronautics
Agricultura	EXACT	agricultura	EXACT	Agriculture	agriculture
Anatomia	EXACT	anatomía	EXACT	Anatomie	anatomy
Antropologia	EXACT	antropología	EXACT	Anthropologie	anthropology
Arqueologia	EXACT	arqueologia	EXACT	Archéologie	archeology
Arquitectura	EXACT	arquitectura	EXACT	Architecture	architecture
Astrologia	EXACT	astrologia	EXACT	Astrologie	astrology
Astronomia	EXACT	astronomía	EXACT	Astronomie	astronomy

TABLE 1 – A fragment of domain labels with an "exact" correspondence – 61 domains were mapped to an equivalent domain.

3) Domain analysis: Example entries are presented from the domain GEOLOGY. Using the DLPC as the baseline, this domain was found to have branches that were considered subdomains of a generic domain. GEOLOGY include CRYSTALLOGRAPHY, MINERALOGY, and PALAEONTOLOGY. The corresponding dictionary definitions for each of these terms were compared to clarify, if possible, the underlying reasoning for these subdivisions.

The multilingual domain map constructed in this study will support future standardisation efforts. Standardisation of the domain labelling process and associated encoding tasks are required in order to achieve structured, organised, accessible, and interoperable lexical resources.

[1] http://wndomains.fbk.eu/hierarchy.html
[2] http://www.gutenberg.org/files/12513/12513-h/12513-h.htm

APPENDICES

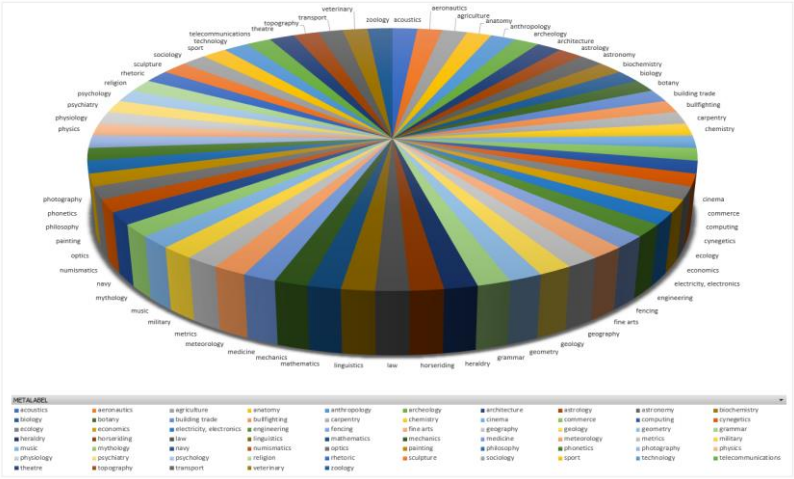

Fig. 1 DLPC vs. DLE – Correspondence between domain labels (65)

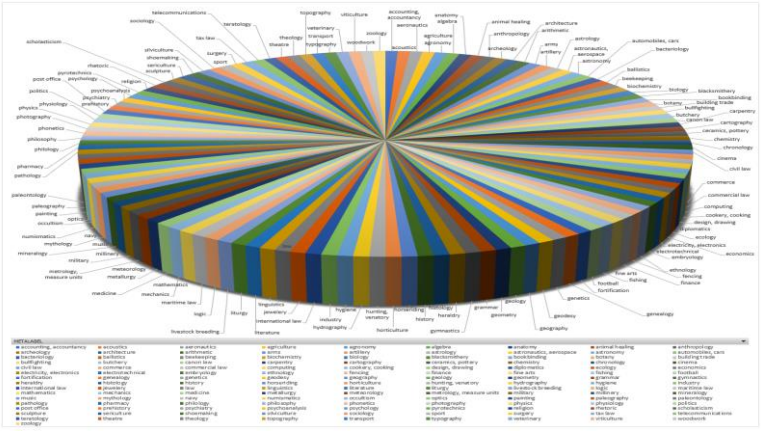

Fig. 2 DLPC vs. DAF – Correspondence between domain labels (136)

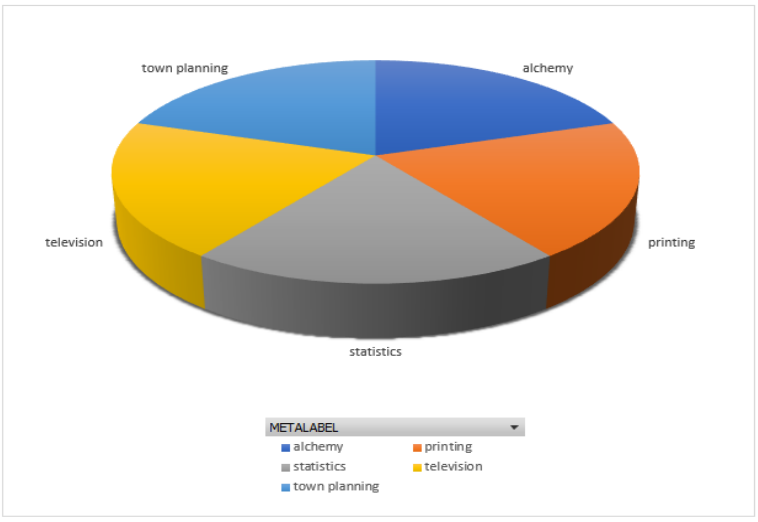

Fig. 3 DLE vs. DAF – Correspondence between domain labels (5)

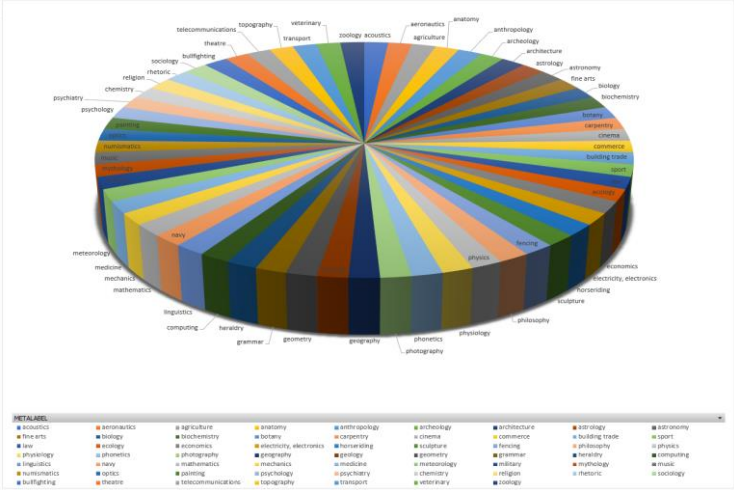

Fig. 4 DLPC vs. DLE vs. DAF – Correspondence between domain labels (61)

Acknowledgements

Research was financed by 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 by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 731015 (ELEXIS). The authors would also like to thank ILex (Institute of Lexicography of the RAE) for granting us access to their data and for hosting Ms. Salgado's 3-week research visit within the scope of an ELEXIS grant. We would also like to thank the Académie Française for sharing domain lists from the current version of their digital dictionary (February 2020, 9th edition) – work completed up to letter S (savoir).

Keywords: Academy of Sciences dictionaries, domain label, Lexicography, Terminology

References

Dictionaries

Dicionário da Língua Portuguesa Contemporânea. 2001. João Malaca Casteleiro (coord.), 2 vols. Lisboa: Academia das Ciências de Lisboa and Editorial Verbo. New digital edition under revision, Ana Salgado (coord.).

Diccionario de la Lengua Española (24.ª ed.). Real Academia Española, 2001–2020, www.rae.es/rae.

Dictionnaire de l'Académie Française (9.ª ed.). Académie Française, 2020, http://www.dictionnaire-academie.fr/.

Other literature

Ahumada, I. (ed.) (2002). Diccionarios y lenguas de especialidad. Jaén: Universidad de Jaén.

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

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In COLING 2004, Workshop on Multilingual Linguistic Resources, Geneva, Switzerland, August 28, 101–108.

Boulanger, J. C. and L'Homme (1991). Les techonolectes dans la pratique dictionnaire générale: quelques fragments d'une culture. In Meta, vol. 36(1), 23–40.

Gella, S., Strapparava, C., and Nastase, V. (2014). Mapping WordNet Domains, WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Iceland: Reykjavik, 1117–1121.

Hausmann, F. J. (1989). Die Markierung in eineim allgemeinen einsprachigen Wörterbuch: eine Übersicht. In F. J. Hausmann, O. Reichmann, H. E. Wiegand and L. Zgusta (eds.). Wörterbücher. Ein internationales Handbuch zur Lexikographie, 649–657. Berlin: Walter de Gruyter.

International Organization for Standardization. 2009. ISO 704: Terminology Work – Principles and Methods. Geneva: ISO.
International Organization for Standardization. 2019. ISO 1087–1: Terminology Work – Vocabulary – Part 1: Theory and Application. Geneva: ISO.
Magnini, B., and Cavaglià, G. (2000). Integrating Subject Field Codes into WordNet. In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (eds.). Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May–2 June 2000, 1413–1418.
Ptaszynski, M. O. (2010). Theoretical Considerations for the Improvement of Usage Labelling in Dictionaries: A Combined Formal-Functional Approach. In International Journal of Lexicography, Volume 23, Issue 4, December 2010, 411–442. DOI: https://doi.org/10.1093/ijl/ecq029.
Salgado, A., and Costa, R. (2019). Marcas temáticas en los diccionarios académicos ibéricos: estudio comparativo. RILEX. Revista Sobre Investigaciones léxicas, 2(2), 37–63. DOI: https://doi.org/10.17561/rilex.v2.n2.2.
Salgado, A., Costa, R., and Tasovac, T. (2019). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. Lexicography ASIALEX 6, 133–156. DOI: https://doi.org/10.1007/s40607-019-00061-x.
Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary Making. Cambridge: Cambridge University Press.
Wiegand, H. E. (1984). On the Structure and Contents of a General Theory of Lexicography. In R.R.K. Hartmann (ed.), LEXeter'83. Tübingen: Max Niemeyer, 13–30.