LexModel – Core Terminology for Lexicography

Rute Costa, Christophe Roche, Ana Salgado

LexModel – Core Terminology for Lexicography[1]

Rute Costa[1], Christophe Roche[1,2], Ana Salgado[1,3]
[1] NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal
[2] Condillac Research Group – LISTIC Lab Université Savoie Mont Blanc
[3] Academia das Ciências de Lisboa, Lisbon, Portugal
rute.costa@fcsh.unl.pt
roche@univ-savoie.fr
anasalgado@fcsh.unl.pt

1. Introduction

This article is the result of the work that has been undertaken in the context of the ELEXIS project European Lexicographic Infrastructure [Horizon 2020 – ID 731015] related to the WP5 which aim is to propose an ELEXIS curriculum[2]. In this ELEXIS curriculum design, we were responsible for the module "Standards for Representing Lexical Data: An Overview".

Through this article, we propose the definition of a core terminology for lexicography aiming to achieve a data model. It comprises two main parts: in the first one, Standards and Formats, we describe the most common standards and formats used by the lexicographic community in order to become familiar with the best practices for representing lexicographic data; in the second part, we present definitions of concepts designated by the terms that make up the core terminology to lexicographic work. After defining the concept in formal language using UML (Unified Modelling Language) diagrams, the proposed definitions are given in natural language. Finally, we illustrate a lexicographic article modelled with Ontolex Lemon, TEI/TEI Lex-0 and LMF.

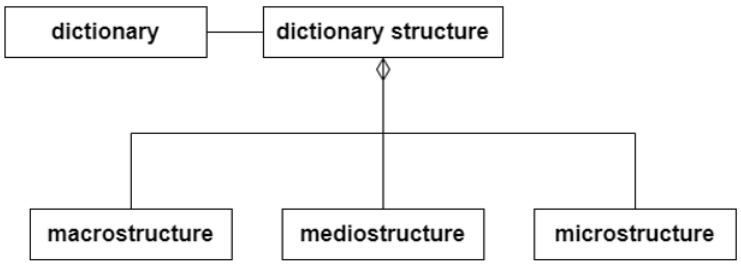
The main objective is to give the formal and natural language definitions of the core concepts, putting them side by side, to facilitate the understanding of a larger community dealing with dictionaries, traditional lexicographers, and NLP and ontology communities. The set of definitions is expressed by resorting to context-free grammar and to ISO 704 (2009) best practices such as the UML notation for defining concepts or ISO 24156-1 (2014).

2. The core terminology for lexicographic work: some examples

In the final document we are working on, we will present 19 definitions of concepts. The concepts are defined using three different approaches in order to serve several purposes. The graphical representation allows identifying the main concepts and their relationships when the formal language (context-free language) guarantees consistency. Finally, the definition in natural language provides a human-readable format.

To exemplify, we present 4 definitions out of the 19 that will be included in the end document: dictionary structure, lexicographic article, lemma and part of speech.

[1] This article is under development.
[2] ELEXIS_D5_2_Guidelines_for_Producing_ELEXIS_Tutorials_and_Instruction_Manuals



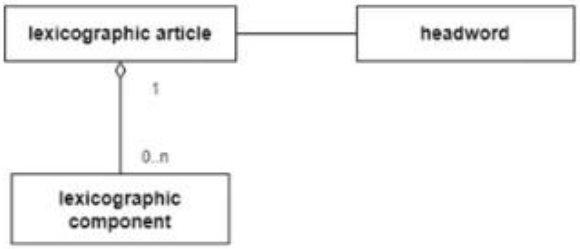
dictionary structure
structure containing a *macrostructure* (1.4), a *microstructure* (1.5) and a *mediostructure* (1.6)

Syn. megastructure

Note 1 to entry: The dictionary structure or megastructure – the dictionary as a whole, referring to the general structure of the parts that compose it – comprises two different sections: the first is the main body of the dictionary and the second is its outside matter.
Note 2 to entry: The outside matter includes the front, middle and back matter.

DIcStruct-> MacroStruct MedioStruct MicroStruct

Figure 1: dictionary structure

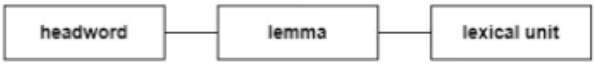


lexicographic article
structured set of *lexicographic components* (1.9) related to a *headword* (1.10)

Syn. entry

LA -> HW LCC*
HW: non terminal for head word
any LA contains one and only one headword and any number of lexicographic component components

Figure 2: lexicographic article

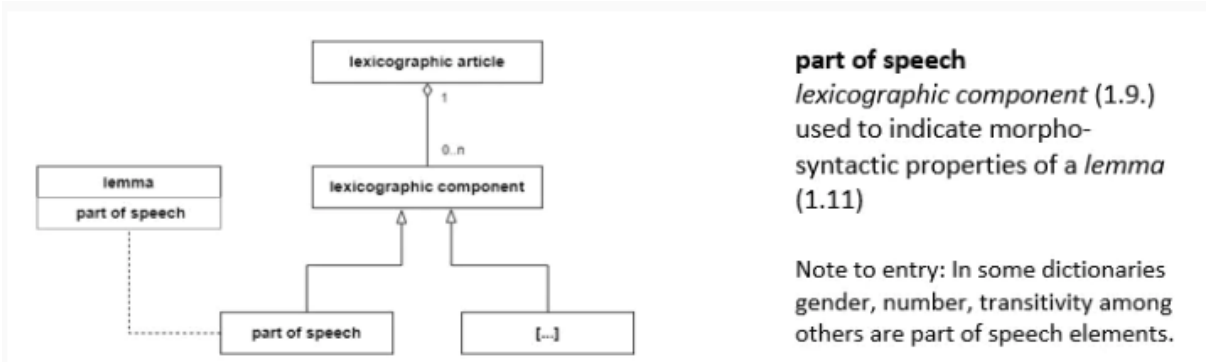


lemma
conventional representation of a *lexical unit* (1.8) that serves as a *headword* (1.10)

Note 1 to entry: A canonical form is a type of conventional representation.
Note 2 to entry: The relevant grammatical conventions may vary between languages and from disciplines such as terminology work (3.5.1), lexicography or documentation. [Source: ISO:24611:2012].
Note 3 to entry: In European languages, the lemma is usually the singular if there is a variation in number, the masculine form if there is a variation in gender and the infinitive for all verbs. [ISO 24611:2012].

Lemma -> Char*

Figure 3: lemma



part of speech
lexicographic component (1.9.) used to indicate morpho-syntactic properties of a *lemma* (1.11)

Note to entry: In some dictionaries gender, number, transitivity among others are part of speech elements.

Figure 4: part of speech

3. Lexicographic article

According to the core vocabulary we have defined, we propose a minimal microstructure of a lexicographic article in XML.

```
<lexicographic_article>
        <headword> … </headword>
        <part_of-speech> … </part_of_speech>
        <sense>
                <lexicographic_definition> … </lexicographic_definition>
                <example> … </example>
                <note> … </note>
        </sense>
        <sense>
                <lexicographic_definition> … </lexicographic_definition>
                <example> … </example>
                <note> … </note>
        </sense>
</lexicographic_article>
```

Figure 5: Minimal microstructure of a lexicographic article

4. LexModel – Core Terminology for Lexicography

The LexModel defines the core terminology for organising, structuring, and representing monolingual dictionaries and has been elaborated for teaching, training, and learning purposes in the context of ELEXIS curriculum design.

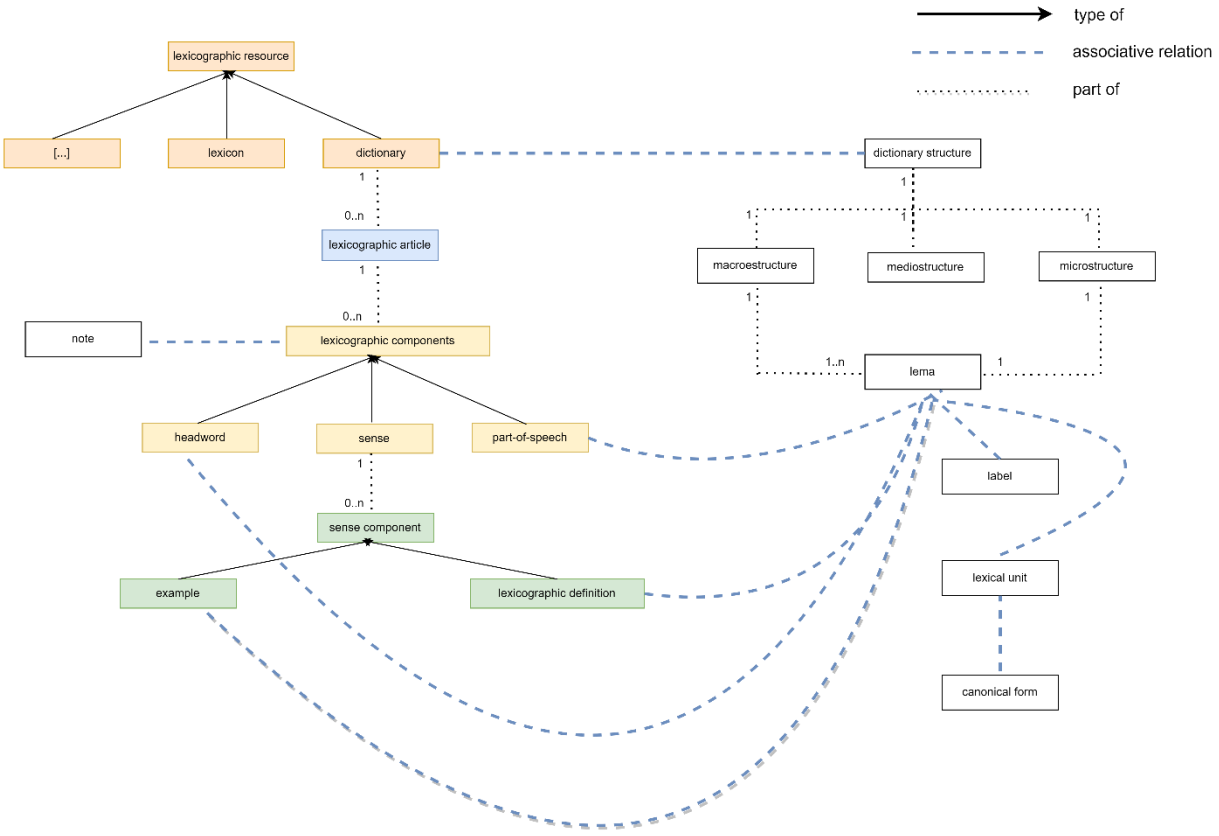


Figure 6: LexModel

References

Abromeit, F., Chiarcos, C., Fäth, C., & Ionov, M. (2016). Linking the tower of Babel: modelling a massive set of etymological dictionaries as RDF. In J. McCrae et al. (Eds.), Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources, Portoroz, Slovenia (pp. 11–19). Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-LDL2016_Proceedings.pdf.

Bański, P., Bowers, J., & Erjavec, T. (2017). TEI Lex-0 guidelines for the encoding of dictionary information on written and spoken forms. In Kosem, I., Tiberius, C., Jakubíček, M., Kallas, J., Krek, S., & Baisa, V. (Eds.), Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference (pp. 485–494). Brno: Lexical Computing CZ s.r.o.

Bohbot, H., Frontini, F., Luxardo, G., Khemakhem, M., & Romary, L. (2018). Presenting the Nénufar Project: A diachronic digital edition of the Petit Larousse Illustré. In GLOBALEX 2018 – Globalex workshop at LREC2018, May 2018, Miyazaki, Japan (pp. 1–6). Retrieved from https://hal.archives-ouvertes.fr/hal-01728328.

Bosque-Gil, J., Gracia, J., & Gómez-Pérez, A. (2016). Linked data in lexicography. Kernerman Dictionary News, 24:19–24. Retrieved from https://lexicala.com/wp-content/uploads/2021/03/kdn24_2016.pdf.

Bowers, J., Herold, A., Romary, L., Tasovac. T. (2021). TEI Lex-0 Etym – Towards terse recommendations for the encoding of etymological information. Preprint. Retrieved from https://halinria.fr/hal-03108781.

Budin, G., Majewski, S., & Mörth, K. (2012). Creating lexical resources in TEI P5. A schema for multi-purpose digital dictionaries. Journal of the Text Encoding Initiative [Online], 3. doi:10.4000/jtei.522.

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report. W3C Community Group Final Report. Retrieved from https://www.w3.org/2016/05/ontolex/.

Costa, R., Roche, C, and Salgado, A. (2022). Standards for Representing Lexicographic Data: An Overview. Version 1.0.0. DARIAH-Campus. [Training module]. https://elexis.humanistika.org/id/REhOykBU7pRute Costa, Christophe Roche and Ana Salgado (2022). Standards for Representing Lexicographic Data: An Overview. Version 1.0.0. DARIAH-Campus. [Training module]. https://elexis.humanistika.org/id/REhOykBU7pPs5zOAENdahPs5zOAENdah.

Costa, R., Salgado, A., Khan, A., Carvalho, S., Romary, L., Almeida, B., Ramos, M., Khemakhem, M., Silva, R., & Tasovac, T. (2021). MORDigital: the advent of a new lexicographical Portuguese project. In I. Kosem et al. (Eds.), Electronic lexicography in

the 21st century: post-editing lexicography. Proceedings of the eLex 2021 conference (pp. 312–324). Brno: Lexical Computing CZ. ISSN 2533-5626.

Declerck, T., McCrae, J., Navigli, R., Zaytseva, K., & Wissik, T. (2019). ELEXIS – European lexicographic infrastructure: Contributions to and from the linguistic linked open data. In Kernerman, I., & Simon, K (Eds.), Proceedings of the 2nd GLOBALEX Workshop. GLOBALEX (GLOBALEX-2018) Lexicography & WordNet located at 11th Language Resources and Evaluation Conference (LREC 2018), Miyazaki Japan (pp. 17–22). Paris: ELRA. Retrieved from https://www.dfki.de/fileadmin/user_upload/import/9709_elexis-european-lexicographic.pdf.

Ide, N. M., & Véronis, J. (1995). Text Encoding Initiative: Background and Contexts. Cambridge, MA: The MIT Press.

Khan, A., Romary, L., Salgado, A., Bowers, J., Khemakhem, M., & Tasovac, T. (2020). Modelling etymology in LMF/TEI: The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a use case. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), 11–16 May (pp. 3172–3180). France: Marseille.

ISO 1087. (2019). Terminology Work – Vocabulary – Part 1: Theory and Application. Geneva: International Organization for Standardization.

ISO 1951. (2007). Presentation/representation of entries in dictionaries – Requirements, recommendations and information. Geneva: International Organization for Standardization.

ISO 24156-1. (2014). Graphic notations for concept modelling in terminology work and its relationship with UML – Part 1: Guidelines for using UML notation in terminology work. Geneva: International Organization for Standardization.

ISO 24613-1. (2019). Language resource management – Lexical markup framework (LMF) – Part 1: Core model. Geneva: International Organization for Standardization.

ISO 24613-2. (2020). Language resource management – Lexical markup framework (LMF) – Part 2: Machine Readable Dictionary (MRD) model. Geneva: International Organization for Standardization.

ISO 24613-3. (2021). Language resource management – Lexical Markup Framework (LMF) – Part 3: Etymological Extension. Geneva: International Organization for Standardization.

ISO 24613-4. (2021). Language resource management – Lexical Markup Framework (LMF) – Part 4: TEI serialisation. Geneva: International Organization for Standardization.

ISO 24613-5. (2018). Language resource management – Lexical markup framework (LMF) – Part 5: Lexical base exchange (LBX) serialization. Geneva: International Organization for Standardization.

ISO 704. (2009). Terminology work – Principles and methods. Geneva: International Organization for Standardization.

Klimek, B. & Brümmer, M. (2015). Enhancing lexicography with semantic language databases. Kernerman Dictionary News, 23, 5–10. Retrieved from https://www.kdictionaries.com/kdn/kdn23_2015.pdf.

McCrae, J. P., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation, 46(6), 701–709. doi:10.1007/s10579-012-9182-3.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). TheOntoLex-Lemon Model: development and applications. In Proceedings of eLex 2017, pages 587–597.

McCrae, J. P., Tiberius, C., Khan, A. F., Kernerman, I., Declerck, T., Krek, S., Monachini, M., & Ahmadi, S. (2019). The ELEXIS interface for interoperable lexical resources. In Proceedings of the eLex 2019 conference. Biennial Conference on Electronic Lexicography (eLex-2019) Electronic lexicography in the 21st century. October 1–3 Sintra Portugal (pp. 642–659). Brno: Lexical Computing CZ, s.r.o. Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_37.pdf.

McGuinness, D.L. & van Harmelen, F. (2004). OWL Web Ontology Language Overview. W3C Recommendation.

Roche, C. (2015). Ontological definition. In Handbook of Terminology: Volume 1, pp. 128–152. Edited by Hendrik J. Kockaert and Frieda Steurs. John Benjamin Publishing Company, Amsterdam/Philadelphia, 2015.

Romary, L., & Tasovac, T. (2018). TEI Lex-0: A target format for TEI-Encoded dictionaries and lexical resources. In Proceedings of the 8th Conference of Japanese Association for Digital Humanities (pp. 274–275). Retrieved from https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf.

Salgado, A. (2021). Terminological Methods in Lexicography: Conceptualising, Organising and Encoding Terms in General Language Dictionaries. (Doctoral dissertation).

Salgado, A., Costa, R., Tasovac, T., & Simões, A. (2019). TEI Lex-0 In Action: Improving the encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem et al. (Eds.), Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference, 1-3 October 2019 (pp. 417–433). Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o.

Tasovac, T., Salgado, A., & Costa, R. (2020). Encoding polylexical units with TEI Lex-0: A case study. Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research, 8(2), 28–57. doi:10.4312/slo2.0.2020.2.28-57. e-ISSN 2335-2736.

Tasovac, T., Romary, L., Bański, P., Bowers, J., Does, J. de, Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Petrović, S., Salgado, A., e Witt, A. (2018). TEI Lex-0: A baseline encoding for lexicographic data. Version 0.8.5. DARIAH Working Group on Lexical Resources. Retrieved from https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Tiberius, C., Costa, R., Erjavec, T., Krek, S., McCrae, J., Roche, C., Tasovac, T. (2020). D1.2. Best practices for lexicography – intermediate report. 50 p. Report H2020-INFRAIA-2016-2017 Grant Agreement No. 731015 ELEXIS – European Lexicographic Infrastructure.