Originally published in: Kosem, Iztok/Zingano Kuhn, Tanara (eds.): Electronic lexicography in the 21st century (eLex 2019): Smart Lexicography. Book of abstracts. Sintra, Portugal, 1–3 October 2019. - Lexical Computing CZ s.r.o.: Brno, Czech Republic, 2019. Pp. 66-68

A Web of Loans: Multilingual Loanword Lexicography with Property Graphs

Peter Meyer, Mirjam Eppinger

Leibniz-Institut für Deutsche Sprache, Mannheim
E-mail: meyer@ids-mannheim.de, eppinger@ids-mannheim.de

Introduction

The Lehnwortportal Deutsch (2012 seqq.) serves as an integrated online information system on German lexical borrowings into other languages, synthesizing an increasing number of lexicographical dictionaries and providing basic cross-resource search options. The paper discusses the far-reaching revision of the system's conceptual, lexicographical and technological underpinnings currently under way, focussing on their relevance for multilingual loanword lexicography.

Data structures

Roughly in the spirit of Měchura (2016), the new Lehnwortportal system combines traditional XML-based digital lexicography with graph-based interlinking of information. It is special in relying on a property graph database, using the Turing-complete Apache Tinkerpop Gremlin query language (Rodriguez, 2015). The overall architecture is best suited for manually curated resources and is to be distinguished from a more NLP-oriented RDF-based Linked Data approach as discussed, for example, in (Gracia, Kernerman & Bosque-Gil, 2017). The graph features three main types of nodes (lexical units, their senses, entries in lexicographical resources). It expresses relations between lexical units such as borrowing, derivation, diachronic development, and variation; in addition, it relates lexical units to their word senses and to the entries in the portal's resources they appear in. Distinct but interconnected subgraphs cleanly separate the data found in the individual original sources from the manually edited, homogenized and interlinked portal data layer that is supposed to represent genuinely linguistic information (Meyer & Eppinger, 2018).

Accessing and editing the graph

Apart from a traditional entry/XML-based presentation, the revised Lehnwortportal provides intuitive and powerful real-time access to the entire graph via an innovative visual query builder (Meyer, 2018) that allows users to find arbitrarily complex graph constellations (n-tuples of words). Amongst other things, the query system offers Boolean operators to describe alternative or non-existing paths (e.g. borrowing histories)

as well as comparisons such as “is homonymous to” or “has same PoS as”. Users can easily navigate the curated graph by following edges between graph nodes visually or hypertextually. Authorized lexicographers may directly edit the subgraphs around search results.

Lexicographical process

The non-linear nature of graph editing and the requirement to keep the portal’s XML resources in sync with the graph creates the need for intertwined ‘graph-augmented’ lexicographical tools and practices. Amongst other things, bookkeeping of changes and additions to the data, including versioning, should itself be graph-driven, taking advantage of the graph query system and the possibility to add temporary editing attributes to search results. Changes in the underlying resources must percolate in a well-defined manner in the graph and lead to the automatic flagging of nodes and edges whose modifications might be lexicographically incorrect and must therefore be checked manually. Especially for the digitization/integration of new lexicographical data, the Lehnwortportal features a ‘graph-first’ input editor.

New resources

The revised Lehnwortportal, freely accessible from 2021, will host a number of new dictionaries of German loanwords in European languages, amongst others, English (Pfeffer & Cannon, 1994), French (Le Trésor de la Langue Française Informatisé) and Hungarian (Benkő & Büky, 1994). The paper includes a conceptual comparison to two related projects on Dutch (van der Sijs, 2015) and Italian (Heinz, 2017) loanwords in other languages.

Keywords: property graph; loanword lexicography; lexicographical process

References

Benkő, L. & Büky, B. (1993). *Etymologisches Wörterbuch des Ungarischen.* Budapest: Akad. Kiadó.

Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). Toward Linked Data-Native Dictionaries. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.* Brno: Lexical Computing, pp. 550-559. Available at: https://elex.link/elex2017/proceedings-download.

Heinz, M. (ed.) (2017). *Osservatorio degli italianismi nel mondo: punti di partenza e nuovi orizzonti. Atti dell’incontro OIM (Firenze, 20 giugno 2014).* Firenze: Accademia della Crusca.

Le Trésor de la Langue Française Informatisé. Accessed at: http://atilf.atilf.fr/

Lehnwortportal Deutsch. Ed. by Institut für Deutsche Sprache, Mannheim. Accessed at: http://lwp.ids-mannheim.de. (16 August 2019)

Měchura, M. (2016). Data structures in lexicography: from trees to graphs. In A. Horák, P. Rychlý & A. Rambousek (eds.) *Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016*. Brno: Tribun EU, pp. 97-104.

Meyer, P. (2019). Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein interaktiv-visueller Query Builder für Property-Graphen. In P. Sahle (ed.) *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*. Frankfurt am Main, pp. 312-314. Available at: https://zenodo.org/record/2600812.

Meyer, P. & Eppinger, M. (2018). fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts, 17-21 July, Ljubljana*. Ljubljana: Znanstvena založba, pp. 1017-1022.

Pfeffer, J. A. & Cannon, G. (1994). *German Loanwords in English. An Historical Dictionary*. Cambridge, New York, Melbourne: Cambridge University Press.

Rodriguez, M. A. (2015). The Gremlin Graph Traversal Machine and Language. In J. Cheney & Th. Neumann (eds.) *Proceedings of the 15th Symposium on Database Programming Languages (DBPL 2015)*. New York: The Association for Computing Machinery, pp. 1-10.

van der Sijs, N. (2015). *Uitleenwoordenbank, hosted by the the Meertens Instituut*. Accessed at: http://www.meertens.knaw.nl/uitleenwoordenbank. (16 August 2019)