Peter Meyer

LEHNWORTPORTAL DEUTSCH: A NEW ARCHITECTURE FOR RESOURCES ON LEXICAL BORROWINGS

Abstract This paper presents the *Lehnwortportal Deutsch*, a new, freely accessible publication platform for resources on German lexical borrowings in other languages, to be launched in the second half of 2022. The system will host digital-native sources as well as existing, digitized paper dictionaries on loanwords, initially for some 15 recipient languages. All resources remain accessible as individual standalone dictionaries; in addition, data on words (etyma, loanwords etc.) together with their senses and relations to each other is represented as a cross-resource network in a graph database, with careful distinction between information present in the original sources and the curated portal network data resulting from matching and merging information on, e. g., lexical units appearing in multiple dictionaries. Special tooling is available for manually creating graphs from dictionary entries during digitization and for editing and augmenting the graph database. The user interface allows users to browse individual dictionaries, navigate through the underlying graph and 'click together' complex queries on borrowing constellations in the graph in an intuitive way. The web application will be available as open source.

Keywords Multilingual lexicography; lexical borrowings; graph database

1. Introduction

This paper presents the *Lehnwortportal Deutsch*, a freely accessible publication platform for resources on German lexical borrowings in other languages. The system is curated and hosted at the Leibniz Institute for the German Language (IDS) and will go live in 2022. The new system is a complete redesign of a web portal online since 2012 (Lehnwortportal 2012 sqq.). It will host digital-native sources as well as digitized paper dictionaries. The approach of the portal is in a certain sense the inversion of the direction of observation of conventional loan dictionaries. It makes it possible to comparatively examine different types of language contact situations and socio-cultural developments (change of rule, migration, technological innovation, etc.) in the light of the dynamics of lexical borrowing processes into the languages concerned. For German, language contact has historically arisen in a number of very different settings, among them long-lasting contact at population borders (e. g., German – Slovenian, German – Polish), contact through emigration (e. g., German – Romanian, German – American English), and contact through socio-cultural influence and elite exchange (e. g. German – Japanese, German – British English, German – Tok Pisin) (cf. Meyer/ Engelberg 2011 for more discussion).

The platform will initially cover lexicographical data on English, French, Dutch, Swedish, Portuguese, Slovene, Czech, Slovak, Standard Polish, the Silesian and Cieszyn dialects of Polish, Hungarian, Turkish, and Tok Pisin. Dictionaries on German loanwords in the dialects of Polish and in the East Slavic languages Belarusian, Ukrainian, and Russian are currently being compiled specifically for the new platform (Meyer/Hentschel 2021; Meyer 2015, to appear). In the current project phase, there is a strong focus on European languages; the choice of resources is strongly constrained by the necessity to acquire publications rights. The selection of resources also reflects an effort to present as wide a variety of different loanword lexicographic approaches as possible. Most dictionaries record borrowings into the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

written or standard language. In addition, however, the new portal will also include resources with a focus (i) on a specific dialect of a language (Cieszyn dialect of Polish); (ii) on the entire dialect continuum of a language (Polish); (iii) on subjective usage frequencies vis-à-vis native/ standard vocabulary (for Silesian Polish); (iv) on borrowing history across multiple languages (Polish and the three East Slavic languages Ukrainian, Belarusian and Russian). Several contributions to the *Lehnwortportal Deutsch*, all related to Polish, have their origins in a long-standing cooperation of the IDS with the Institute of Slavic Studies at the Carl von Ossietzky University of Oldenburg.

2. Data Model

In an approach different from similar projects on Dutch (Uitleenwoordenbank 2015) and Italian (Osservatorio degli Italianismi nel Mondo), the new platform preserves its lexicographical sources in their original digital format and presentation.

In addition, the lexical units treated in the resources – loanwords, etyma, variants and derivatives thereof – and their relations to each other (such as borrowing, variation, derivation etc.), to their senses, and to their containing entries are modelled as a single large cross-resource network, technically a property graph (Rodriguez 2015).

The data model (Meyer/Eppinger 2019) consists of two interconnected graph ‘layers’ or subgraphs, one layer representing native dictionary data on a per-entry basis and the other one manually curated cross-resource information. If, for example, a lexical unit, such as a German etymon borrowed into multiple languages, appears in multiple resources, it should be represented as a single ‘merged’ graph node on the curated layer. In case the individual resources provide contradictory information on some property of the word, a lexicographically informed choice has to be made for the curated node, but through its connections to the various nodes on the native dictionary layer the conflicting source data is preserved as well.

Granular semantic versioning for documents (cf. SEMVER) will be enforced on various data levels (the portal, its dictionaries, their entries, and the graph nodes). Updates to the contents of the *Lehnwortportal* will periodically culminate in new releases; previous releases of the lexicographical data, including individual dictionary entries and graph data on lexical items (‘infoboxes’, see below), will remain accessible to the user through a unique and permanent URL. Editorial work is always done on a dedicated ‘upcoming’ release that is accessible and modifiable only for authorized users after login.

3. Backend and tooling

While the previous system used a relational database to emulate a graph (Meyer/Engelberg 2011), the new platform leverages the query capabilities and scalability of a native graph database with multi-model capabilities. The server backend of the application, with basic role-based administration tools, will be published as Open Source. As a basic principle, all data are stored in the database either as nodes in the graph or as generic documents, in order to make consistent and atomic updates as easy as possible. This includes configuration information on all relevant levels as well as lexicographical data such as XML documents, scan images and binary resources for multimedia content. Besides generic functionality for data and version management the application includes the following components.

- There is a dedicated tool for easy manual creation of subgraphs, corresponding to separate entries of a lexicographical resource – typically a paper dictionary – to be digitized, and also for assigning the entry to correctly aligned parts of image scans in the case of paper dictionaries (Meyer/Eppinger 2018).
- An annotation tool assists in linking word senses to one or even multiple 'keywords' in a pre-computed large set of multilingual word embeddings such as ConceptNet Numberbatch (Speer/Chin/Havasi 2018), for the purpose of easy 'Google-style' onomasiological search through a large number (>10000) of available German keywords (Meyer/Tu 2021).
- The platform provides visual graph editing functionality that takes advantages of the possibility to set properties, such as the editing status of words or their properties, for strictly internal purposes. Generic graph database queries can algorithmically set internal 'flags' on graph nodes that must be manually checked. This makes it possible to use the graph capabilities for lexicographical bookkeeping purposes.

4. Online presentation

Lexicographical content is presented in a responsive Single Page Application that provides access to the individual dictionaries in a uniform, consistent manner, even though entry presentation and microstructure may vary considerably between dictionaries. Entries can be selected through traditional headword lists and optional filters. In addition to the dictionary-specific entry presentation, the lexical units treated in an entry, which are represented as nodes on the curated graph layer (see above), can be displayed in 'infoboxes' that detail all information available on the word in question. Infoboxes provide grammatical information and word senses, but also list related words, possibly in other resources, sorted according to their (borrowing/derivation/variation/…) relationships. Clicking on such a related word opens its infobox, such that users may navigate through the graph by simply following links. Infoboxes can also provide links to external resources, such as Google n-gram time series (Michel et al. 2011) or other online dictionaries, and enable authorized users to add comments on specific words.

End users may perform queries on the graph to create custom lists of lexical units, similar to a faceted search familiar from online retail catalogues. In the most elementary case, formulating a query just means adding some filter for words in a certain language, or in a specific dictionary, or with a certain part of speech etc. For their queries, users may use both globally available properties (such as part of speech, meaning) and resource-specific criteria. This approach is naturally extended to cover multi-word constellations specified through properties of the words themselves and their – possibly indirect or mediated – relations to each other. For advanced purposes, a visual graph query builder can be used to define arbitrarily complex configurations in the graph (Meyer 2019). This includes configurations that must be formalized via Boolean operators on paths in the graph. As a simple example, one might want to search for loanwords borrowed into language X that have not also been borrowed into language Y.[1]

[1] Queries with possibly nested Boolean operations on paths in a graph are not supported by most graph database query languages, at least not in full generality. The *Lehnwortportal* uses Gremlin, a low-level open-source graph traversal language (Rodriguez 2015) which is Turing-complete and thus fulfills this requirement.

All search options will be available as components of a single generic low-threshold search facility. Query results can be displayed in multiple ways: as simple word lists augmented with a configurable set of additional word-related information; in tabular form (words with their properties) with filtering and sorting options; or in aggregated form as basic statistical visualizations, such as bar charts for part of speech distribution, timelines for the distribution of dates of first attestation (where available) or symbolic maps for geographical distribution of languages. For each word, its 'infobox' (see above) is accessible with a click, as well as an interactive graph visualization for the relevant subgraph representing the search result in question.

5. Conclusion and Future Directions

The new *Lehnwortportal Deutsch* explores the possibilities (Engelberg/Meyer 2015) and limitations (Meyer 2017) of a graph-based approach to editing, curating, presenting and querying heterogeneous, multilingual, highly interlinked lexicographical data. Using a multi-model graph database not only for storing cross-resource data on words and their relationships, but also for all kinds of lexicographical source data such as XML and images, and for configuration and administration information streamlines the editorial process and tooling significantly and helps to ensure data consistency. The most time-consuming part of the portal creation process is the manual digitization of existing dictionaries, which requires a translation of etymological information, often not sufficiently explicit, into the rigorous language of property graphs. Due to inconsistencies in the source dictionaries themselves and to the wealth of different etymological constellations, it is difficult or impossible to formulate hard rules for the translation process. In some cases, the scholarly character of exposition makes it hard even for experts to extract the relevant bits of information from a lengthy etymological discussion.

Besides the integration of new dictionaries and, possibly, manually collected data that fills gaps in previously digitized resources, there are plans for including new kinds of properties of lexical data, such as phonemic or morphological analyses, in the future.

The reasons for working with a property graph database instead of, say, an RDF triplestore are mainly practical. To this day, RDF is rarely used as the native format for dictionary creation, mainly for the lack of adequate tooling; property graphs are a more convenient and easily human-readable way of organizing network-like information. The graph data format is ideally suited for data integration into the Linguistic Linked Open Data (LLOD) infrastructure (cf. Declerck et al. 2020), as serializing property graphs to RDF is a straightforward procedure.

References

Declerck, T./McCrae, J./Hartung, M./Gracia, J./Chiarcos, Ch./Montiel, E./Cimiano, Ph./Revenko, A./Sauri, R./Lee, D./Racioppa, S./Nasir, J./Orlikowski, M./Lanau-Coronas, M./Fäth, Ch./Rico, M./Elahi, M.F./Khvalchik, M./Gonzalez, M./Cooney, K. (2020): Recent developments for the linguistic linked open data infrastructure. In: 12th International Language Resources and Evaluation Conference (LREC 2020). https://doi.org/10.5281/zenodo.3736949 (last access: 23-03-2022).

Engelberg, S./Meyer, P. (2015): Das Lehnwortportal Deutsch als kontaktlinguistisches Forschungsinstrument. In: Kelih, E./Fuchsbauer, J./Newerkla, S. M. (eds.): Lehnwörter im Slawischen: Empirische und crosslinguistische Perspektiven. Frankfurt a. M./Berlin/Bern/Bruxelles/New York/Oxford/Vienna, pp. 149–170.

Lehnwortportal Deutsch (2012 sqq.): lwp.ids-mannheim.de (last access: 23-03-2022).

Meyer, P. (2015): Aligning word senses and more: tools for creating interlinked resources in historical loanword lexicography. In: Kallas, J./Kosem, I./Krek, S. (eds.): Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 Conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton, pp. 198–210.

Meyer, P. (2017): The limits of lexicographical abstraction. Some strengths and problems of the data architecture in the Lehnwortportal Deutsch. In: Heinz, M. (ed.): Osservatorio degli italianismi nel mondo. Punti di partenza e nuovi orizzonti. Atti dell'incontro OIM Firenze, Villa Medicea di Castello 20 giugno 2014. Florence, pp. 55–76.

Meyer, P. (2019): Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein Query Builder für lexikografische Property-Graphen. In: Sahle, P. (ed.): Digital Humanities: multimedial & multimodal. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2019), Frankfurt am Main, Mainz, 25.3.2019 – 29.3.2019. Konferenzabstracts. Frankfurt a.M., pp. 312–314. https://doi.org/10.5281/zenodo.2600812 (last access: 23-03-2022).

Meyer, P. (to appear): "Lehnwortportal Deutsch", czyli "Portal zapożyczeń z języka niemieckiego" jako cyfrowe źródło informacji o kontaktach języka polskiego w dziedzinie leksyki [the Lehnwortportal Deutsch, or Portal of German Loanwords in Other Languages, as a digital information source on Polish lexical language contact]. In: Tom pokonferencyjny, VII Światowy Kongres Polonistów, Wrocław [Proceedings of the VII World Congress of Polonists, Wrocław].

Meyer, P./Engelberg, S. (2011): Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen. In: Hedeland, H./Schmidt, Th./Wörner, K. (eds.): Proceedings of GSCL Conference 2011 Hamburg. (= Arbeiten zur Mehrsprachigkeit, Serie B 96). Hamburg, pp. 169–174.

Meyer, P./Eppinger, M. (2018): fLexiCoGraph: creating and managing curated graph-based lexicographical data. In: Čibej, J./Gorjanc, V./Kosem, I./Krek, S. (eds.): Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts, 17–21 July, Ljubljana. Ljubljana, pp. 1017–1022.

Meyer, P./Eppinger, M. (2019): A web of loans: multilingual loanword lexicography with property graphs. In: Kosem, I./Zingano Kuhn, T. (eds.): Electronic lexicography in the 21st century (eLex 2019): Smart Lexicography. Book of abstracts. Sintra, Portugal, 1–3 October 2019. Brno, pp. 66–68.

Meyer, P./Hentschel, G. (2021): Charting a landscape of loans. An e-lexicographical project on German lexical borrowings in Polish dialects. In: Gavriilidou, Z./Mitits, L./Kiosses, S. (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. II, Komotini, Greece. Komotini, pp. 615–621.

Meyer, P./Tu, N.D.T. (2021): A word embedding approach to onomasiological search in multilingual loanword lexicography. In: Kosem, I./Cukr, M./Jakubíček, M./Kallas, J./Krek, S./Tiberius, C. (eds.): Electronic lexicography in the 21st century: post-editing lexicography. Proceedings of the eLex 2021 Conference. 5–7 July 2021 (virtual). Brno, pp. 78–91.

Michel, J.-B./Shen, Y.K./Aiden, A.P./Veres, A./Gray, M.K./Google Books Team/Pickett, J./Hoiberg, D./Clancy, D./Norvig, P./Orwant, J./Pinker, S./Nowak, M.A./Aiden, E.L. (2011): Quantitative analysis of culture using millions of digitized books. In: Science 331, pp. 176–182.

Osservatorio degli Italianismi nel Mondo. http://www.italianismi.org/ (last access: 23-03-2022).

Rodriguez, M. A. (2015): The Gremlin Graph Traversal Machine and Language. In: Cheney, J./Neumann, T. (eds.): Proceedings of the 15th Symposium on Database Programming Languages (DBPL 2015). New York, pp. 1–10.

SEMVER: Semantic Versioning for Documents 1.2.1. https://github.com/nils-tekampe/semverdoc/blob/master/semverdoc.md. (last access: 23-03-2022).

Speer, R./ Chin, J./Havasi, C. (2018): ConceptNet 5.5: An open multilingual graph of general knowledge. arXiv:1612.03975v2 [cs.CL], https://doi.org/10.48550/arXiv.1612.03975 (last access: 23-03-2022).

Uitleenwoordenbank (2015): – Sijs, N. van der (2015): Uitleenwoordenbank, hosted by the Instituut voor de Nederlandse Taal. uitleenwoordenbank.ivdnt.org (last access: 23-03-2022).

Contact information

Peter Meyer
Leibniz-Institut für Deutsche Sprache
meyer@ids-mannheim.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/
Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society.
Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.