Originally published in: Kosem, Iztok/Cukr, Michal/Jakubíček, Miloš/Kallas, Jelena/Krek, Simon/Tiberius, Carole (Eds.): Electronic lexicography in the 21st century: post-editing lexicography. Proceedings of the eLex 2021 conference. 5–7 July 2021, virtual. – Brno: Lexical Computing CZ s.r.o., 2021. Pp. 78-91.

A Word Embedding Approach to Onomasiological Search in Multilingual Loanword Lexicography

Peter Meyer[1], Ngoc Duyen Tanja Tu[1]

[1] Leibniz-Institut für Deutsche Sprache, R5, 6-13 68161 Mannheim Germany
E-mail: meyer@ids-mannheim.de, tu@ids-mannheim.de

Abstract

In this paper we present an experimental semantic search function, based on word embeddings, for an integrated online information system on German lexical borrowings into other languages, the *Lehnwortportal Deutsch* (LWPD). The LWPD synthesizes an increasing number of lexicographical resources and provides basic cross-resource search options. Onomasiological access to the lexical units of the portal is a highly desirable feature for many research questions, such as the likelihood of borrowing lexical units with a given meaning (Haspelmath & Tadmor, 2009; Zeller, 2015). The search technology is based on multilingual pre-trained word embeddings, and individual word senses in the portal are associated with word vectors. Users may select one or more among a very large number of search terms, and the database returns lexical items with word sense vectors similar to these terms. We give a preliminary assessment of the feasibility, usability and efficacy of our approach, in particular in comparison to search options based on semantic domains or fields.

Keywords: onomasiological search; word embeddings; multilingual lexicography; lexical borrowings

1. Introduction

The Lehnwortportal Deutsch (LWPD) is an online platform developed at the Leibniz-Institut für Deutsche Sprache and comprising lexicographical resources on German loanwords in other languages. The LWPD in its entirety realises the concept of a 'reverse loan dictionary' that does not focus on the target languages of the borrowing processes, but on the source language. Besides offering a traditional, lemma-based access to the individual dictionaries, the system provides sophisticated portal-wide cross-resource options to search for lexical units (German etyma, corresponding loanwords, variants and derivatives thereof, etc.).

At present, however, onomasiological access is restricted to simple substring-based searches on the word sense definitions for words as provided in the individual dictionaries. Consequently, a genuine semantic search in the LWPD would be more suitable for research questions like "Which languages have a conspicuously high proportion of German loanwords in certain thematic areas, such as *food* and *drinks*?"

In a project funded by the Fritz Thyssen Stiftung the LWPD is currently being substantially revised on both the backend and the user interface levels (Meyer &

Eppinger, 2019). The new edition will go online in early 2022, featuring a number of newly added resources on German borrowings in English, Dutch, French, Portuguese, Hungarian, Czech and Slovak. The new system will offer a much more powerful and simplified way to search the underlying graph database (Meyer, 2014), which represents the portal data as a network of partially cross-resource relationships between lexical units, through an innovative 'query builder' interface (Meyer, 2019). The semantic search function discussed in this paper will be an integral part of the query builder.

Conceptually, the approach presented below differs from hand-crafted semantic domain taxonomies that are used as search features in similar projects (e.g. van der Sijs, 2015; Osservatorio degli Italianismi nel Mondo) and come with many well-known problems:

(a) Semantic domain definitions are inherently vague and cannot be exhaustive, i.e. there is not a (perfectly) suitable domain for every word sense. This usually leads to senses without domain assignment or, equivalently, to the introduction of a semantically unspecified default 'miscellaneous' domain. Assignment of a word sense to multiple domains is frequently possible due to overlap, but is usually not wanted and must be avoided by arbitrary assignment decisions. If domain schemas are explicitly designed for multiple assignments, then this considerably complicates both the manual annotation process and the burden on the part of the user who has to experiment with combinations of (typically rather broad) domains.

(b) An introspection-based manual annotation procedure will inevitably lead to a complex lexicographical practice of domain assignments, especially if maximal inter-annotator agreement is demanded. This actually requires a considerable amount of *reverse engineering* of that (typically opaque) practice on the part of the user, and will prove difficult for word senses that do not fit easily into one of the domains, implying the annotator assigns them according to subjective intuition or some internal conventions.

(c) It is challenging to find a reasonable middle ground between ease of use and sufficient granularity. If the taxonomy is too coarse, the user might get too many search results, which makes the search inefficient. If, on the other hand, the taxonomy is too fine-grained, the number of categories to choose from becomes impractical and confusing, in particular for casual use.

(d) The domain taxonomy is essentially static. If certain domains turn out to yield unsatisfactory (e.g. counterintuitive) results, there is nothing the user can do apart from trying to get further relevant search results by randomly trying other domains. For lexicographers, any revision of the 'boundary' of a domain may turn out to be a time-consuming process as it involves a possibly large number of reassignments.

Our experimental approach, presented in section 2, is an attempt at addressing the

problems mentioned above. Section 3 discusses the problem of evaluating this approach with regard to its usability and performance as well as the quality of the search results. In section 4, we briefly summarise the pros and cons of our approach in comparison to domain-based searches.

2. Approach

2.1 Basic idea

In the revised LWPD, lexical items (etyma, loans, derivatives, and so on … figuring in the included dictionaries) can be searched for using any number of search criteria in arbitrary Boolean combinations. Basically, the new semantic search function will allow the user to describe the desired 'range' of meanings by entering words that are, in an intuitive sense, similar in meaning or topic. The user actually selects words from a very large given list of frequently used German words (henceforth: 'search keys') and takes advantage of autosuggest functionality during input. This speeds up typing and gives instantaneous feedback on the availability of search keys. Multiple search keys can be combined with each other to describe different aspects of a semantic 'field'. The query returns words with at least one word sense sufficiently close in semantics to the meanings of all search keys provided.

The list of search keys is meant to be of roughly the same order of magnitude as the active vocabulary of a native German speaker. So far, we have experimented with the 10,000 most frequent verbs, nouns and adjectives from DeReWo. DeReWo is a word frequency list based on DeReKo, the world's largest collection of German-language corpora. Note that the list of search keys available to the user can be altered, even radically, at any time, as will become clear in what follows.

2.2 Technical implementation

The technical implementation of our approach is based on word embeddings (Mikolov et al., 2013), a technique to represent the distributional properties of words in large corpora mathematically through vectors, i.e. lists of numbers. A simple measure, the cosine similarity of two vectors, is supposed to represent the semantic similarity of the respective words (Speer et al., 2018). Thus the semantic similarity between the search key and an LWPD word sense can be calculated by computing the cosine similarity between the vector representations of the two objects. The greater the cosine similarity, the more semantically similar the two words are. The maximum cosine similarity is 1.0, the minimum is -1.0. The semantic search function picks out word senses that have a sufficiently high cosine similarity (i.e., close to 1.0) to the search keys input by the user.

In our project, we use the ConceptNet (CN) NumberBatch pre-computed word

embeddings (Speer et al., 2018; we use version 19.08) to map each LWPD lexical unit word sense and each search key to a vector. Note that we could not train custom word embeddings ourselves since we do not have access to the corpus data underlying many of the portal's lexicographical resources. The CN embeddings are trained on multilingual data as well as otherwise known semantic relationships between words. Vectors for all included words of the more than 70 languages present in CN are aligned in one vector space, i.e. similarities can be measured across languages – which is evidently a basic precondition for their use in an LWPD search. As we will see soon, the dataset of embeddings can easily be replaced at a later time, if other pre-computed embeddings turn out to yield better search results.

The basic parts of the database architecture for the semantic search are shown in Figure 1.

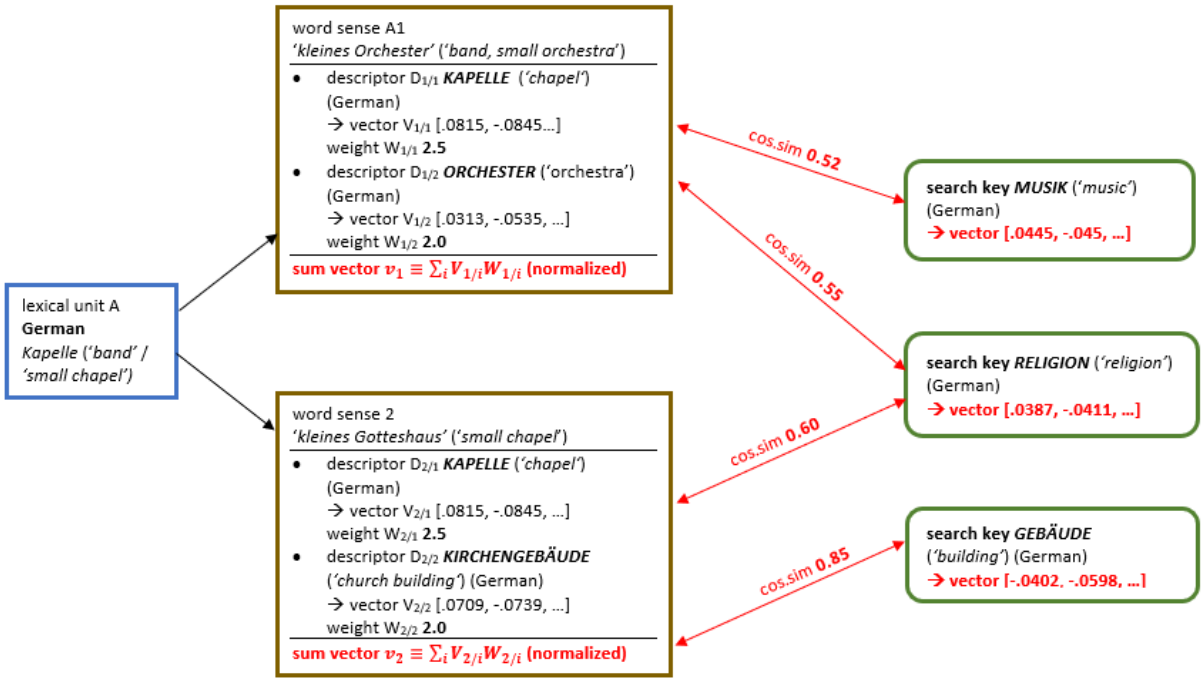

Figure 1: Basic database architecture of assigning embeddings to word senses and search keys.

This architecture is now explained in more detail.

(1) In LWPD, all lexical units are represented as nodes (vertices) in a property graph database. A lexical unit may appear in multiple dictionaries/entries (not shown in Figure 1); this occurs frequently with German etyma.

(2) All word senses of a lexical unit as found in the resources are represented as separate sense nodes in the graph. There can be considerable overlap between sense definitions if the lexical unit appears in multiple sources. No attempt at unifying these sense definitions is made in the LWPD.

(3) Using an in-house web application, student annotators assign to each word sense in the LWPD at least one word from CN, henceforth called a *descriptor* of the sense. The descriptors are supposed to have meanings that are closely related to the word sense in question. For these assignments, the full range of words covered by CN is available, with a vocabulary size of almost 600,000 items available for German alone. In most cases, a default descriptor is provided in advance; in the most elementary case this is simply the word the word sense is related to. For manual editing, the annotators have a number of tools and rules at their disposal, on which see below. Assigning multiple descriptors helps to overcome the notorious difficulties of word embeddings, in particular the fact that embeddings are not context-sensitive and do not differentiate in cases of polysemy and homonymy. For example, the etymon *Reif* appears in the present LWPD exclusively in the sense of 'hoop, bracelet'; just assigning the CN word *reif* to this sense would obscure the fact that there is a homonymous *Reif* meaning 'hoarfrost' and an adjective *reif* 'ripe, mature' – the latter since CN words are case-insensitive. So a second descriptor like German *ring* 'ring' can help to disambiguate. If multiple descriptors are used, they have to be *labelled* by the annotators according to their function. Labels are selected from a predefined list and include 'disambiguating word with similar meaning', 'hypernym', 'cohyponym' and others. For example, the CN words *bräme* ('trimming'), *verbrämung* ('trimming') and *pelzbesatz* ('fur trimming') might be assigned to the Polish word *bramik* ('fur trimming'). The latter CN word would get the label 'synonym', the first two CN words the label 'hypernym'.

(4) Each descriptor label is mapped onto a number representing the *weight* of the descriptor for the word sense it is assigned to. For example, hypernyms might get mapped to the integer 2 and synonyms to the number 2.5 (if a word sense has only one descriptor, weights play no role; formally, the weight of a solitary descriptor is always 1). This allows us to test (and change between) different mapping schemes in order to find the one that gives optimal results.

(5) The weighted and normalised sum of the vectors belonging to the CN descriptors yield the vector representation of the word sense. Thus, each word sense node in the LWPD graph has one such vector as a property.

(6) The search keys available to the users are selected as explained above, e.g. from a frequency list of lemmatised German words with relevant part of speech. They must be words in CN; but in practice this is not a serious restriction due to size of the CN data. Though it would seem natural not to restrict the available choices at all and use the entire German CN vocabulary, this would result in a disturbing amount of noise presented to the user. Each search key is represented as a node in the graph which has its CN vector as a property.

(7) The cosine similarity between all word sense vectors and all search key vectors

is computed; if it is above a certain threshold, an edge (i.e. a relation) between the word sense and the search key is stored in the graph and assigned the cosine similarity as a property. Consequently, no edge is stored between the word sense and the search key if their cosine similarity is only slightly above 0**.** The threshold can be defined arbitrarily but should exclude very low similarities in order to reduce noise in the search results; ultimately it is a matter of practical experience.

The annotators follow a complex, tool-guided procedure for assigning descriptors and labels in a meaningful and consistent way. Note that the notion of inter-annotator agreement is ill-defined in this context since the number of plausible alternative assignments is, in general, simply too high. The following remarks give a brief sketch of a still evolving practice.

(a) Default assignments 1: If an LWPD word is contained in CN, the word itself is automatically assigned to all of its word senses as its descriptor. For example, the Slovene word *bager* ('excavator') is contained in CN, so the assigned CN word is *bager*. If the LWPD word has more than one word sense, all its senses are marked for later manual revision, which means they are prioritised for a manual check because it is very likely that further differentiation among the senses is necessary. To give an example, the Hebrew word *Zup* has the two senses 'Suppe' ('soup') and 'Abschmecken einer Flüssigkeit' ('seasoning a liquid'). The first sense could be covered by the German CN word *Suppe* ('soup') corresponding to the etymon of *Zup*, the second one by the CN word *abschmecken* ('(to) season').

(b) Default assignments 2: If an LWPD word w is not included in CN, but there is an LWPD word w^* with an etymological or variational relationship to it that is included, then this CN word is taken as the default descriptor for the word senses of w (see (a) above for an example). These assignments are marked for manual review later. Information on the relationship between words is available in the LWPD graph database. For example, the Slovene loanword *ravbati* ('(to) rob') is not included in CN, but its German etymon *rauben* ('(to) rob') is, so *rauben* becomes the default descriptor for the senses of *ravbati*.

(c) Flagging of highly polysemous CN descriptors: The in-house tool warns annotators of polysemous descriptors, suggesting the use of additional descriptors for disambiguation purposes. It is not a trivial task to automatise the detection of polysemy. Typical lexicographical resources such as Wiktionary or WordNet-type databases exhibit a level of sense differentiation that is too granular for our purposes. Among the strategies that we are trying out to detect problematic cases of polysemy in German CN words are the following: (i) GermaNet (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010) partitions its synsets into different 'semantic fields'. If the synsets containing a certain CN

word are distributed among multiple semantic fields, then we assume significant polysemy. (ii) Consulting the lemmatisation of a reliable reference dictionary of German such as the DWDS, if the CN word corresponds to multiple headwords, we assume significant polysemy. The identification of significantly polysemous words from other languages is an open issue.

(d) Manual editing: Where default assignments are either not possible or introspectively misleading, appropriate descriptors have to be selected in a 'manual' fashion by searching for CN words that have a close semantic relationship to the LWPD word (e.g. hypernyms, synonyms, etc.), using resources such as OpenThesaurus, DWDS, and Wortschatz Universität Leipzig.

2.3 Performing queries

As explained above, semantic queries for words in the upcoming LWPD are specified by one or more search keys. An autocomplete function makes it easier to find and enter the search keys.

A typical user query may look like this: If you are interested in finding out whether German terms for certain types of dishes have been borrowed in the languages available in the LWPD's dictionary, you can use specific search keys to do so. In a domain-based semantic search, you would first have to make sure that a suitable domain exists. In our semantic search system, you could just use the search keys *Speise* ('dish') and *flüssig* ('liquid') if you want to get terms for liquid dishes present in the LWPD. As a search result you will obtain, among other things, *Suppe* ('soup') and *Mus* ('pulp'). If you are interested in sweet dishes, then you just have to enter *Speise* ('dish') and *süß* ('sweet') as search keys and you obtain among others *Nachtisch* ('dessert'), *Süßigkeit* ('candy') and *Zimtstern* ('star-shaped cinnamon cookie'). Thus, a user can search for very specific word fields without consulting any *a priori* taxonomy.

Technically, the semantic search is part of a traversal of the graph database. The database will search for word sense nodes whose cosine similarity to all of the search key nodes provided by the user is greater than a certain threshold. The search result list contains the LWPD words connected to these word sense nodes. The user may alter the threshold in the query to influence the size of the result set and obtain results that are more or less 'strict'.

A very similar approach has already been successfully used for search engine optimisation (Castro Fernandez et al., 2018; Kuzi et al, 2016; Fernandez et al., 2008) but not for semantic searches of lexicographic resources.

3. Evaluation

3.1 Usability and performance

The quality of a semantic search can be measured in terms of two properties: 1) Usability and 2) performance. (Elbedweihy et al., 2012)

(1) Usability: In our onomasiological search, search queries are entered using natural language search keys, so no query language needs to be learned. It also allows anyone to easily execute semantic search queries without having to read a manual beforehand. In addition to this, due to the autosuggesting input facility, the user does not have to invest much time in finding out which search keys are available at all and in formulating his search queries. In contrast, with a domain-based search, one must first become familiar with the taxonomy before starting a search. Furthermore, the searches are highly flexible. Thus, users can add or alter a search key if they want to filter the results of the previous search or found that the previous search was incomplete.

(2) Performance: The cosine similarities between the LWPD word senses and the search keys are all precomputed and stored in the graph database, if the cosine similarity is above a certain threshold. Since both the cosine similarities and the search keys stored in the graph database are indexed, a traversal from a search key to 'matching' LWPD words is possible in (approximately) constant time, and therefore very fast.

3.2 Quality of the search results

The quality of the search results of many semantic searches is evaluated by comparing the results of different search engines for the same query (e.g. Tümer et al., 2009; Uma Devi & Meera Gandhi, 2015). In our case, however, this is not possible because the data of lexicographical resources with a semantic search function differ from each other, which means that they are trivially providing different search results for the same query.

Moreover, the notion of recall of the search results is ill-defined in the case of the system presented here. The recall is calculated as the quotient of the relevant search results and that of *all* relevant items from the LWPD, i.e. those lexical units from the LWPD that *should* appear in the search results. However, the relevant search results would have to be determined by a human annotator, which has several disadvantages: (a) there are no fixed criteria for deciding whether a lexical item is 'really' a relevant search result, so subjective decisions are necessary; (b) an exhaustive search for relevant search results would be too time-consuming even for a small fraction of search keys.

The precision of the search results seems to be somewhat less problematic and could be tackled in a similar way as in Chauhan et al. (2013) and Mohamed and Shokry (2020). The precision is calculated as the quotient of the relevant search results and the number of all search results. Thus, it indicates the proportion of relevant search results in relation to all search results – it is not necessary to determine *all* possibly relevant items in the LWPD. In practice, however, it is still almost impossible to decide whether a result offered by the system should be considered relevant, e.g. if you select the search key *Speise* ('dish'), is *Koch* ('cook') relevant? What about *Service* ('(coffee) set')? Operationalising the evaluation of search result quality beyond taking samples from user studies is clearly an avenue for future research.

Unfortunately, a thorough evaluation of LWPD's onomasiological search will have to wait until at least a considerable subset of our data is available. We hope to complete the annotation of word senses for all German etyma by the end of 2021.

To get a first impression of the quality of the search results, we conducted a small study on the German etyma that are represented in the LWPD in its current incarnation, simulating possible search queries by looking for suitable words in the lexicographical sense definitions of these etyma. Of the 3,709 'meta-etyma' that serve as headwords in the Dictionary of German Etyma in the present database of the LWPD, 2,074 appear as CN words and also figure as lexical units in at least one GermaNet synset (we used GermaNet 14.0). For each such etymon E, we collected its word sense definitions as given in the LWPD dictionaries. All words in these definitions were POS-tagged and lemmatised with a standalone version[1] of the GATE DictLemmatizer plugin. For 1,668 etyma, at least one lemmatised word W was found that (i) belongs to the NN, ADJA or VV* POS-classes most relevant for searches and (ii) appears both in CN and in at least one GermaNet synset. For each such word W we determined the pair of one synset containing E and one synset containing W that has maximum semantic similarity $S_{E,W}$ according to the information-content-based measure by Lin (1998), assuming that the semantics of words W in a sense definition for a word E bears significant similarity to a word sense of E. The resulting 4,676 pairs turned out to be, in hindsight, a surprisingly noise-free collection of pairs of clearly semantically related terms such that the words W appearing in the definitions for the respective E did indeed very often appear to be good candidates for search keys relevant to E.

We then calculated, for each E-W pair, the CN-based cosine similarity between E and W and compared it to the $S_{E,W}$ measure introduced above. The results are shown in Figure 2. The more similar a word W in the definition of an etymon E is according to GermaNet, the higher, on average, is the cosine similarity between these two words. For highly GermaNet-related words, the average cosine similarity goes up to a

[1] The software is available at http://staffwww.dcs.shef.ac.uk/people/A.Aker/activityNLPProjects.html .

remarkable 0.65. It must be emphasised that these numbers constitute at best anecdotal evidence of the power of our approach to semantic search, but given the fundamentally different ways in which Lin's measure on GermaNet synsets and cosine similarity of word embeddings treat semantic similarity, they nevertheless indicate a basic and non-trivial consistency of search result quality with our theoretical expectations.

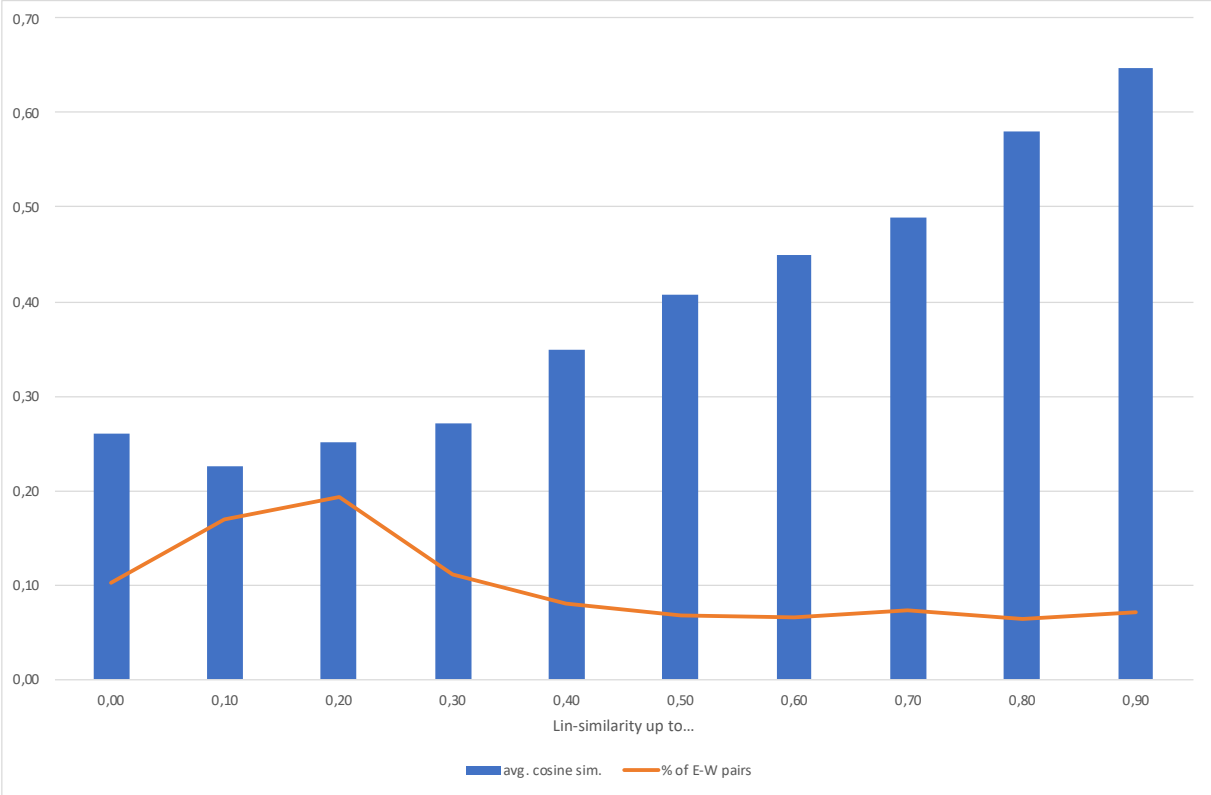

Figure 2: Average cosine similarity (blue bars) between German etyma E in the LWPD and words W in their definitions as a function of the Lin-measure based similarity of the corresponding maximally semantically similar GermaNet synsets (x-axis). The leftmost bar represents a maximum Lin-similarity between 0.0 and 0.1, and so on. The orange line indicates the percentage of E-W pairs falling in the respective class; so for example the eighth column reads "6.5% of all E-W pairs [orange line] have a Lin-similarity between 0.8 and 0.9 [x-axis position] of their respective synsets; the average cosine similarity of E and W in this class is 0.58 [blue bar].".

4. Conclusion

The experimental approach to onomasiological access in a multilingual lexicographical resource outlined in this paper is still in an early stage of implementation. It offers possible solutions to many of the issues of traditional 'domain-based' search strategies, sketched in section 1. Taking up the points listed there, we can wrap up our discussion with the following observations.

(a) Lexicographical annotators gain enormous flexibility in characterising word senses through a huge number of descriptor words. The downside to this is the

curious fact that, as noted above, annotator agreement is not a useful validation criterion anymore; in addition, annotators cannot assess the implications of their descriptor assignment choices for future users. It is, however, possible to give the annotators some feedback on the 'effect' their assignments have by showing them which other lexical units in the LWPD the assigned descriptors are semantically similar to and would be retrieved using the assigned descriptors in a query.

(b) Instead of having to reconstruct a lexicographical practice of domain assignments, the user is offered a much more open, even playful access to semantic search. Guided by autocomplete functionality and without prior familiarisation with a system of domains, users can experiment with any combinations of search keys to delimit and change (narrow down or open up) the scope of their queries. Thus, this kind of semantic search fits very well into the concept of the LWPD, since it is a lexical resource aimed at scientists as well as interested laypeople.

(c) The fundamental problem of having to decide on a more or less fixed set or taxonomy/hierarchy of semantic domains in advance of the whole annotation process simply disappears.

(d) As said above, it takes a lot of effort to change the taxonomy in a domain-based search or just redefine the 'boundaries' of a given domain. In contrast, the word embedding approach is highly dynamic. (i) The set of search keys can be altered in any conceivable way any time, including additional languages (as long as the keys are included in CN, which is very likely, because the CN embeddings are trained on a very big database). (ii) The scheme of mapping descriptor labels onto weights can be adjusted as needed. (iii) The pretrained set of multilingual embeddings can be exchanged for another one. In this case, only word senses with descriptors absent from the new embeddings must be annotated anew. It is not to expected that this concerns a sizeable fraction of the word senses. (iv) Of course, assignments for individual word sense can be revised any time. In all cases, all it takes for the changes to take effect is a recomputation of the vectors and cosine similarities in the database.

In the end, the most desirable state of affairs would most certainly that of offering users a combination of different semantic search options. Finding out which option is the best for which usage scenario remains a topic for further research.

5. References

Castro Fernandez, R., Mansour, E., Qahtan, A. & Elmagarmid, A. K. (2018). Seeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery. In IEEE (eds.) *Proceedings of the 34th International Conference on Data*

Engineering, ICDE 2018. Paris, pp. 989–1000. Available at: https://ieeexplore.ieee.org/document/8509314/.

Chauhan, R., Goudar, R., Sharma, R. & Chauhan, A. (2013). Domain ontology based semantic search for efficient information retrieval through automatic query expansion. In R. Kher, N. Gondaliya, M. Bhesaniya, L. Ladid & M. Atiquzzaman (eds.) *Proceedings of the International Conference on Intelligent Systems and Signal Processing, ISSP 2013*. Gujarat, pp. 397–402. Available at: http://ieeexplore.ieee.org/document/6526942/.

ConceptNet NumberBatch. Accessed at: https://github.com/commonsense/conceptnet-numberbatch. (06 April 2021)

DeReKo: *Das deutsche Referenzkorpus* Accessed at: https://www1.ids-mannheim.de/kl/projekte/korpora/. (06 April 2021)

DeReWo: *Die Deutsche Referenzkorpus Wortliste*. Accessed at: https://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html. (06 April 2021)

DWDS: *Digitales Wörterbuch der Deutschen Sprache*. Accessed at: http://www.dwds.de. (06 April 2021)

Elbedweihy, K., Wrigley, S. N., Ciravegna, F., Reinhard, D. & Bernstein, A. (2012). Evaluating semantic search systems to identify future directions of research. In R. García-Castro, L. Nixon & S. Wrigley (eds.) *Proceedings of the Second international Workshop on Evaluation of Semantic Technologies*. Heraklion, Greece, pp. 25-36. Available at: https://www.zora.uzh.ch/id/eprint/63315/.

Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E. & Castells, P. (2008). Semantic Search Meets the Web. In IEEE (eds.) *Proceedings of the 2008 International Conference on Semantic Computing*. Santa Monica, USA, pp. 253–260. Available at: http://ieeexplore.ieee.org/document/4597199/.

Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German.. In: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid. Available at: https://www.aclweb.org/anthology/W97-0800.pdf.

Haspelmath, M. & Tadmor, U. (2009). The Loanword Typology project and the World Loanword Database. In: M. Haspelmath & U. Tadmor (eds.): *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: De Gruyter, pp. 1–34.

Henrich, V. & Hinrichs, E. (2010). GernEdiT - The GermaNet Editing Tool. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.) *Proceedings of the Seventh Conference on International Language Resources and Evaluation*. Malta, pp. 2228–2235. Available at: https://www.aclweb.org/anthology/L10-1180/.

Kuzi, S., Shtok, A. & Kurland, O. (2016). Query Expansion Using Word Embeddings. In Association for Computing Machinery, New York, NY, United States (eds.) *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. Indianapolis Indiana USA, pp. 1929–1932. Available at: https://dl.acm.org/doi/10.1145/2983323.2983876.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *Proc. of Conf. on Machine Learning*, pp. 296–304.

LWPD: Lehnwortportal Deutsch. Leibniz-Institut für Deutsche Sprache, Mannheim. Accessed at: http://lwp.ids-mannheim.de/. (06 April 2021)

Meyer, P. (2014): Graph-Based Representation of Borrowing Chains in a Web Portal for Loanword Dictionaries. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the 16th EURALEX International Congress*. Bolzano, pp. 1135–1144. Available at: http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX%202014_gesamt.pdf.

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.

Meyer, P. (2019). Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein Query Builder für lexikografische Property-Graphen. In: P. Sahle (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.: Zenodo, pp. 312-314.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR abs*/1301.3781.

Mohamed, E. H. & Shokry, E. M. (2020). QSST: A Quranic Semantic Search Tool based on word embedding. In *Journal of King Saud University - Computer and Information Sciences*.

OpenThesaurus. Accessed at: https://www.openthesaurus.de/. (06 April 2021)

Osservatorio degli Italianismi nel Mondo. Accessed at: http://www.italianismi.org. (06 April 2021)

Speer, R., Chin, J. & Havasi, C. (2018). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In: arXiv:1612.03975 [cs].

Tümer, D., Shah, M. A. & Bitirim, Y. (2009). An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia. In IEEE Computer Society (eds.) *2009 Fourth International Conference on Internet Monitoring and Protection*. Venice/Mestre, Italy, pp. 51–55. Available at: http://ieeexplore.ieee.org/document/5076348/.

Uma Devi, M. & Meera Gandhi, G. (2015). Wordnet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports Domain. In *Journal of Computer Science*, 11(2), pp. 361–371.

van der Sijs, N. (2015). Uitleenwoordenbank, uitleenwoordenbank.ivdnt.org, hosted by the Instituut voor de Nederlandse Taal. Accessed at: http://uitleenwoordenbank.ivdnt.org/. (06 April 2021)

Wortschatz Universität Leipzig. Abteilung *Automatische Sprachverarbeitung* am

Institut für Informatik der Universität Leipzig. Accessed at: https://corpora.uni-leipzig.de/ (06 April 2021)

Zeller, J. P. (2015). The semantic fields of German loanwords in Polish. In *Studies in Polish Linguistics*, pp. 153–174.