Published in: Heinz, Matthias (Hrsg.) (2017): Osservatorio degli italianismi nel mondo. Punti di partenza e nuovi orizzonti. Atti dell'incontro OIM Firenze, Villa Medicea di Castello 20 giugno 2014. Firenze: Accademia della Crusca. S. 55–76.

PETER MEYER

THE LIMITS OF LEXICOGRAPHICAL ABSTRACTION. SOME STRENGTHS AND PROBLEMS OF THE DATA ARCHITECTURE IN THE *LEHNWORTPORTAL DEUTSCH*

1. INTRODUCTION

The idea of a database-driven 'inverted' loanword dictionary (ENGELBERG 2010) that documents borrowings from a given donor language into many recipient languages goes back to at least the 1970s (cf. KARAULOV 1979). Apart from technical considerations, the major obstacle to such an enterprise for any major language is the large number of recipient languages to be considered. So far, the only comprehensive lexicographical resource of this kind available is for Dutch (VAN DER SIJS 2010). The underlying database for this monumental work has been compiled in a huge effort by the author; the data are available through the print publication and as a PDF file of the book; a powerful online search interface, the *uitleenwoordenbank* (www.meertens.knaw.nl/uitleenwoordenbank/), is available online (VAN DER SIJS 2015). Apart from the *Osservatorio degli italianismi nel mondo* that is currently in development (OIM 2014, HEINZ-GÄRTIG 2014), the only other project of a somewhat similar intended scope seems to be the *Lehnwortportal Deutsch* (http://lwp.ids-mannheim.de; cf. MEYER-ENGELBERG 2011) that aims to provide free access to a growing number of dictionaries of German loanwords in other languages. The web portal is maintained and hosted by the Institute for the German language (IDS) and is online in a preliminary version since late 2012. In a first stage of this project, dictionaries on German loans in three Slavic languages have been integrated; an ongoing research project develops and integrates dictionaries on Polish-mediated German loans in the East Slavic languages Russian, Belarusian and Ukrainian (Meyer 2014a). An increasing number of loanword resources on Germanisms in other language families is currently being acquired or prepared for inclusion.

The overall structure and basic conception of the *Lehnwortportal Deutsch* (henceforth, LWP) has been described in several publications (cf. Meyer 2013a, Meyer 2014b) and will not be reviewed here in depth. Apart from traditional lemma-based access to the individual resources through a uniform web user interface, the web portal features an 'inverted' dictionary whose headwords (henceforth, *metalemmata*) are German etyma; the entries of this dictionary mainly consist of pointers to entries in the various loanword dictionaries of the portal, thus allowing users to trace the way of a specific German word into the different recipient languages. The inverted dictionary is not simply the product of an automated rearrangement of the digital or digitized data presented in the loanword dictionaries. The German etyma as they actually appear in the loanword dictionary entries – often in some dialectal and/or diachronic form – are mapped in a manual, lexicographically informed process onto 'normalized' metalemmata, typically a diasystematically corresponding form in contemporary standard New High German, using a dedicated in-house software tool.

The major innovation of the LWP is the data organization metaphor underlying the various cross-resource search options (cf. Meyer 2013b) offered for advanced purposes. Both the portal-specific metalemmata and the actual object-language words (etyma, loanwords including morphophonological variants, derivatives etc. listed in the entries – not only headwords and not necessarily lexemes) as they are presented or 'recorded' in the entries of the individual loanword dictionaries are represented as the vertices of a directed acyclic graph (DAG) whose edges represent various relations between words, such as “etymon x is mapped to metalemma y”, “loanword x is borrowed from etymon y”, “x is a derivative of y”. For the sake of brevity, the words corresponding to graph vertices will henceforth be referred to as ‘recorded words’. Since diasystematically equatable etyma from different resources and/or entries are mapped onto the same metalemma and hence form part of the same connected subgraph, the graph provides a cross-resource abstraction layer put on top of the idiosyncratic data organisation found in the different loanword dictionaries. The graph may either be searched directly using a portal-specific, declarative, German-like query language, or by filling in expandable HTML forms whose input is translated into graph queries behind the scenes. The DAG-based data organisation is ideally suited for modelling arbitrarily long chains of borrowing processes where, e.g., a German etymon is first borrowed into Polish, the Polish loanword is later borrowed into Ukrainian, from there into Russian and so forth (Meyer 2014a). The DAG representation is also used for the dynamic creation of the portal’s metalemma entries in the LWP web application (cf. fig. 1, p. 73). See Spohr 2012 for more references on graph-based endeavours in lexicography.

2. A SKETCH OF THE PORTAL ARCHITECTURE

The LWP data model has been designed with a number of rather specific requirements in mind. A major objective of the portal was to make a growing range of existing dictionaries of German loanwords in other languages available under a single roof and with a uniform access structure. Apart from dedicated third-party funded projects, the IDS Mannheim has no resources for compiling such dictionaries on its own, so it is only to be expected that the legacy dictionaries to be included in the portal come in a variety of original data formats (XML, plaintext, image-digitized, …), with very different entry structure and with incompatible formats of diasystemic, grammatical and other information. It is important to note that the resources are taken 'as is', although there are plans to add a versioning system to the portal that will allow designated lexicographers to make corrections and additions.

From the outset, the LWP was meant to provide more than a simple compilation or mashup of lexicographical resources. Experts should be able to perform detailed, powerful and fast searches in a steadily growing lexicographical portal database that makes systematic use of *ex post facto* interlinking of legacy resources on the level of individual words, not just entries or documents as a whole. Updating and augmenting entries should be possible and be handled separately from changes in portal-specific cross-referencing information.

To fulfil these requirements, two different representations of the lexicographical data are used side by side. Each entry in one of the loanword dictionaries is a standalone XML document; each loanword dictionary retains its own XML schema. No attempt has been made to create a kind of super-schema that purports to fit all microstructural needs of the individual dictionaries. Even the merely four dictionaries included so far in the LWP display remarkable idiosyncrasies that would have been impossible to foresee in the design process for such a super-schema. Trying to find an all-encompassing XML schema for all dictionaries would be a hard task and force constant readjustment. Just two examples of particularities: (a) In the Standard Polish loanword dictionary of the LWP (de Vincenz-Hentschel 2010), derivatives of a given Polish lemma are often annotated with those word senses of the base lexeme that are semantically 'compatible' with the meaning of the derivative. (b) The dictionary of German loanwords in the Cieszyn dialect of Polish (Menzel-Hentschel 2005) systematically specifies etymologically related *lexical parallels* to the lemma in other languages of the same linguistic area (Polish, Czech, German; both standard languages and dialects), irrespective of the exact and often unclear borrowing history (i.e., which language borrowed from which).

Schema-specific XSL transformations are used to produce

HTML views of the entries that are at least close to the original entry presentations, so that, apart from minor design changes, the dictionaries may be consulted in their original appearance.

The XML documents are integrated into the portal database without significant alterations; however, all XML elements corresponding to recorded words are provided with an additional XML attribute, viz. a globally unique ID (UUID), for cross-referencing purposes. The list of metalemmata and different sorts of cross-referencing added for the LWP are coded in separate XML documents that make heavy use of the UUIDs just mentioned. These XML documents are not edited manually, but through a complex in-house GUI application that takes care of keeping the link consistency intact.

In principle, the cross-linked XML documents as such could be used for portal-wide search operations. But searching XML documents with varying schemas and recursively traversing ever longer chains of cross-references between XML documents would soon become too slow as the portal grows larger.

As *relations between recorded words* are the greatest common factor both of different microstructures in loanword dictionaries and of cross-referencing XML documents, a natural solution to this and other technical problems is to represent the portal's lexicographical content as a network, a *directed acyclic graph*, of relations between the recorded words. Such a graph can be stored and accessed in a scalable and efficient manner in a dedicated NoSQL graph database such as Neo4j (cf. Webber-Eifrem-Robinson 2013). Currently, the LWP emulates a graph database through a conventional relational database. The graph is derived in a fully automated, dictionary-specific process from the underlying XML resources mentioned above: Structural configurations between XML elements – expressible by relative XPath expressions – are translated into graph edges; word-specific information (diasystemic, grammatical, and other data on recorded words) is converted to node attributes in a unified data format. The graph as such is not an independent resource and must be recreated or adjusted every time an underlying XML resource changes; thus, no complex and error-prone synchronisation between XML documents and the graph structure is needed.

3. strengths

With its inter-dictionary cross-references and the uniform layout and access structure of the portal and its component dictionaries, the LWP is a standard example of a dictionary net in the sense of Engelberg-Müller Spitzer 2013. Moreover, its graph-based data representation adds a layer of domain-specific *uniform access semantics* for all portal resources, a layer that is created through a process of abstracting from the idiosyncrasies of individual dictionary macro- and microstructures and reframing the lexicographical data in a common graph-based metalanguage that enables users to search for information without having to consider the structural diversity of the various lexicographical sources.

- In comparison to XML-based solutions, the data architecture is highly *scalable* and can handle complex cross-resource searches for a growing number of component dictionaries without considerable performance penalty.
- The architecture is *flexible* and accommodates to a wide range of entry microstructures in legacy dictionaries instead of imposing a single lexicographical data model.
- The duality of arbitrarily many interlinked XML documents on the one side and a graph structure automatically extracted from these documents on the other allows the portal-makers to *cleanly separate* dictionary-specific content from portal-added information (by coding different content in different XML documents) without sacrificing a uniform cross-resource view on the data where the boundaries between the different sources of information are virtually invisible.

- The portal-wide part of the data model is *granular*: Since the graph nodes represent individual recorded words, not whole entries, cross-references can and must be formulated at the word level.
- The graph is *extendable*; arbitrarily many additional attributes can be assigned to both nodes (words) and edges (relations). These attributes may contain user-generated plaintext comments; versioning information in the case of updates to the original entries; information on the degree of plausibility of, say, a borrowing relationship as indicated by the source; semantic or ontological classification as a complement to a traditional lexicographical meaning definition; etc. There are no restrictions on how this additional information is encoded and represented in, e.g., separate XML source documents, as long as the graph can automatically be constructed using all available documents. This paves the way for innovative bookkeeping techniques as far as versioning and user-generated additions to the resources are concerned, since all versions of the lexicographical information could easily be co-present in one and the same graph, as long as all nodes and edges are systematically provided with versioning information.
- The graph structure remains semantically *consistent* even when different lexicographical sources give contradictory information on a particular word, say, specify different gender for a German etymon noun. The simple reason is that words recorded in different resources are, as a matter of principle, mapped onto different vertices of the graph. Any statements on cross-resource identities are added content that might be coded as an identity-expressing edge by the portal staff. German etyma are a special case; their diasystematic identity is coded by mapping them onto the same metalemma. Metalemmata serve as simple pointers to dictionary-specific etyma and do not possess attributes

of their own, so no problems with conflicting information arise. Search requests may easily return inconsistent data from different resources, so that an informed evaluation and decision will be left to the user.

- The DAG has a *simple* conceptual structure that can be mapped easily onto description standards such as TEI and onto semantic web standards such as RDF (every graph edge is indeed basically isomorphic to an RDF triple) to ensure interoperability with other applications and the possibility of cross-linking with other data sets; cf. Wandl Vogt-Declerck 2013 for an example of publishing lexicographical resources in the Linked Open Data framework.
- The graph structure is well suited for *visualisation* and is already being used for illustrative purposes in the LWP, as connected subgraphs can be displayed interactively for each German metalemma. However, the simple visual metaphor of graphs also lends itself to exploitation for novel and intuitive kinds of searches, similar to, e.g., the interactive construction of graphs representing search queries in treebanks as described in Voormann-Lezius 2002.

4. problems

Somewhat paradoxically, most of the conceptual, lexicographical, and practical difficulties that the LWP faces ensue from features that have been described above as advantages of the basic conception of the portal.

A major source of trouble that will quickly become obvious even to occasional users of the portal is the *heterogeneity* of the resources included, not only as far as reliability, quality and volume of the provided information is concerned.

- Loanword dictionaries vary in the *criteria of inclusion/exclusion* of etyma, loan translations etc. The two Polish-related dictionaries in the LWP (Menzel-Hentschel 2005, de Vincenz-Hentschel 2010) do not include loanwords whose German etyma are of Latin or Greek origin; many yet-to-be-included resources beg to differ on this point. If a user queries the portal for languages that have borrowed such etyma, she or he will probably get a thoroughly misleading picture that largely depends on the particular decisions of the various dictionary editors. There is, of course, no proper general solution for this in a context where legacy resources have to be taken "as is" and might, in addition, differ considerable as to their age and intended comprehensiveness.

- Both systematic and occasional *lacunae* in the lexicographical data pose another major problem. One of the currently three component dictionaries of the LWP (Menzel-Hentschel 2005) does not indicate part of speech or any other grammatical information on etyma and loans. This makes the design of advanced searches difficult: At present, the mere choice of including part of speech or gender as a criterion in a portal-wide search implies that no results from Menzel-Hentschel 2005 will be displayed. Two possible workarounds come to mind, neither of which is entirely satisfactory: The user might add a disjunctive criterion ("POS is either noun *or unspecified*"; such an option is not implemented yet) or such a disjunctive criterion is added implicitly by the search engine. Adopting the latter solution would only be sensible if search results with unspecified POS are somehow marked as not completely conforming to the criteria specified.

  An example for an occasional type of lacuna would be missing information on the meaning of recipient-language derivatives from loanwords in cases where well-known productive derivational patterns are involved. From the point of view of print lexicography, this is

a reasonable way of saving space and avoiding redundancy, but it is a problem for anyone using meaning paraphrases as a search criterion in the portal. A possible solution for this particular problem could be to simply copy the meaning(s) associated with the base of the derivative and flag this via an additional attribute of the derivative. Such artefacts can quickly make the search system unwieldy for end users, however.
Sometimes textual condensation routinely used in print lexicography produces lacunae in a digitized version, for example in cases where a previous word sense definition is referred to later in the entry using words such as ‘idem’ or its German equivalent ‘dss’. Usually, the exact reference cannot be established algorithmically, such that the only way out is to fill in the full definition manually.

- As briefly explained above, the German *metalemmata* constitute the backbone for cross-resource links in the graph. Operationalizing the process of individuating and assigning metalemmata as well as defining their interrelations has turned out to be very difficult, mainly because it is often unclear whether it is appropriate to ‘identify’ two German etyma given in different resources. It is clear that the main criteria for this must rely on linguistic reconstruction in historical linguistics, but often details remain unclear or seem to be irrelevant to the majority of portal users. To give an example, the Standard Polish dictionary in the LWP (de Vincenz-Hentschel 2010) gives the following New High German etyma for Polish *sztuciec* ‘case for storage’ (http://lwp.ids-mannheim.de/art/wdlp/lemma/sztuciec): *Stutz*, *Stutzen*, *Stütze* ‘stump, vessel’. Of these three forms, only *Stutzen* remains in active use in Standard German (though there is an unrelated homonym *Stütze* ‘support’), and in addition it is difficult to pinpoint the precise morphological and historical-phonological relationships between these variants. Should all three forms or just *Stutzen* be included as metalemmata? Should dialectal variants that come to be integrated into the standard language in specialised meanings be represented as separate metalemmata, or, more generally, how do we determine the maximal admissible 'diasystematic distance' between variant forms mapped onto the same metalemma? Comparable problems are of course faced by all projects that try to integrate heterogeneous lexicographical resources into a dictionary net through some concept of meta-index, cf. Herold-Geyken-Lemnitzer 2012.

Moulding entry data into the graph structure of the LWP is a complex process of interpretation of both explicit and implicit lexicographical information. Once again, users must be aware that the advanced search options operate on a formalized, artificially homogenized view of the original data.

- Most obvious is the case of *dictionaries without a systematic microstructure* whose entries are written in a rather narrative or even discursive style, as in the Slovene dictionary of the LWP (Striedter-Temps 1963) that has been included into the LWP through image-digitization of the entries. For the construction of the graph, additional manual extraction of lexicographical information into a fixed XML structure was necessary. This process was difficult and error-prone as many details, especially regarding the exact delimitation of different entry sections, determination of the scope of qualifiers, and the interpretation of implicit references within the entry required careful human inspection and often enough remained unclear even to trained philologists. It goes without saying that such issues are intrinsic to any attempt at retrodigitizing unstructured text, but they are much more apparent in a setting where each piece of information must be assigned to precisely one word (vertex) or word-word-relation (edge).

- As stated above, the *set of attributes and their values is uniform* across the portal, which entails the mapping of all dictionary-specific grammatical and diasystematic classifications and categories onto portal-wide 'ontologies'. Thus, although denominations for dialects and older forms of German such as 'New High German' (NHG), associated with slightly differing, yet not always explicitly specified, temporal and areal boundaries by different authors, must be equated for the portal's purposes. Users searching for loans from NHG etyma will therefore not get words borrowed in a specific period in the history of High German, but words borrowed from *whatever was classified as a 'New High German' etymon* in the various source dictionaries, including dialectal specifications such as 'Bavarian-Austrian' that, strictly speaking, leave the temporal component unspecified.
- Relations between words are often *left unspecified* in the source dictionaries when they are supposed to be obvious to the intended audience. A good example is the entry on the Polish noun *kształt* 'shape etc.' in de Vincenz-Hentschel 2010 (http://lwp.ids-mannheim.de/art/wdlp/lemma/kształt) that notes two variant forms of the word, viz. *kształt* and *kstałt*. There is a productive adjective in *-owny* derived from the loanword, also listed in two variant forms, *kształtowny* and *kstałtowny*. The precise relations, namely that *kształtowny* is derived from *kształt* and *kstałtowny* from *kstałt*, are left unspecified, i.e. they are not recoverable from the XML markup of the entry. This enforces a rather contrived algorithmic translation from XML to DAG, e.g. by encoding both adjectives as derived from both nouns, or (this is the current solution in the LWP) encoding both adjectives as derived from the 'canonical' lemma variant. In every conceivable solution to this

problem, the interpretation of the relation (edge type) "x is a derivative of y" becomes blurred and opaque. – A similar case where information was left unspecified because it was unknown to the dictionary compiler can be found in the Slovene dictionary (Striedter-Temps 1963): Often enough, it is not clear whether a word is a Slovene derivative of a loanword borrowed from a German etymon *e* or a Slovene loanword borrowed from a German derivative of *e*, cf. Slovene *drukati/druk/drukar* 'to print/(a) print/printer' vs. German *drucken/Druck/Drucker* (http://lwp.ids-mannheim.de/art/st/lemma/drukati). The two different borrowing pathways – which may even occur simultaneously, with Slovene speakers alternatively borrowing the German derivative or forming a derivative on their own according to a productive Slovene pattern – necessitate a rather crude operationalization. In this specific case, a borrowed derivative has been assumed whenever the German derivative is mentioned in the entry.

As a general principle, the direction of graph edges in the LWP is intended to convey a uniform semantics: In a relation $x \rightarrow y$ the target node word y may to be thought of as being 'dependent on' and usually temporally subsequent to the source word node x. The following relation types have been implemented so far in the LWP data model: metalemma → German etymon; German etymon → loanword in target language; lexeme → derivative lexeme; lexeme → compound lexeme; word → variant form (orthographical, phonological variant); lexeme → lexical parallel (see above for explanation). The relation type 'metalemma → German etymon' is special as it does not literally indicate a temporal sequence, even though it might be read as typically implying a certain linguistic provenance that is, however, not made explicit in the metalemma, because the latter is, as stated above, a commonly used Standard NHG lexeme, wherever possible. The main advantage of this approach, apart from technical considerations, is that this semantics transitively extends to arbitrarily long chains ('paths') of connected unidirectional edges as they arise, inter alia, with borrowing chains. In some cases, however, this semantics is compromised when the edge direction is either unclear (there is not enough evidence nor any suitable operationalization to decide on the 'correct' direction) or irrelevant (semantically, there is no directedness). Such a situation will lead to search results that are misleading at best because the graph structure seems to be erroneous:

- The *lexical parallels* adduced in Menzel-Hentschel 2005 leave the actual borrowing pathway(s) unspecified, as stated above, mostly because in many cases they cannot be reconstructed with certainty on the basis of the evidence available. Hypotheses on plausible borrowing histories are left to a free text commentary. To take an example, Cieszyn Polish *(h)ajbisz* has German *Eibisch* 'marshmallow (the plant, Althaea officinalis)' as its etymon and Czech *ibišek, ajbiš* as lexical parallels. The DAG subgraph for this entry (http://lwp.ids-mannheim.de/art/meta/lemma/Eibisch) has edges such as "German etymon *Eibisch* → Cieszyn Polish loanword *ajbisz*" and "Cieszyn Polish loanword *ajbisz* → Czech parallel form *ajbiš*", suggesting a borrowing chain *Eibisch* → *ajbisz* → *ajbiš*. The free text commentary reveals that the Cieszyn Polish form has "obviously" been borrowed from Czech, so the chain should rather be *Eibisch* → *ajbiš* → *ajbisz*. This is a simple and clear case; for a really complicated example see http://lwp.ids-mannheim.de/art/wdlt/lemma/bachórz. Things become even more difficult where multiple borrowings, possibly along different pathways and concerning only certain derivatives or even word senses of derivatives, are concerned. In all of these cases, the simple graph metaphor cannot but fail on its semantic promise, all the more since the 'correct' relations cannot be reconstructed from the digital source because they are specified, if at all, in the commentary. The only clean way of dealing with this problem is to exclude lexical parallels from normal portal-wide search queries.

- The relation of being an *orthographical or phonological variant* is, generally speaking, a symmetric one and therefore clashes with the directedness requirement of the DAG. In a group of multiple variants, the only way of connecting each node to every other one in the group would be to introduce arbitrary edge directions that avoid cycles. For the time being, the LWP always uses an operationalization of electing one member of the group as the 'canonical' variant and introducing only variance-edges from this node to all other group members. These 'non-canonical' variants are not linked to any other nodes; any other edges linking the 'canonical' variant to other nodes outside the group must then be interpreted as potentially also or only 'applicable' to all or some of the 'non-canonical' variants. For example, in the case of Polish *kształt* vs. *kstałt* and *kształtowny* vs. *kstałtowny* discussed above, the only edges introduced are: *kształt* → variant *kstałt*; *kształt* → derivative *kształtowny*; *kształt* → derivative *kstałtowny*. That the last edge mentioned is really only 'applicable' to the 'non-canonical' variant *kstałt* remains unexpressed and would not be deducible from the source data anyway. One remedy in these cases would be to form a non-directed, strictly local 'hyperedge' that spans all variants and combines them to a sort of 'hypernode' that all other edges attach to. Such an approach is possible with modern graph databases.
- A further problem arises with the *directedness of morphological derivation*. Often enough, a decision on the 'correct' direction is notoriously difficult and theory-dependent, for example with conversion (zero derivation). This is particularly relevant for establishing the direc-

tion of derivational edges between metalemmata, as the source loanword dictionaries will usually define a direction for derivation relations within the recipient language.

Sometimes, the fundamental decision to take recorded words as graph nodes and, therefore, as basic building blocks of the data model, leads to a representation that is, surprisingly, not granular enough and makes it impossible to code certain relations adequately. The most important case is that of 'multiple borrowing' where "one and the same etymon" has been borrowed at different times or, more generally, in different diasystematically related forms, into a target language. This leads to (a) word senses of a loanword that are attributable to only one of these forms and (b) to derivative formations that also relate to only one of these forms. By way of example, the LWP's Polish dictionary de Vincenz-Hentschel 2010 features an entry on Standard Polish *waga* 'scales' where the NHG and the Middle High German form of the etymon are separately associated with some of the 24 (!) word senses of the loanword, albeit in a non-exhaustive and overlapping fashion. The same kind of association with word senses is also coded for the derivatives of *waga*. At least in principle, this can lead to a state where a certain loanword derivative is 'compatible' with only one or at least not with all given forms of the etymon. This is not the case with *waga*, where all derivatives are associated with at least one word sense that is itself connected to both etyma. In another case, however, the entry on Polish *lump* 'wretch' (http://lwp.ids-mannheim.de/art/wdlp/910) lists two related German etyma, viz. *Lump* 'wretch' and the compound noun *Lumpenzucker* 'bad sugar'; all derivatives only relate to a word sense that itself is associated only with *Lump*, not with *Lumpenzucker*. In this case, a path in the graph such as *Lumpenzucker* → borrowed as *lump* → derivative adjective *lumpowski* leads to the wrong conclusion, viz. that of *lumpowski* being indirectly related, amongst other things, to the etymon *Lumpenzucker*. Note that splitting loanwords into homonyms is not an option for de Vincenz-Hentschel 2010 since word senses, as we saw in the case of *waga*, can often be linked to multiple etyma.

5. conclusion

Interlinking a growing number of digital resources has become a major topic of research in the digital humanities. The present paper has demonstrated some of the conceptual, philological, and practical difficulties that arise even in the context of manually cross-referencing resources from a rather narrow content domain and providing a common access structure for them. As a matter of fact, the main strength of the LWP and its main problems come by the same name, viz. abstraction.

For most of the issues sketched in this paper, there is no ready-made solution. Three strategies come to mind that can help to live with what may be perceived as a thoroughly unsatisfactory situation. First, user-generated content and the possibility for experts to provide additions and corrections can help to overcome some of the graph's shortcomings, if only in the long run. Second, users must be made aware at every turn of the pitfalls of taking search results at face value. The graph-based abstraction used in the LWP is 'leaky' in a sense well-known to computer scientists: in many cases, thorough knowledge of the abstraction process and a careful inspection of the resources underlying the abstraction is necessary to interpret query results adequately. And third, the search engine should take a maximally generic approach unless otherwise specified – it is better to have too many results (and let the human user sort out the false positives) than to use restrictive search strategies that possibly sort out relevant hits. As far as the LWP is concerned, the strategy adopted for the advanced HTML form based search is roughly as follows: the user may specify search criteria for a German etymon and a loanword related to this etymon. Now the search engine does not search for matching etymon-loanword pairs that are connected through a path in the graph but instead looks for German metalemmata that are connected to both the etymon and the loanword – the simple rationale for this setup being that every node in the graph is linked to at least one metalemma. This approach neutralizes the problems with the semantics of graph relations discussed above: for example, not every variant of an etymon is connected to every variant of a loanword that is etymologically 'connected' to it; but all variants of the loanword and all variants of the etymon are connected to the same German metalemma, because otherwise there would be no etymological 'connection'. In many cases, this strategy yields too many hits, but this is preferable over a situation where careful consideration of the DAG-construction algorithm would be needed in order to formulate a maximally exhaustive search query.

Peter Meyer
Institut für Deutsche Sprache, Mannheim

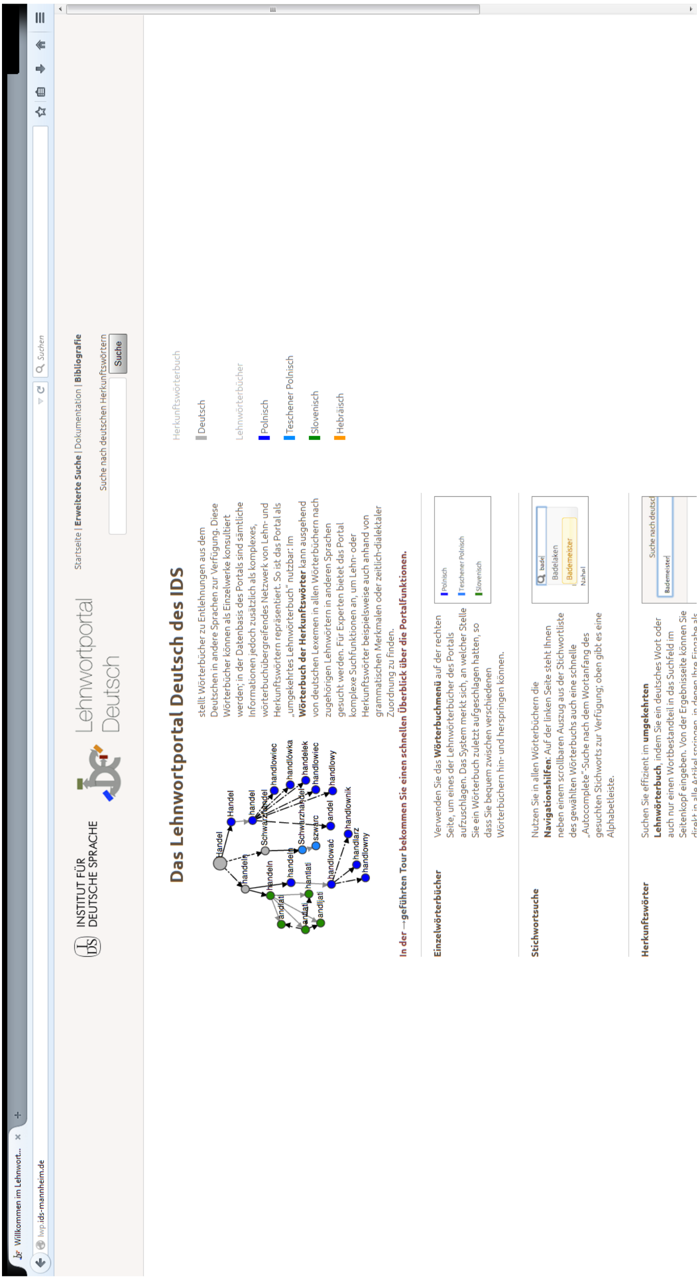

Fig. 1. Homepage of the *Lehnwortportal Deutsch* (http://lwp.ids-mannheim.de)

References

de Vincenz-Hentschel 2010 = Andrzej de Vincenz - Gerd Hentschel, *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts*, Oldenburg, BIS-Verlag ("Studia slavica Oldenburgensia", 20) (http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp [14/07/2015]).

Engelberg 2010 = Stefan Engelberg, *An inverted loanword dictionary of German loanwords in the languages of the South Pacific*, in A. Dykstra - T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress*, Leeuwarden, 6-10 July 2010, Leeuwarden, Fryske Akademy, pp. 639-47.

Engelberg-Müller Spitzer 2013 = Stefan Engelberg - Carolin Müller-Spitzer, *Dictionary Portals*, in R.H. Gouws *et al.* (eds.), *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie*, Supplementary Volume: *Recent Developments with Focus on Electronic and Computational Lexicography*, Berlin-Boston, de Gruyter, pp. 1023-35.

Heinz-Gärtig 2014 = Matthias Heinz - Anne-Kathrin Gärtig, *What a multilingual loanword dictionary can be used for: Searching the* Dizionario di italianismi in francese, inglese, tedesco *(DIFIT)*, in A. Abel - C. Vettori - N. Ralli (eds.), *Proceedings of the XVI EURALEX International Congress. The User in Focus*, 15-19 July 2014, Bolzano/Bozen, Bolzano, EURAC research, pp. 1099-107 (http://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_085_p_1099.pdf [14/07/2015]).

Herold-Geyken-Lemnitzer 2012 = Axel Herold - Alexander Geyken - Lothar Lemnitzer, *Integrating lexical resources through an aligned lemma list*, in Ch. Chiarcos - S. Nordhoff - S. Hellmann (eds.), *Linked data in linguistics*, Berlin/Heidelberg, Springer, pp. 35-44.

Karaulov 1979 = Jurij Nikolaevič Karaulov, *Obratnyj slovar' zaimstvovanij kak sposob isučenija lingvoėkologii*, in «Izvestija Akademii Nauk SSSR», Serija Literatury i Jazyka, 38/6, pp. 552-62.

Menzel-Hentschel 2005 = Thomas Menzel - Gerd Hentschel, *Wörterbuch der deutschen Lehnwörter im Teschener Dialekt des Polnischen*, 2., ergänzte und korrigierte elektronische Ausgabe (http://www.bkge.de/Publikationen/Online/Woerterbuecher/Deutsche_Lehnwoerter_im_Teschener_Dialekt [14/07/2015]).

Meyer 2013a = Peter Meyer, *Ein Internetportal für deutsche Lehnwörter in slavischen Sprachen. Zugriffsstrukturen und Datenrepräsentation*, in S. Kempgen *et al.* (eds.), *Deutsche Beiträge zum 15. Internationalen Slavistenkongress*, Minsk 2013, München, Sagner ("Die Welt der Slaven", Sammelbände, Band 50), pp. 233-42.

Meyer 2013b = Peter Meyer, *Advanced graph-based searches in an Internet dictionary portal*, in I. Kosem *et al.* (eds.), *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*, 17-19 October 2013, Tallinn, Estonia, Ljubljana-Tallinn, Trojina, Institute for Applied Slovene Studies - Eesti Keele Instituut, pp. 488-502 (http://eki.ee/elex2013/proceedings/eLex2013_34_Meyer.pdf [14/07/2015]).

Meyer 2014a = *Graph-based representation of borrowing chains in a web portal for loanword dictionaries*, in A. Abel - C. Vettori - N. Ralli (eds.), *Proceedings of the XVI EURALEX International Congress. The User in Focus*, 15-19 July 2014, Bolzano/Bozen, Bolzano, EURAC research, pp. 1135-45 (http://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_088_p_1135.pdf [14/07/2015]).

Meyer 2014b = Peter Meyer, *Von XML zum DAG. Der lexikographische Prozess bei der Erstellung eines graphenbasierten Wörterbuchportals*, in M.J. Domínguez Vázquez - F. Mollica - M. Nied Curcio (eds.), *Zweisprachige Lexikographie zwischen Translation und Didaktik*, Berlin/Boston, de Gruyter, pp. 303-21.

Meyer-Engelberg 2011 = Peter Meyer - Stefan Engelberg, *Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen*, in H. Hedeland - T. Schmidt - K. Wörner (eds.), *Multilingual Resources and Multilingual Applications* (= "Arbeiten zur Mehrsprachigkeit/Working Papers in Multilingualism", Folge B, Nr. 96), Hamburg, Universität Hamburg, pp. 169-74.

OIM = *Osservatorio degli italianismi nel mondo*, progetto dell'Accademia della Crusca, coord. Luca Serianni - Matthias Heinz, con la collaborazione di Marco Biffi - Domenico De Martino - Nicoletta Maraschio - Giovanni Salucci - Gesine Seymer - Harro Stammerjohann *et al.* [Firenze, Accademia della Crusca, 2014-]: www.italianismi.org [30/11/2015].

Spohr 2012 = Dennis Spohr, *Towards a Multifunctional Lexical Resource. Design and Implementation of a Graph-based Lexicon Model*, Berlin-Boston, de Gruyter, ("Lexicographica", Series Maior, 141).

Striedter-Temps 1963 = Hildegard Striedter-Temps, *Deutsche Lehnwörter im Slovenischen*, Wiesbaden, Harrassowitz.

van der Sijs 2010 = Nicoline van der Sijs, *Nederlandse woorden wereldwijd*, Den Haag, SDU Uitgever (http://www.meertens.knaw.nl/uitleenwoordenbank/resources/2010.pdf [1/8/2017]).

van der Sijs 2015 = N. van der Sijs, *Uitleenwoordenbank*: www.meertens.knaw.nl/uitleenwoordenbank/ (hosted by the Meertens Instituut) [1/8/2017].

Voormann-Lezius 2002 = Holger Voormann - Wolfgang Lezius, *TIGERin - Grafische Eingabe von Benutzeranfragen für ein Baumbank-Anfragewerkzeug*, in S. Busemann (ed.), *KONVENS 2002.* 6. Konferenz zur Verarbeitung natürlicher Sprache. Proceedings. DFKI Document D-02-01, Saarbrükken, Deutsches Forschungszentrum für Künstliche Intelligenz (http://konvens2002.dfki.de/cd/pdf/04P-Vorrman-Lezius.pdf [14/07/2015]).

Wandl Vogt-Declerck 2013 = Eveline Wandl-Vogt - Thierry Declerck, *Mapping a Traditional Dialectal Dictionary with Linked Open Data*, in I. Kosem *et al.* (eds.), *Electronic lexicography in the 21st century: thinking outside the paper*, *Proceedings of the eLex 2013 conference*, 17-19 October 2013, Tallinn, Estonia, Ljubljana-Tallinn, Trojina, Institute for Applied Slovene Studies - Eesti Keele Instituut, pp. 460-71 (http://eki.ee/elex2013/proceedings/eLex2013_32_Wandl-Vogt+Declerck.pdf [14/07/2015]).

Webber-Eifrem-Robinson 2013 = Jim Webber - Emil Eifrem - Ian Robinson, *Graph Databases*, Sebastopol, CA, O'Reilly.