Originally published in: Rehm, Georg/Witt, Andreas/Lemnitzer, Lothar (Eds.):
Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference 2007.
– Tübingen: Narr Francke Attempto, 2007. Pp. 315-323.

Thorsten Trippel

The Use of Graphs as a Data Structure for Reusing Lexical Resources

Abstract:

Lexical resources are often represented in table form, e. g., in relational databases, or represented in specially marked up texts, for example, in document based XML models. This paper describes how it is possible to model lexical structures as graphs and how this model can be used to exploit existing lexical resources and even how different types of lexical resources can be combined.

1 Descriptive Structures in Lexical Resources

In general there are two different views on lexical resources: the first is viewing a lexicon as a special sort of text with typographic markup, the second is representing the structure of the lexicographic information. Some of the typographic markup may imply structural information, such as bold face for lexicon headwords in dictionaries, inclusion in slashes for transcriptions, etc. Another kind of lexicon is taken into account here, namely the lexicon in which the lexical information is marked up according to the information type it provides. This is usually referred to as lexicon data categories (comp., for example, ISO 12620, 1999, and its revision currently under consideration for Draft International Standard status, as well as ISO 1951, 1997). The data categories are usually described in terms of their microstructure and macrostructure, sometimes also by using a mesostructure or megastructure (see, for example, Hartmann, 2001, or Gibbon, 2002). Though the terms *lexicon* and *dictionary* are sometimes strictly separated from each other, they are here without loss of generality – and without a further discussion of reasons for separating them or regarding them as synonymous – interchangeable, as both can be described by these structures and contain similar information items in many respects. The term *dictionary* will in doubt be used for the book dictionary or modern renderings of the same for human use, while *lexicon* is used more generally, also to include termbanks and lexical resources for computer systems and for linguistic theories.

1.1 Lexicon Structures and the Treatment of Ambiguity

The microstructure of a lexicon is defined here as the sequence of data categories in each lexicon entry, i. e., the order and the selection of data categories. The lexicon macrostructure is a different aspect of the lexicon, describing the structure of the lexicon entries in relation to each other, i. e., the sequence of lexicon entries, such as the lexicon articles ordered alphabetically by the headword of the article. The lexicon mesostructure is the interrelation of lexicon entries, such as crossreferences and references to sketch-grammars, lexicon metadata as stored in the *front matter* of the lexicon, etc. The megastructure, as defined by Hartmann (2001), is not a structural property of the lexicon entries or their relation to each other, but the link to the organization of a lexicon – especially a print dictionary – as a whole and constitutes the link to the representation of lexicons. The megastructure describes the sequence of *front matter* with lexicon metadata and additional information, such as foreword and usage description, *lexicon body* with microstructure and macrostructure organization of lexicon articles and *back matter*, which corresponds to the front matter.

Other descriptions of lexicons such as the Multilingual ISLE Lexical Entry (Atkins et al., n. a.) or the Text Encoding Initiative's description of lexicon articles in print dictionaries (Sperberg-McQueen and Burnard, 2001) do not explicitly model these structures, but assume a form-meaning pair of the lexicon. In this case the form is meant to be the primary key to the lexicon article. This also applies to the emerging ISO standard for semantic lexicons called the Lexicon Markup Framework (ISO 24613, 2006, currently a Committee Draft standard). These models are appropriate for specific types of lexicons which are based on the form-meaning dichotomy. In those models especially the representation of the meaning is highly structured, allowing various forms of semantic description. They do not account for other types of lexical resources, such as form-form based lexicons – for example a pronunciation dictionary, mapping orthography to phonemic transcription (see for example Wells, 1990). Another class of lexicons not accounted for are meaning based lexical resources such as semantic networks or terminological lexicons (described for example by Wüster, 1991).

Whichever framework is used for the description of lexical resources, the common features of lexicon entries are not explicitly modeled as similarities. The missing explicity of similarities results in problematic cases often referred to in lexical semantics (comp. Cruse, 1986). These problematic cases are essentially cases of ambiguity, for example, for homographs or polysemous words where the form-meaning relation is not a 1:1 relation, i. e., one wordform has exactly one meaning, but where there is a 1:n relation, i. e., one wordform has two or more meanings.

In print lexicons authors often distinguish between homonyms and polysemous words in the way they treat words with different meaning but the same spelling. Homonyms are often represented by appending a number to the repeated headword with distinct lexicon entries for each distinguished headwords, while for polysemous words the different meanings are enumerated within the same entry along with different definitions, wordclasses or usage examples. Both correspond also to the sometimes used expression of different *readings* of a word.

Very similar to the treatment of homonyms in print dictionaries is the handling of synonymy in termbases. In termbases the ambiguity is a 1:n meaning-form relation, i. e., one meaning has more than one form. In terminology, the disambiguation is achieved by assigning a usage attribute, saying for example that one form is preferred or deprecated (see, for example, Arntz

and Picht, 1989, and Sager, 1990 for an introduction to terminology organization). Hence, the treatment of ambiguity in one way or the other is possible by either using a modification of the lexicon article's headword or by including a parallel structure under the headword distinguished by specific discriminating features.

1.2 The Parallel Use of Different Lexical Resources

Structural differences in handling ambiguity is only one area of problems when using in parallel lexical resources originating from different sources. It can at least be assumed that two lexicons that contain some similar data categories and some common lexicon entries can be combined with each other. A combination of lexical resources can become necessary, for example, due to a merger of publishing houses or the creation of a larger lexicon base in a Natural Language Processing (NLP) context. One example of a combined use of different lexical resources is the use of a bilingual lexicon in parallel with a monolingual dictionary by a human user where the user looks for a translation of a word and then 'double checks' in a monolingual lexicon to make sure that the word has been used correctly. In this case both lexical resources share a formal description, e. g., the orthography of a word in one language, but provide additional mutually exclusive information. Another case is the use of two lexicons of the same kind, e. g., two bilingual lexicons of the same language pair because of different coverage. In this case the microstructure will be rather similar without taking into consideration the order of data categories, and taking into account that some data categories can be omitted by one lexicon editor or the other.

The combination of lexical resources imposes certain restrictions, i. e., it is not possible to append one lexicon to the other and apply the macrostructure (ordering) to the new resource, even if the data category sequence is normalized by reordering and normalizing the data categories by including previously omitted data categories with default values, e. g., an empty value or another type of null value, or by deleting data categories that are available only in one of the resources. Neither the loss of information, nor the inclusion of default values is acceptable, but even in this case the problem would remain that there could be many duplicates in the resulting lexical database, if both lexicons had the same coverage even every lexicon entry could be double. One way of solving this problem would be manual or automatic post-editing such as filtering out identical entries (for procedures see, for example, Heinrich, 2004). However, filtering is not a structural, but a heuristic approach, even though it is admitted that with NLP procedures these can show good results.

2 Modeling Lexicons as Graph Structures

A conclusive model for lexicons needs to fulfill certain requirements, i. e., it has to

- allow the redundancy free combination of lexical resources without information loss (unification of lexicon resources),
- provide a way of representing the lexicon structures, i. e., the lexicon microstructure, mesostructure and macrostructure (structure mapping),
- define a way of disambiguating ambiguities such as homonyms and synonyms.

As the table structure does not provide for unification and the disambiguation implies the change of the content – especially the modification of headwords but also the addition of data categories for classification purposes – the table structure does not seem to be appropriate. Instead, a graph structure is proposed for lexicons (cf. Trippel, 2006).

2.1 Definition of Lexicon Graphs

A graph in general consists of nodes – also called vertices – and edges, often drawn as arcs between certain nodes (for a formal definition, see for example Wilson, 1972). In a graph model for the lexicon, the *Lexicon Graph Model*, every lexical information item, i. e., every distinct value of one data category in at least one lexicon entry, is represented by one and only one node. For example, for every orthography of a word there exists one node, every part of speech used in the lexical resource is represented by a node, every definition is a node, etc. Nodes are typed, so as to distinguish wordclasses from orthography. However, some data categories are of the same type, for example, a data category *antonym* is in many lexical resources an implicit crossreference to an orthography of another word. That means that the content of this data category is of the same type as the orthography.

Without any edges the nodes are disconnected, i. e., they do not have any relation to one another. But the lexical resources impose a relation between certain items of lexical information, i. e., between some nodes, by assigning a definition to an orthographic representation of a word in a lexicon entry. These relations are the edges of the model. However, not every edge is the same, there are classes of edges. The class of edges relating an orthography to a definition is for example different from the class of edges connecting a definition to its antonym's definition. This shows that there can be relations between the same type of nodes, hence to distinguish the classes from each other, a typisation for the edges is also introduced.

A third type of element of this Lexicon Graph should be included as well: a class of nodes that is very particular in the sense that it represents structured information as represented in the front matter of lexicons, which can consist of sketch-grammars, or metadata, etc. To distinguish these special lexical information items but on the other hand to include them as they may be referenced from the lexicon body, but are not part of the lexicon information, external knowledge representation nodes are included. They can consist of links to external information as well as whole embedded hierarchies such as ontologies. In terms of relations they can be treated as regular nodes of the graph.

The nodes of the Lexicon Graph do not need to be of textual type, in fact the Lexicon Graph allows the process of including information from different modalities, for example, by referencing multimodal information by ways of Uniform Resource Identifiers (URI). In this sense the Lexicon Graph Model is related to the formal models of RDF(S) (Lassila and Swick, 1999, and Binckley and Guha, 2003).

2.2 Disambiguation in a Lexicon Graph

Figure 1 shows a simplified Lexicon Graph from the unification of two lexicon entries, one being originally for the word *unity*, the other for *drill*. These arbitrary words do not seem to be related, but there is at least one common feature, i. e., both can be of the wordclass *noun*.

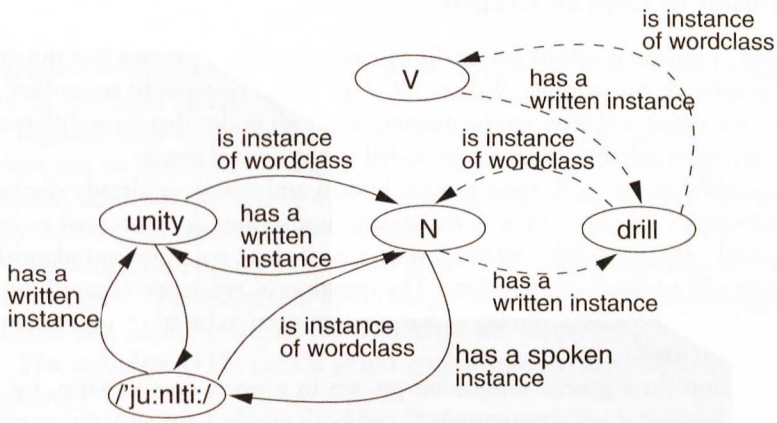

Figure 1: Sample Lexicon Graph with five nodes; this graph represents entries from two different lexicons, where the dashed relations are from one lexicon, the solid ones from another

For disambiguation in a lexicon graph, no changes need to be applied to any nodes, ambiguity is omnipresent, as most nodes have different edges. Ambiguity can be seen if edges are of the same type and lead to nodes of the same type. In figure 1 such ambiguity can be seen for the nodes *N* having an arc to an orthography *unity* and *drill*; *drill* is also ambiguous, having a 'part of speech arc' to *N* and *V*. The other nodes may be connected to more then one node, but these nodes are distinguished by their different type.

For disambiguation of different 'readings' it becomes necessary to take into account discriminating edges, i. e., if one node n_1 has two edges of the same type leading to two different nodes n_2, n_3 which have the same type, another node n_4 has to be used for disambiguation. There has to be an edge connecting n_1 and n_4 but there is only one edge between n_4 and n_2 or n_4 and n_3. If there is no such n_4, a disambiguation based on the lexical information provided is not possible. In figure 1, the disambiguation can be done by selecting the orthography of a noun that has a specific transcription, i. e., selecting the `N`-node, selecting any orthographic form that is associated with it. The result would be ambiguous, because there are two orthographic forms. This results in a disambiguation process where that orthography node is selected, which is related to a pronunciation `/ju:nIti:/`, if this pronunciation is also related to the part of speech `N`. By this way the node `drill` can be ruled out. The wordclass of the node `drill`, however, cannot be disambiguated as there is no other node available.

The latter case is a special case of ambiguity which cannot be resolved, because there are conflicting values of nodes, such as an orthographic form that is linked to two different wordclasses. A query for the wordclass of `drill` has to state the ambiguity that is inherent and cannot be resolved without the extension of the lexicon.

2.3 Unification of Lexicon Graphs

The unification of lexicon graphs is essentially a graph unification process (for the definition of the union of graphs see, for example, Wilson, 1972, p. 18). If two graphs are unified, the nodes that have the same value and type are maintained, as are all nodes that have different values or types. Edges that exist only in one graph are added to the unified graph.

Conflicting values in the unification process lead to ambiguity, as already discussed in the context of ambiguity resolution. Hence, the lexicon unification does not lead to information loss. For practical purposes a different issue arises, i. e., if there is an efficient algorithm for the implementation of a unification procedure. The question of efficiency of graph unification is not addressed here, as there are a number of features that need to be taken into account, besides the implementation itself.

One of the factors for a general unification process in a worst case scenario, i. e., all nodes of two graphs to be unified are interconnected, and both graphs have only one common node, the algorithm would need to compare all nodes $\#n_1$ of the first graph, all edges $\#e_2$ of the first graph with $\#n_2$ nodes of the second and $\#e_2$ of the second, resulting in $n_1 * n_2 + e_2 * e_1$ operations. The number of operations would increase with the number of nodes involved dramatically, the algorithm would be rather inefficient, being limited by a non-polynomial increase of calculations with an increasing number of nodes.

The worst case scenario is not the real world scenario in the case of the lexicon graph, as – speaking in microstructure terms – some lexical data categories are closed classes, reducing the number of nodes significantly, and not all nodes are in relation to one another (if an `unrelated-` relation is not defined). It needs to be investigated if these restrictions are sufficient to proof that there is a limit to the calculation given by a polynomial function.

2.4 Lexicon Structure Mapping

The lexicon structures, such as the microstructure, are not maintained in the graph though there has not been any information loss. The way of maintaining the structure has to be represented differently: the structures are available in distinct subgraphs of the graph. A subgraph can be selected by a query to the graph, selecting a first node to enter the graph – for example, a type of node such as the orthography nodes – and the selection of edges of a specific type with attached nodes leading from these nodes.

The result is that the Lexicon Graph is independent from a concrete lexicon microstructure and macrostructure, though the mesostructure is explicitly defined in the lexicon. The microstructure and macrostructure hence become ways of defining subgraphs, allowing to operate on a common underlying lexical structure, the lexicon graph. The traditional structures are in this case only concurring selection processes. A *lexicon* in the traditional sense is then defined as a *lexicon graph* together with a *subgraph selection function*, which contains the structure of the lexicon. If two lexicons are merged, the resulting Lexicon Graph can extend the original lexicons, together with the original structure defined in a subgraph selection function, if the lexical data categories are the same. The worst case in this scenario would be that the data categories are so varied and the overlap of the two lexicons would be so small that no extension would be the result; however, the lexical information would be maintained.

3 Implementation

The implementation of the Lexicon Graph is done by a simple document grammar available as a DTD and an XSchema (Thompson et al., 2004). In both, the elements of nodes and edges are defined together with a type attribute. For processability they also receive an identifier, so that the edges can be defined as being edges of a specific type between specific nodes. The implementation, however, is working with directed graphs to ease the definition of types and to allow non-symmetrical relations between nodes.

The elements defined are the nodes of lexical information, special external information nodes and edges between the different nodes, each class defined in their own container, each element receiving a type and identifier attribute. The edges are modeled by the start and end node identifiers. The complete DTD for the graph implementation is available online (`http://www.lexicongraph.de`).

The merging program was implemented in XQuery(Boag et al., 2003), as were some lexicon views, for example, to create the view of the test lexicons. These XQuerys are also available from `http://www.lexicongraph.de`. To store and process the data the XML database Tamino (2003) was used, though for small graphs the storage in the file system with operations using the saxon XQuery processor (Kay, 2004) was tested as well.

4 Evaluation

To evaluate the Lexicon Graph approach, a test set of lexicon entries was created based on different lexicons. For test purposes sample entries from lexicons as different as, for example, the Bielefeld Verbmobil reference lexicon (Gibbon and Lüngen, 2000), the Bonn Machine-readable Pronunciation dictionary (Portele et al., 1995), and the Oxford Advanced Learner's Dictionary (Crowther, 1995), were used. The test consisted of

- a transfer of each of the sample lexicon entries into a lexicon graph, usually using a structured XML rendering of the lexicons, which were for the machine readable lexicons generated automatically from the available resource,
- a creation of an XQuery which produces the subgraph relevant for each of the basic lexicons,
- merging the lexicon graphs, and
- testing the views of the graphs on the unified graph.

Every test showed that there is no information loss by regenerating the basic entries. Based on the unified lexicon, further lexicon views were created, for example, for a pronunciation lexicon (pronunciation-orthography mapping), part of speech lexicon (orthography-wordclass mapping) and others. These views showed, as expected, an extension of the original lexicons without the addition of further ambiguities. These lexicons were also used to define use-cases for disambiguation based on discriminating features as described in Trippel (2006). All tests indicate that the model fulfills its requirements of not causing information loss.

5 Summary and Future Work

The Lexicon Graph is an appropriate data structure for different kinds of lexicons and allows the reuse of lexical resources in other contexts than the original intended ones. For this purpose, a graph is defined in which the nodes consist of the lexical information, i. e., the unique values of all distinct lexical data categories. The edges in the graph represent the relations between those nodes resulting from being part of the same lexicon entry. The structures of lexicons are subgraph structures, which can be extracted by querying the graph. After the unification of different graphs, the result can be an extended lexicon, if the original data categories are compatible, if not there is at least no information loss. It was discussed that the lexicon graph can serve as the basis of other types of lexicons as well, which have an extended coverage than the non-unified lexicon graphs.

Open issues addressed are the investigation of the complexity of lexicon graph unification and if the constraints imposed by a lexicon are sufficient to allow efficient unification processes. For this purpose the Lexicon Graph needs to be further formalized and its mathematical properties need to be investigated. Another issue would be to extend the Lexicon Graph by publicly available and machine processable lexical resources to create a larger graph. The resulting graph could be the basis for other lexicon views and be made publicly available, if the views allow for an effective implementation, i. e., if the complexity of the subgraph selection permits it.

Bibliography

Arntz, Reiner and Picht, Heribert (1989): *Einführung in die Terminologiearbeit.* Hildesheim, Zürich, New York: Olms.

Atkins, Sue; Bel, Nuria; Bouillon, Pierrette; Charoenporn, Thatsanee; Gibbon, Dafydd; Grishman, Ralph; Huang, Chu-Ren; Kawtrakul, Asanee; Ide, Nancy; Lee, Hae-Yun; Li, Paul J. K.; McNaught, Jock; Odijk, Jan; Palmer, Martha; Quochi, Valeria; Reeves, Ruth; Sharma, Dipti Misra; Sornlertlamvanich, Virach; Tokunaga, Takenobu; Thurmair, Gregor; Villegas, Marta; Zampolli, Antonio and Zeiton, Elizabeth (n. a.): "Standards and Best Practice for Multilingual Computational Lexicons and MILE (the Multilingual ISLE Lexical Entry)". Deliverable d2.2-d3.2 isle computational lexicon working group, International Standards for Language Engineering (ISLE), Pisa. `http://www.ilc.cnr.it/EAGLES96/isle/clwg_doc/ISLE_D2.2-D3.2.zip`.

Binckley and Guha (2003): "RDF Vocabulary Description Language 1.0: RDF Schema". `http://www.w3.org/TR/rdf-schema`.

Boag, Scott; Chamberlin, Don; Fernandez, Mary F.; Florescu, Daniela; Robie, Jonathan and Siméon, Jérôme (2003): "XQuery 1.0: An XML Query Language". `http://www.w3.org/TR/xquery/`, W3C Working Draft 02 May 2003.

Crowther, Jonathan (editor) (1995): *Oxford Advanced Learner's Dictionary of Current English.* Oxford: Oxford University Press, 5th edition.

Cruse, David A. (1986): *Lexical Semantics.* Cambridge: Cambridge University Press.

Gibbon, Dafydd (2002): "Prosodic Information in an Integrated Lexicon". In: *Proceedings of the Speech Prosody 2002 Conference*, edited by Bel, B. and Marlien, I. Aix-en-Provence: Laboratoire Parole et Langage, pp. 335–338.

Gibbon, Dafydd and Lüngen, Harald (2000): "Speech Lexica and Consistent Multilingual Vocabularies". In: *Verbmobil: Foundations of Speech-To-Speech Translation*, edited by Wahlster, Wolfgang, Berlin, Heidelberg, New York: Springer, pp. 296–307.

Hartmann, Reinhard R. K. (2001): *Teaching and Researching Lexicography*. Applied Linguistics in Action. Harlow: Pearson Education.

Heinrich, Ines (2004): *Entwicklung eines Konzepts zur Erkennung und Bereinigung doppelter Datensätze in einer Terminologiedatenbank*. Master's thesis, Hochschule Anhalt für angewandte Wissenschaften (FH), Köthen.

ISO 12620 (1999): "Computer Applications in Terminology – Data Categories". Technical Report, ISO. International Standard.

ISO 1951 (1997): "Lexicographical Symbols and Typographical Conventions for Use in Terminography". Technical Report, ISO. International Standard.

ISO 24613 (2006): "Language Resource Management – Lexical Markup Framework (LMF)". Technical Report, ISO. Committee Draft International Standard.

Kay, Michael (2004): "Saxon Basic Edition". `http://www.saxonica.com`, Version 8.0 B.

Lassila and Swick (1999): "Resource Description Framework (RDF) Model and Syntax Specification". `http://www.w3.org/TR/REC-rdf-syntax`.

Portele, T.; Krämer, J. and Stock, D. (1995): "Symbolverarbeitung im Sprachsynthesesystem HADIFIX". In: *Proceedings of 6. Konferenz Elektronische Sprachsignalverarbeitung*. Wolfenbüttel, pp. 97–104.

Sager, Juan C. (1990): *A Practical Course in Terminology Processing*. Amsterdam, Philadelphia: Benjamins.

Sperberg-McQueen, C.M. and Burnard, Lou (editors) (2001): *The XML Version of the TEI Guidelines*, The TEI Consortium, chapter "Print Dictionaries". TEI P4 edition. `http://www.tei-c.org/P4X/index.html`.

Tamino (2003): "Tamino XML Server". Version 4.1.1. Database by Software AG, Darmstadt, Germany.

Thompson et al. (2004): "XML Schema Part 1: Structures Second Edition". `http://www.w3c.org/TR/xmlschema-1/`.

Trippel, Thorsten (2006): *The Lexicon Graph Model: A Generic Model for Multimodal Lexicon Development*. Saarbrücken: AQ Verlag.

Wells, John C. (1990): *Longman Pronunciation Dictionary*. Harlow: Longman.

Wilson, Robin J. (1972): *Introduction to Graph Theory*. Edinburgh: Oliver and Boyd.

Wüster, Eugen (1991): *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Bonn: Romanistischer Verlag, 3rd edition.