Linguistic Applications of Formal Concept Analysis[⋆]

Uta Priss

School of Computing, Napier University, u.priss@napier.ac.uk

1 Introduction

Formal concept analysis as a methodology of data analysis and knowledge representation has potential to be applied to a variety of linguistic problems. First, linguistic applications often involve the identification and analysis of features, such as phonemes or syntactical or grammatical markers. Formal concept analysis can be used to record and analyze such features. The line diagrams of concept lattices can be used for communication among linguists about such features (see section 2).

Second, modeling and storage of lexical information is becoming increasingly important for natural language processing tasks. This causes a growing need for detailed lexical databases, which should preferably be automatically constructed. Section 3 describes the role that formal concept analysis can play in the automated or semi-automated construction of lexical databases from corpora.

Third, lexical databases usually contain hierarchical components, such as hyponymy or type hierarchies. Because formal concept lattices are a natural representation of hierarchies and classifications, lexical databases can often be represented or analyzed using formal concept analysis. This is described in section 4.

It should be remarked that because this paper appears in a collection volume of papers on formal concept analysis, the underlying notions, such as formal concept, formal object and attribute, and lattice, are not further explained in this paper. The reader is referred to Ganter & Wille (1999) for detailed information on formal concept analysis.

2 Analyzing Linguistic Features with Formal Concept Analysis

Linguists often characterize datasets using distinct features, such as semantic components, phonemes or syntactical or grammatical markers, which can easily be interpreted using formal concept analysis. An example is Kipke & Wille's (1987) paper which applies formal concept analysis to semantic fields. In that paper, data from an analysis of the German and English words for the semantic field of "bodies of water" is modeled using different types of concept lattices.

A different but also feature-based analysis is conducted by Großkopf (1996) who analyzes verb paradigms of the German language and by Großkopf & Harras (1999) who analyze speech act semantics of German verbs. Großkopf & Harras use formal contexts which have verbs as objects and semantic features that characterize the speech

[⋆] This is a preprint of a paper published in Ganter; Stumme; Wille (eds.), Formal Concept Analysis, Foundations and Applications. Springer Verlag. LNAI 3626, 2005, p. 149-160. ©Springer Verlag.

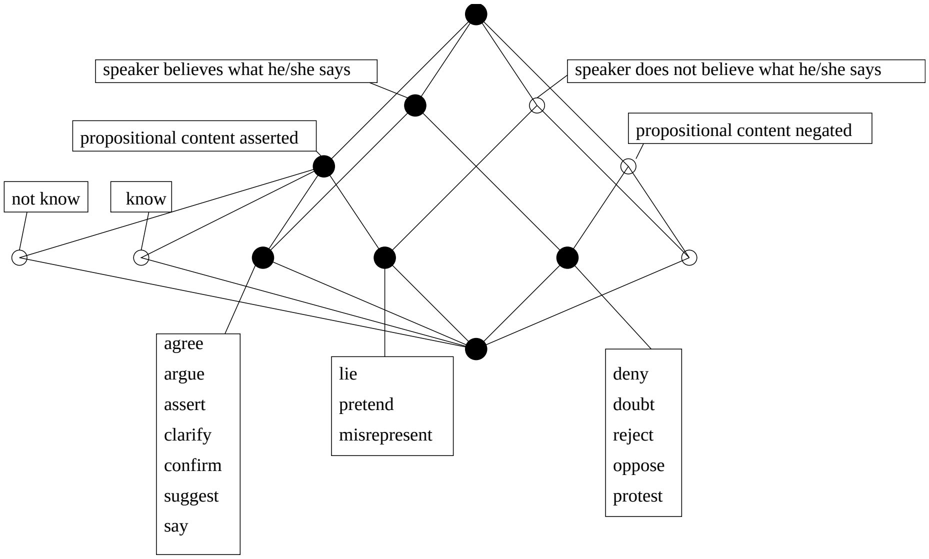
speaker believes what he/she says
speaker does not believe what he/she says
propositional content asserted
propositional content negated
not know
know
agree
argue
assert
clarify
confirm
suggest
say
lie
pretend
misrepresent
deny
doubt
reject
oppose
protest

Fig. 1. Verb speech acts in analogy to Großkopf & Harras (1999)

acts of the verbs as attributes. Figure 1 shows an English example which is equivalent to of one of Großkopf and Harras's German examples. The lattice shows the attitudes of speakers for some verbs of assertion. For verbs, such as "agree" and "lie", a certain propositional content is asserted. But in one case (assert) the speaker believes the content, in the other case (lie) the speaker does not believe the content. Both cases are different from verbs, such as "deny", for which the speaker neither believes the content nor pretends to believe it.

The lattice in the example is a conceptual scale which can be applied to different sets of verbs. In the example, some of the attributes ("know" and "not know") are not relevant for the chosen set of verbs. But these attributes are relevant when comparing verbs of knowledge acquisition (e.g., "learn") with verbs of assertion. Großkopf and Harras built a Toscana system which facilitates browsing through a large selection of German verbs using different conceptual scales. A researcher can interactively analyze the semantic structures of sets of verbs using such a system. Großkopf and Harras's paper contains further examples and some linguistic results achieved with these methods.

It is somewhat surprising that formal concept analysis is not used more frequently for the analysis of linguistic features. It may be that linguists are still not widely aware of these methods. There has been a recent interest in formal concept analysis in areas of formal linguistics. In modern theories of grammar, such as head-driven phrase structure grammar, HPSG (Pollard & Sag, 1994), lexical knowledge is organized in hierarchies of feature structures of classes of lexical entries. According to Sporleder (2002) and Osswald & Petersen (2002), formal concept analysis may be a suitable method for deriving such hierarchies of lexical information automatically. But so far, their research is more theoretical than practical. No detailed examples of applications have yet been published in this area.

3 The Use of Formal Concept Analysis in Natural Language Processing Tasks

This section focuses on an approach developed by Basili et al. (1997). Because this approach has significance for many applications both in linguistics and in AI, we cover this paper in greater detail. We first explain the underlying problems of natural language processing before describing the actual implementation in section 3.3.

3.1 The limits of automated word sense disambiguation

It is a central task for any natural language processing application to disambiguate the words encountered in a text and to determine to which lexical entries they belong. Normally this involves both the use of a lexicon as well as the use of a syntactic parser. But there is an inverse relationship between lexical and syntactic processing: a simple and shallow lexicon requires a detailed syntactic parser whereas a detailed lexicon requires only a shallow parser. The tendency in recent linguistic research has been to focus on improvements in the representation and detail of lexica. But there may be a limit to how much detail can be represented in a lexicon. For example, Basili et al. (1998) claim that "in any randomly chosen newspaper paragraph, each sentence will be likely to have an extended sense of at least one word, usually a verb, in the sense of a use that breaks conventional preferences and which might be considered extended or metaphorical use, and quite likely not in a standard lexicon."

The following examples from Pustejovsky (1991) illustrate the problem: in the sentence "the woman finished her beer", the verb "finish" is synonymous to "stopped drinking" whereas in "the man finished his cigarette" it is synonymous to "extinguished". From a common sense viewpoint, it is obvious that the shared meaning between both sentences is the act of finishing an activity. But a detailed lexical representation should include the logical inferences which are different. A natural language processing application might need to 'know' that an appropriate activity after "finishing" is "ordering a new one" in the case of beer, and "lighting another one" in the case of cigarettes. This information could either be stored with "cigarettes" and "beer" or with "finishing". It could also be stored with more general terms ("alcoholic beverage") or in rule format. This example shows that the issues are complicated.

A second example, is "she looked through the window" as opposed to "he painted the window". In the first case, "the window" stands for "the glass of the window". In the second case, it stands for "the frame of the window". A natural language processing application might need to 'know' that glass is transparent and more likely to break and that frames can be made of a variety of materials. Essentially, a lexical database might need to store detailed descriptions of windows. But because many different types of windows exist, this would be an enormous amount of information.

Priss (2002) asserts that the types of problems outlined in these examples may require more than linguistic knowledge because even humans do not store this information in a linguistic format. Humans parse such linguistic ambiguities by utilizing common sense non-linguistic knowledge about the world. Ultimately, automated natural language processing may need to be combined with other methods, such as perceptive, robotic interfaces or associative interfaces which facilitate better co-operation between humans and computers.

3.2 Lexical databases and ontological knowledge

There is no clear dividing line between lexical databases and AI ontologies. Traditionally, lexical databases are often less formalized than AI ontologies. For example, WordNet (Fellbaum, 1998) is less formalized than the AI ontology, CYC (2003). But contemporary linguistic representations of lexical knowledge can be as formalized and logic-based as formal ontologies.

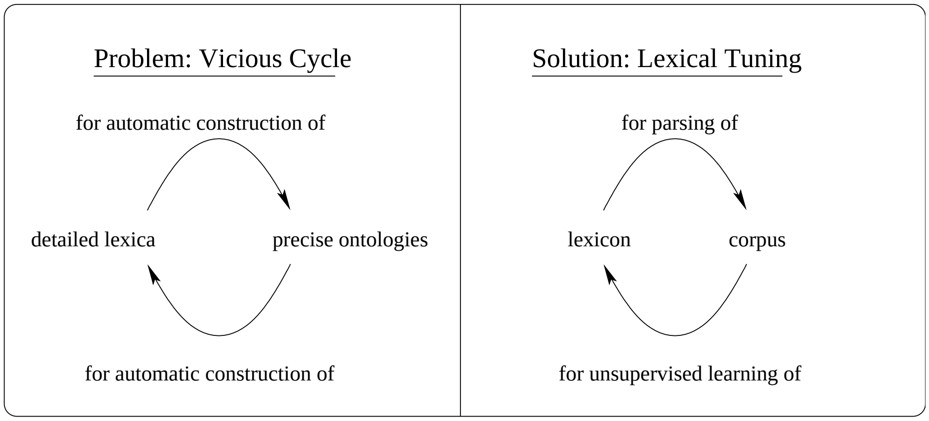

Fig. 2. A vicious cycle and a possible solution

As depicted in the left half of figure 2, lexical databases and AI ontologies form a vicious cycle because the construction of detailed lexica requires precise ontological information, and vice versa. Linguistic structures convey information about the knowledge domains which are expressed in a language. For example, Scottish people have more words for rain than people who live in dryer climates. Other often quoted examples are kinship terms, color terms (not just cultural differences but also the number of color terms used, for example, in the paint industry) and food vocabulary. Thus lexica can serve as a source of information for AI ontology engineers.

Ontologies are useful for natural language processing, because differences in meaning correspond to syntactic differences, such as whether a verb is used transitively or intransitively. Before processing text, it is helpful, to first construct a model of the underlying semantic structures and to describe an underlying AI ontology. Because of the large size of lexica and because of the ever changing nature of language, both the construction of lexica and of AI ontologies should be semi-automatic to achieve sufficient depth. Large corpora, such as the World Wide Web, provide an almost unlimited resource for lexical input. But the vicious cycle holds for them as well: sophisticated parsing presupposes already existing sophisticated lexical databases and vice versa. As Basili et al. (2000) state: words can be more easily disambiguated if one knows their selectional restrictions, but selectional restrictions can be more easily identified if one knows the disambiguated meanings of their words.

3.3 Bootstrapping of linguistic information from large corpora

Fortunately, vicious cycles can sometimes be dissolved by bootstrapping. In this case, Basili et al. (1997) claim that a lexicon of any quality can be used as a starting point. As depicted in the right half of figure 2, the lexicon is improved by corpus-driven unsupervised learning from a corpus. The improved lexicon is then re-applied to the parsing of the corpus which improves the learning from the corpus and so forth. This form of bootstrapping is used for "lexical tuning" (Basili et al., 1997), which is the process of refining lexical entries to adapt them to new domains or purposes. Basili's approach, which is described in the remainder of this section, is based on formal concept analysis. Formal concept analysis provides an excellent tool for lexical tuning because the duality between formal objects and attributes is used to represent linguistic dualities, such as verb frames and noun phrases or corpus-derived sentence parts and rule-based lexical structures, which drive the bootstrapping process.

Verbs are more difficult to disambiguate than nouns or adjectives. Verb subcategorization frames are the result of classifying verbs based on their argument structures. For example, the verb "multiply" usually has a mathematical meaning (e.g., "multiply 3 by 4"), if it occurs with a direct object, and a biological meaning (e.g., "the bacteria multiplied"), if it occurs without a direct object. The prepositional modifiers and syntactic constituents that co-occur with a verb indicate which subcategorization frame is used. The frames often correspond to different senses as recorded in a lexicon. A verb in a corpus can often be matched to its sense by comparing its subcategorization frame with the frames in the lexicon entry for that verb.

Basili et al. (1997) describe the following machine-learning method for extracting verb subcategorization frames from a corpus. After parsing the corpus, a formal context is derived for each verb. The objects of the formal context are all phrases that contain the verb. The attributes are the arguments of the verbs, such as direct or indirect objects and prepositions. Phrases that have the same argument structure are clustered in the extension of an object concept.

At this stage the concept lattices only represent possible subcategorization frames. Natural language processing as used for the parsing and for the automatic identification of the verb argument structures from the corpus does not yield error-free results. Furthermore because language is often ambiguous, there can never be absolutely perfect results. Thus after generating the lattices, it still needs to be identified which concepts in each lattice represent the senses of the verb which are most acceptable for the corpus.

Basili et al. describe methods to assign weights to the nodes of the concept lattices of each verb, which facilitate the selection of the most relevant subcategorization frames. These weights can be trained using a training subset of a corpus. Probabilistic models for prepositions can also be derived that correspond to the specific domain of the corpus. These can then be incorporated into the calculations. Using these methods, the most probable verb frames can be selected from the context.

For improved precision, Basili at al. point out that instead of automatic extraction, a human lexicographer could also use the lattice for each verb as a decision tool in manual lexicon construction. Since estimates state that constructing a lexical entry (in English) by hand takes at least one hour, using the automatically derived lattices for each verb as a guideline can be a significant time-saver.

3.4 Possible applications for ontology engineering

Basili at al. (1997) experimentally evaluate their approach for the purpose of lexical tuning and conclude that it improves on other techniques. Their approach has subsequentially been successfully implemented in a variety of projects of lexical tuning (for example, Basili et al. (1998)). We believe that this approach is very promising and can have many more applications. For example, conceptual graphs (Sowa, 1984) are very similar to verb subcategorization frames. Concept lattices can be automatically derived, that have the core concepts from a set of conceptual graphs as objects and the relations as attributes. These can be used to analyze the text from which the conceptual graphs are derived. More importantly, they can be used to design concept hierarchies based on the structure of the arguments, which can be used by ontology engineers.

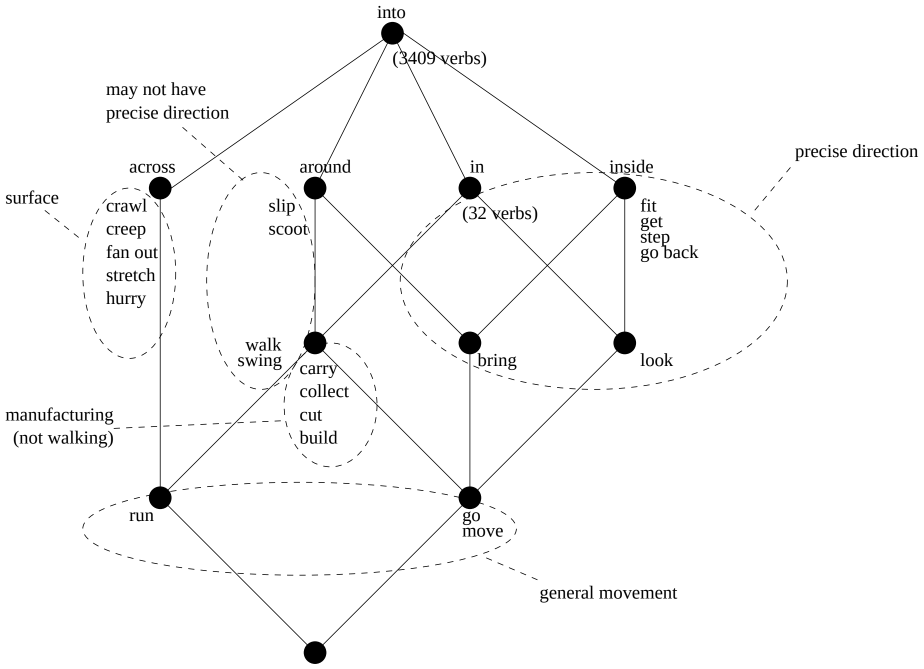

Fig. 3. Clustering of verbs from a corpus

Figure 3 shows an example, which is derived from the semantically tagged Brown corpus that is available for WordNet 1.6. The underlying algorithm is even simpler than Basili's approach: instead of determining the full argument structure of the phrases, only the first preposition that follows a verb is considered. While Basili's approach constructs a separate lattice for each verb, our algorithm is applied to all verbs from a corpus. The attributes of the lattice in this example are the prepositions “into”, “across”, “around”, “in” and “inside”. The objects of the lattice are all verbs in a subset of the corpus which are at least once followed by “into”. Even this basic algorithm clusters the verbs in potentially meaningful groups, which are indicated by the circles. The circles are not meant to be automatically derivable. Instead they are open to human interpretation of the data.

It is conceivable to apply this algorithm to the complete Brown corpus (or other corpora) and to obtain an event hierarchy that can be useful for ontology engineering. Ontologies often focus on noun hierarchies and overlook hierarchies based on verbs, event structures or argument structures. A promising exception is Basili et al.'s (2002) paper which develops event hierarchies for information extraction and multilingual document analysis.

Another approach is presented by Petersen (2001). She uses formal concept analysis for the automatic acquisition of lexical knowledge from unstructured data. She reports on an application of computing a hierarchy of derivational information for English and German lemmas from the lexical database CELEX. Osswald & Petersen (2002) describe why formal concept analysis is suitable for automatic induction of classifications from linguistic data. A similar approach is also described by Sporleder (2002).

4 Representing Lexical Databases

Apart from bootstrapping, there are other possibilities to avoid the problem that common sense knowledge may not be completely representable in linguistic structures. Instead of attempting to represent as much information as possible in detailed logical formalisms, it may be sufficient to represent some information in an easily human-readable format, which can then be browsed through and explored by human users. Thus humans and machines co-operate in information processing tasks.

With respect to linguistic applications, an important but also challenging task is to construct interfaces for lexical databases. As mentioned before, the structures in lexical databases are often not very different from AI ontologies, such as CYC. For example, WordNet contains a noun hierarchy, which is similar in structure to taxonomies in ontologies, to classification systems or to object-oriented type hierarchies. Thus representations of hierarchies in lexical databases have similar applications as AI ontologies and classifications. They can be used to browse through collections of documents in information retrieval, to visualize relationships in textual information, to aid in the structuring and classification of scientific knowledge and they can serve as an interlingua.

The following sections describe how to formalize lexical databases in terms of formal concept analysis and how to use the formalizations for an analysis of semantic relations and for comparing and merging of lexical databases.

4.1 Formalizing lexical databases, thesauri or ontologies

Roget's Thesaurus and the lexical database, WordNet, have both been formalized with formal concept analysis methods. Roget's Thesaurus contains a six-level classification. At the lowest level, words are grouped that are either synonymous or closely related according to some semantic field, such as animals or food items. Because of polysemy, many words occur multiple times in the thesaurus. Thus one could construct a concept lattice using words as objects and their thesaurus classes as attributes to explore polysemy. But a lattice of the whole thesaurus would be far too large to be visualized. Therefore, Wille (1993) describes a method of constructing smaller, so-called "neighborhood" lattices. The semantic neighborhood of a word consists of all words that share some meanings with the word, that means all words which co-occur with the original word in at least one bottom-level class of the thesaurus. The set of objects of a neighborhood lattice consists of such a semantic neighborhood of a word. The attributes are either all bottom-level thesaurus classes of the original word, or all bottom-level thesaurus classes of all the words in the neighborhood. The choice depends on whether a larger or a smaller semantic environment is intended.

As an example, figure 4 shows the neighborhood lattice of "over" according to Old (1996). The lattice shows that there are two separate clusters of meanings of "over": the temporal/completion/direction senses ("it is over") are distinguished from the other senses. This is visible in the lattice because the six nodes in the right only share top and bottom with the rest of the lattice. The left side of the lattice shows that "over" can be used at different levels of intensity: "addition", "excess", "superiority" and "distance, past". Other similar examples can be found in Sedelow & Sedelow (1993).

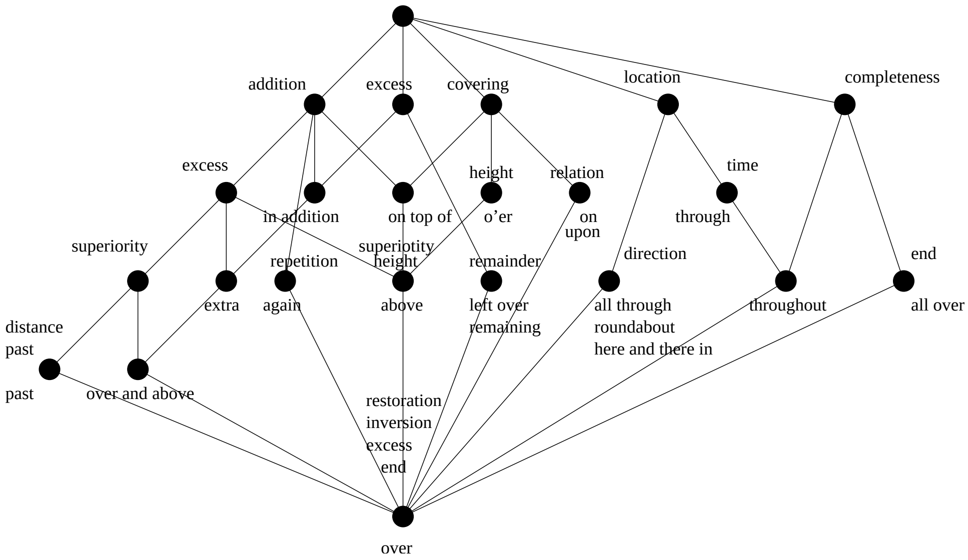

Fig. 4. Neighborhood lattice of "over" (Old, 1996)

In contrast to Roget's Thesaurus, the synonym sets in WordNet occur at all levels of the hierarchy not just at the bottom. Furthermore the hyponymy hierarchy is only one relational structure in WordNet. Not all of the other relations in WordNet are hierarchical. Although the part-whole relation is hierarchical, it is not necessarily meaningful to embed it into a lattice because intersections of parts may not be meaningful. For example, in a "substance-of" relation, ketchup and cake both contain sugar and salt, but is "sugar and salt" a meaningful grouping? For these reasons, WordNet requires a slightly different modeling than Roget's Thesaurus. Priss (1998) develops relational concept analysis as a means for modeling WordNet and similar lexical databases. One advantage of this approach is that semantic relations can be implemented using bases, that means a reduced set of relation instances from which the complete relation can be derived (Priss, 1999).

4.2 Analyzing semantic relations in lexical databases

The formalization of WordNet and Roget's Thesaurus as described in the previous section, can serve as a basis for a linguistic, cognitive or anthropological analysis of the structures and the knowledge that is encoded in such lexical databases. Priss (1996) uses relational concept analysis to analyze meronymy (part-whole) relations in WordNet. Classifications of types of meronymy that are manually derived based on semantic analysis are often fuzzy and never agreed upon among different researchers. Priss shows that relational concept analysis facilitates the derivation of a classification of meronymy that is based on an entirely structural analysis. This classification is less fuzzy but still retains a significant amount of the information contained in manually derived classifications.

Another method of analyzing WordNet using formal concept analysis, identifies "facets" in the noun hierarchy. Facets are regular structures, such as "regular polysemy" (Apresjan, 1973). Figure 5 shows an example of family relationships in WordNet, which are arranged into facets (or scales) in figure 6.

4.3 Comparing or merging lexical databases

Once lexical databases are formalized as concept lattices, the lattices can serve as an interlingua, for creating multilingual databases or to identify lexical gaps among different languages. This idea was first mentioned in Kipke & Wille's (1987) concept lattices of the semantic fields of "bodies of water" in English and German. The lower half of the example in figure 7, which is taken from Old & Priss (2001), shows separate concept lattices for English and German words for "building". The main difference between English and German is that in English "house" only applies to small residential buildings (denoted by letter "H"), whereas in German even small office buildings (denoted by letter "O") and larger residential buildings can be called "Haus". Only factories would not normally be called "Haus" in German. The lattice in the top of the figure constitutes an information channel in the sense of Barwise & Seligman (1997) between the German and the English concept lattice. In the process of manually creating such a channel a linguist must identify the attributes, which essentially describe the difference between

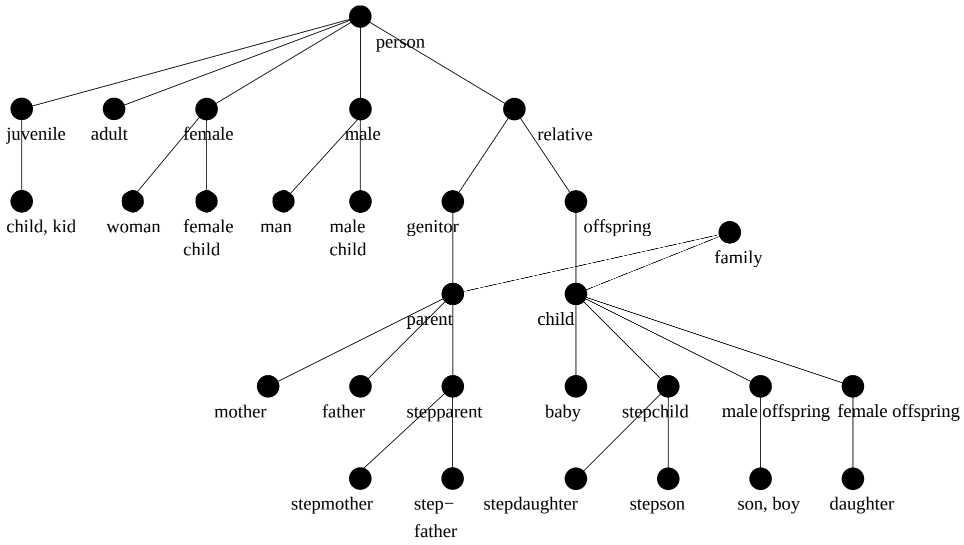

Fig. 5. Family relationships in WordNet

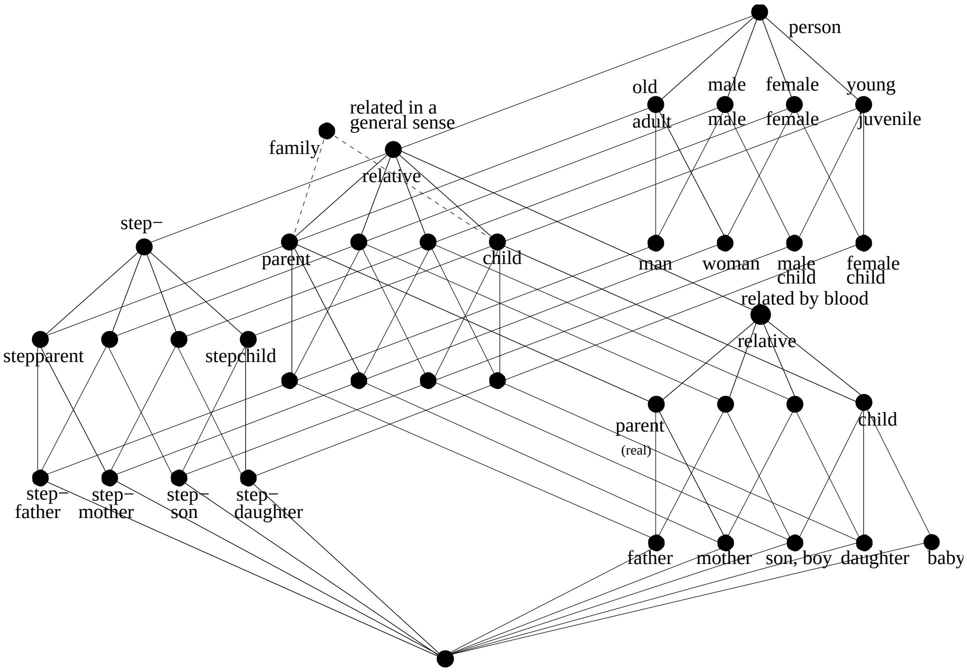

Fig. 6. Facets of family relationships

the word use in the different languages. The concept lattices help to identify the relevant differences. A similar (but more automated) approach is implemented in Janssen's (2002) SIMuLLDA tool, which is a multilingual lexical database application that uses concept lattices as an interlingua.

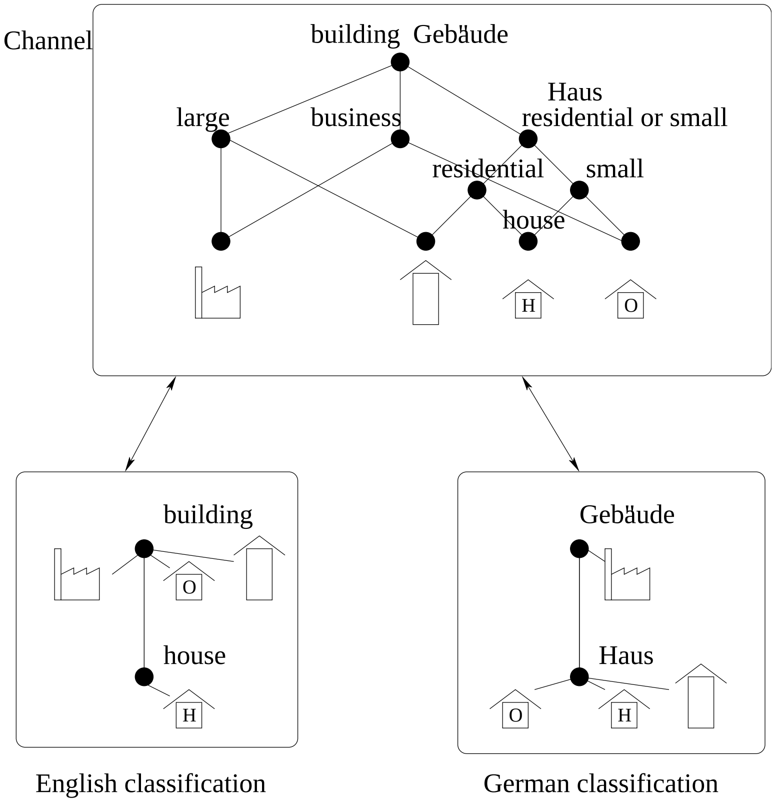

Fig. 7. A concept lattice as an interlingua

5 Conclusion

This paper argues that there are many possibilities to use formal concept analysis for linguistic applications. So far in linguistics, formal concept analysis has been mainly used for the semi-automated construction of lexical databases and for analyses of semantic relations and lexical databases. Because of the close relationship between lexical databases and ontologies, any of these applications has relevance both for linguistics and AI. Hopefully this paper might stimulate some more research into these applications. Specifically Basili's approach described in section 3.3 has a significant potential for further work.

References

1. Apresjan, J. (1973). *Regular Polysemy.* Linguistics, 142.
2. Barwise, J.; Seligman, J. (1997). *Information Flow. The Logic of Distributed Systems.* Cambridge University Press.
3. Basili, R.; Pazienza, M.; Vindigni, M. (1997). *Corpus-driven unsupervised learning of verb subcategorization frames.* AI*IA-97.
4. Basili, R.; Catizone, R.; Padro, L.; Pazienza, M. T.; Rigau, G.; Setzer, A.; Webb, N.; Zanzotto, F. (2002). *Knowledge-Based Multilingual Document Analysis.* In: Proceedings of SemaNet02.
5. Basili, R.; Catizone, R.; Pazienza, M.; Stevenson, M.; Velardi, P.; Vindigni, M.; Wilks, Y. (1998). *An empirical approach to Lexical Tuning.* In: Proceedings of the Workshop "Adapting Lexical and Corpus Resources to Sublanguages and Applications", First International Conference on Language Resources and Evaluation, Granada, Spain.
6. CYC (2001). On-line available at http://www.cyc.com.
7. Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database and Some of its Applications.* MIT press.
8. Ganter, B.; Wille, R. (1999). *Formal Concept Analysis.* Mathematical Foundations. Berlin-Heidelberg-New York: Springer, Berlin-Heidelberg.
9. Großkopf, A., (1996). *Formal concept analysis of verb paradigms in linguistics.* In: Diday; Lechevallier & Opitz (Eds.) Ordinal and Symbolic Data Analysis.
10. Großkopf, A.; Harras, G. (1999). *Begriffliche Erkundung semantischer Strukturen von Sprechaktverben.* In: Stumme & Wille (Eds.) Begriffliche Wissensverarbeitung: Methoden und Anwendungen.
11. Janssen, M. (2002). *SIMuLLDA. A Multilingual Lexical Database Application using a Structured Interlingua.* PhD Thesis, Universiteit Utrecht.
12. Kipke, U.; Wille, R. (1987). *Formale Begriffsanalyse erläutert an einem Wortfeld.* LDV–Forum, 5.
13. Old, L. J., (1996). *Synonymy and Word Equivalence.* In: Proceedings of the Midwest Artificial Intelligence and Cognitive Science Society Conference (MAICS96), Bloomington, IN.
14. Old, L. J.; Priss, U. (2001). *Metaphor and Information Flow.* In: Proceedings of the 12th Midwest Artificial Intelligence and Cognitive Science Conference, p. 99-104.
15. Osswald R.; Petersen, W. (2002). *Induction of Classifications from Linguistic Data.* In: Proceedings of the ECAI-Workshop on Advances in Formal Concept Analysis for Knowledge Discovery in Databases.
16. Petersen, W. (2001). *A Set-Theoretical Approach for the Induction of Inheritance Hierarchies.* Electronic Notes in Theoretical Computer Science 51.
17. Pollard, C.; Sag, I. (1994). *Head-Driven Phrase Structure Grammar.* CSLI Lecture Notes Series, Chicago.
18. Priss, U. (1996). *Classification of Meronymy by Methods of Relational Concept Analysis.* In: Proceedings of the 1996 Midwest Artificial Intelligence Conference, Bloomington, Indiana.
19. Priss, U. (1998). *The Formalization of WordNet by Methods of Relational Concept Analysis.* In: Fellbaum, Christiane (Ed.), WordNet: An Electronic Lexical Database and Some of its Applications, MIT press, 1998, p. 179-196.
20. Priss, U. (1999). *Efficient Implementation of Semantic Relations in Lexical Databases.* Computational Intelligence, Vol. 15, 1, p. 79-87.
21. Priss, U. (2002). *Associative and Formal Concepts.* In: Priss; Corbett; Angelova (eds.), Conceptual Structures: Integration and Interfaces. Proceedings of the 10th International Conference on Conceptual Structures, Springer Verlag, LNAI 2393, p. 354-368.
22. Pustejovsky, J. (1991). *The Generative Lexicon.* Computational Linguistics, 17, 4, p. 409-441.

23. Sedelow, S.; Sedelow, W. (1993). *The Concept concept.* Proceedings of the Fifth International Conference on Computing and Information, Sudbury, Ontario, Canada, p. 339-343.
24. Sowa, J. (1984). *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley, Reading, MA.
25. Sporleder, C. (2002). *A Galois Lattice based Approach to Lexical Inheritance Learning.* ECAI Workshop on ML and NLP for Ontology Engineering.
26. Wille, R. (1993). *The Formalization of Roget's International Thesaurus.* Unpublished manuscript.