Old, L. J. (2003).
An analysis of semantic overlap among English prepositions in Roget's Thesaurus.
In P. Saint-Dizier (Ed.),
Proceedings of the Association for Computational Linguistics SIG Semantics Conference
(ACL-SIGSEM) (pp. 13-19). Toulouse: IRIT.

An Analysis of Semantic Overlap among English Prepositions in Roget's Thesaurus.

L. John Old
School of Computing, Napier University
10 Colinton Road, Edinburgh, EH10 5DT
E-mail: j.old(at)napier.ac.uk

Abstract. Using prepositions, related words from other parts of speech, and senses listed in Roget's Thesaurus, this paper discusses and illustrates the complex relationships between and among prepositions and other basic parts of speech. The pattern of genus and differentiae emerging from the complex relationships between words and senses suggests that prepositions cannot be viewed in isolation, and that a natural, and even optimal, organization of semantics exists that may explain why current methods of classification and partitioning of words and senses sometimes result in confusion.

Keywords: Preposition, word class, part of speech, genus and differentiae, lattice, Formal Concept Analysis, Roget's Thesaurus

1. Introduction

Prepositions, as a class of words, have been referred to as a closed set. The "set" is the set of words that are eligible to be called prepositions. It is closed probably as a consequence of the fact that the words defined as (or classed as) prepositions describe a limited set of concepts (for example spatial and temporal relations) that don't change--unless our consensual reality changes.

Prepositions are not, on the other hand, a stable set. The semantics of individual prepositions is mutable across time, and among related languages. Non-standard or idiomatic use of prepositions can become the standard, while the "correct" or traditional usage goes out of fashion. Or not… An educated Scot uses the word *outwith* (archaic to some[1]) where the average English speaker would instead use *outside of,* or *except*. *Outwith* is a perfectly good preposition and unambiguous to its users.

While an English speaker standing before a house might say that the rear garden is *beyond, to the back of* or *behind* the house, but never *after* the house; a Dutch speaker would say it is "*achter* het huis." *Achter* means "after." It has the same Indo-European language root as *after,* and has the same basic semantics[2] in both languages. Even though Dutch and English are about as close as any two languages can be without being dialects, this preposition has evolved to be used in different ways.

Words that are prepositions do not have a clear semantics even within the same language. Where a teacher speaking American English, referring to a poorly written essay, might tell a student to "do it *over,*" a British teacher would only ever say, "do it *again*."

Even prepositions commonly considered synonyms may vary or disagree in the senses they describe. *Above* can be a synonym of *over* in the sense of "higher up," but not in

[1] From Middle English, according to Webster's 3rd Edition, 1965. Though it is not in Roget's Thesaurus, *outwith the law* (illegal), is.

[2] It is still acceptable English to say, "Take the first turn right *after* the set of lights;" or "After you :)"—but we are more likely to use it in its analogous temporal form: "… after 10 o'clock;" or "… after I get up."

the sense of "across"--one may live *across* the road or *over* the road, but not *above* the road (and still mean the same thing).

Furthermore, there is considerable overlap between the set of words that are called prepositions and words from other word classes (parts of speech). Crystal [1989, p. 92] points out that word classes:

> … are not as nearly homogeneous as the theory implies. Each class has a core of words that behave identically, from a grammatical point of view. But at the "edges" of a class are the more irregular words, some of which may behave like words from other classes.

This paper offers no solutions to this apparent confusion, but attempts to illustrate it as a natural, and even desirable, feature of prepositions—and of language.

The prepositions used here are drawn from the 411 entries found in Roget's International Thesaurus [1962]. This is the "American" edition of the thesaurus. Roget's Thesaurus is used because it groups words of similar meaning together, by part of speech. WordNet [Miller et al., 1993] does the same, and contains a richer set of relations, but does not contain prepositions. The comparisons made here between prepositions and other parts of speech are limited to nouns, verbs, adjectives and adverbs.

2. Overlap among Parts of Speech

Prepositions are a small set compared to other parts of speech. While prepositions are a closed set, nouns are ever increasing as science and technology advance and new words are needed to describe new concepts. Other parts of speech are being added to as well (for example, to be "ENRONed"), though not as rapidly. Table 1 shows the word-count-by-part-of-speech for words in Roget's Thesaurus.

POS	Count	PCent
Noun	69017	57.4%
Adjective	23171	19.3%
Verb	21368	17.8%
Adverb	6346	5.3%
Preposition	411	0.3%

Table 1. Part of speech count of words in Roget's Thesaurus

Other lexicons will have different numbers but the distribution will be about the same. In Table 1 words are counted only once per part of speech. The word *line*, for example, is found as a noun entry in 20 different thesaurus senses but is counted here only once as a noun. The preposition a*fter* is just one of the 411 prepositions counted here. It is also counted once under each of the other parts of speech as its 13 senses, or entries, are spread across all five parts of speech.[3] The difference between entries and words is that an entry represents one sense-instance of a word, while word is a particular string of characters. So *after*, with 13 senses, is represented in Roget's Thesaurus by 13 entries. Four of those senses are prepositional, so the word *after* has four prepositional entries.

Approximately half of the prepositions found in Roget's Thesaurus have more than one sense and so are polysemous. Many of those words are elsewhere in the thesaurus classified under different parts of speech. In Figure 1 the percentage of overlap among parts of speech has been illustrated graphically using pie charts. In this case entries were chosen, rather than words.

Entries found classified under only one part of speech are ignored here, as they do not contribute to the analysis of overlap between parts of speech, and also because the more than 105,000 unique entries in this category (of the total 200,000 thesaurus entries) would make the overlapping entries for smaller parts of speech, invisible. So *betwixt*, for example, which occurs only as a preposition, is ignored. *After*, which occurs in all five word classes, is included in the calculations for all five pie charts.

The arrows serve as a rough indicator of the main allegiance owed by a word class to another word class. For example, verbs and nouns share a high percentage of words (77% and 87% respectively), indicated by a thick, double-headed arrow; 47% of prepositions are also adverbs (indicated by a thick arrow) and

[3] *After* is found as a synonym of *afternoon* and *evening* in one nominal thesaurus sense.

32% are also adjectives (indicated by a narrower double-lined arrow); and 57% of adverbs are also adjectives (indicated by a thick arrow). The relative proportions shown here are not normalized numbers for each word class (for example there are many more nouns and verbs than prepositions), but a clear indication, at least, is present in the illustration.

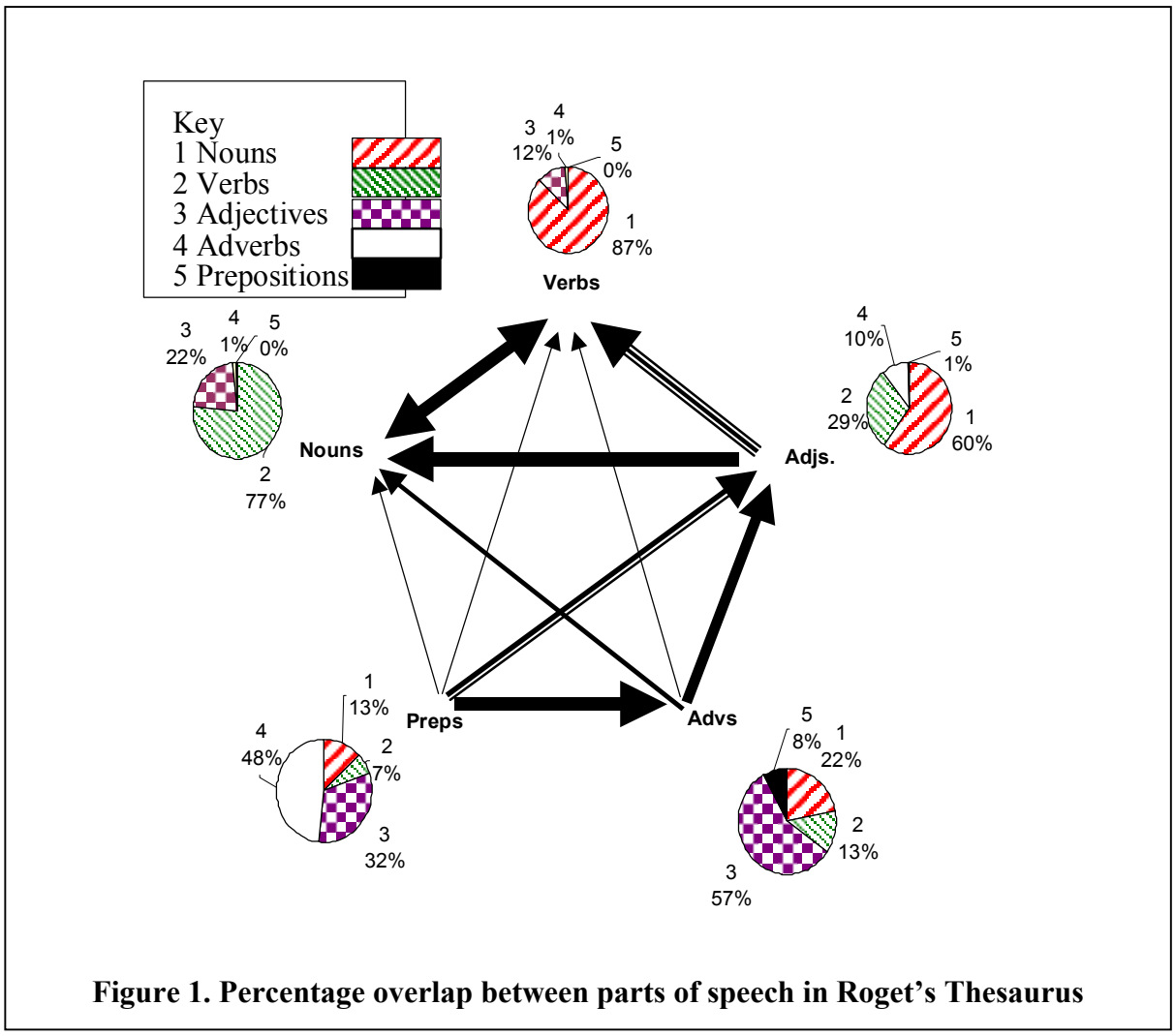

Figure 1. Percentage overlap between parts of speech in Roget's Thesaurus

Note that among the different parts of speech only adverbs (that is, 8% of adverbs that occur in other parts of speech) are also found as prepositions in any significant numbers. Those same entries constitute the 48% of entries represented on the Prepositions pie chart labeled "4 48%" (in white).

Table 2. Number of words shared between prepositions and other parts of speech.

POS	Overlap
Adverb	137
Adjective	87
Noun	33
Conjunction	18
Verb	12

In real numbers, 287 words classified as prepositions in Roget's Thesaurus are also found in senses other than those classed as prepositional. For example, 33 of these words also occur as nouns[4]. Table 2 shows the actual overlap in terms of word-counts (including conjunctions). These overlaps are formed with 198 of the 411 prepositions. There are a further 213 prepositions that do not overlap with any other part of speech.

[4] An *over* (Nn) is a cricket term for a period of play—but that sense is not included in this American edition of Roget's Thesaurus. Examples of verbs are "to *further* a cause" and "to *near* a conclusion."

3. Part of Speech Overlap for the Preposition *Over*

In Figure 2 the overlap between prepositions that occur as synonyms of *over* in various senses with various parts of speech can be seen represented as a "concept lattice" [Wille, 1982]. This forms a kind of topology of *over,* its senses, and the word that are found accompanying it in those senses--its synonyms. The lattice includes only "shared" synonyms of *over*--those words that occur with *over* in more than one sense. As with Figure 1, the words that have been omitted occur in only one part of speech and do not contribute to the connectivity or overlap between parts of speech, or senses. They would however differentiate or discriminate senses which otherwise contain identical sets of words. This is discussed further under the Section, *Genus and Differentiae*, below.

is a lattice, not a tree. The nodes/circles are called *concepts* and are labeled above by the index numbers of the senses and below by words found in those senses. Index numbers are of the form:

Category#:Paragraph#:Sense.

Though a concept is defined as the set of all of its attributes (words) and all of its objects (senses), for economy of representation words and index numbers are used as labels only once. Words label the lowest concept in which they occur and index numbers label the highest concept in which they occur. Thus a lattice is a partial ordering, where concepts higher in the lattice structure are labeled by senses that contain more synonyms, and concepts lower in the lattice are labeled by senses that contain fewer synonyms. Symmetrically, concepts lower in the lattice are labeled by words that have

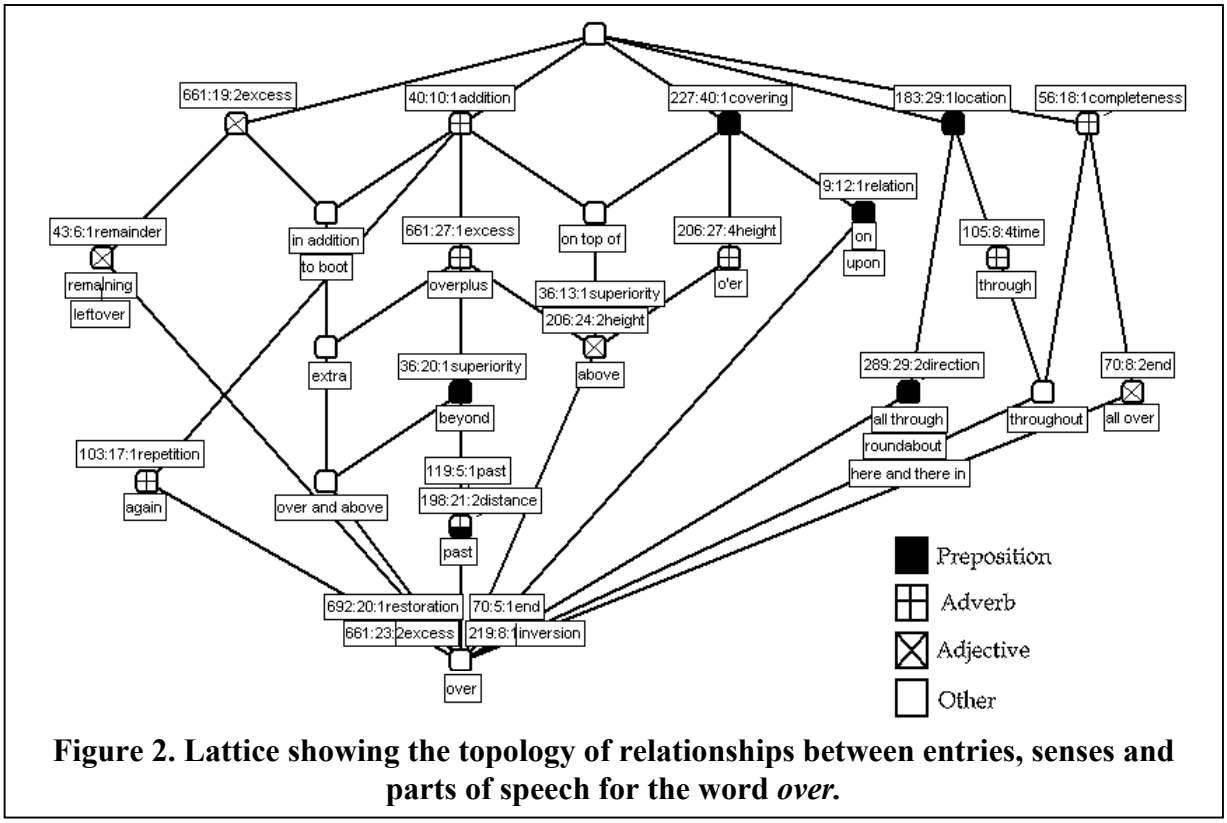

Figure 2. Lattice showing the topology of relationships between entries, senses and parts of speech for the word *over.*

A concept lattice is generated automatically from a relation between two sets, *objects* and *attributes*. In this example the *objects* are words from Roget's Thesaurus while their *attributes* are the senses of the words. A polysemous word can occur in more than one sense (as several entries) and a sense can contain more than one word—hence the graph structure formed

more senses, and concepts higher in the lattice are labeled by words that have fewer senses. No information is lost through this method of labeling only once per word and once per sense--the complete sets of senses and words can be read from the lattice as illustrated in the following examples.

Senses are read off the concept lattice top down. To the top and right of the centre of

the lattice can be seen sense 227.40.1, a prepositional sense from Category 227, *Covering.* This sense of *over* contains the following set of entries that share more than one sense with *over*: {*on top of, on, upon, above, over, o'er*}. These entries can be found on the lattice by following the lines (or links) down from the *Covering* concept, as follows: the concept below and to the left is labeled with *on top of*; the concept below and to the middle is labeled with *o'er*; following the link down to the right there is a concept labeled with *upon* and *on;* and finally, the concept below and linked to both the lower-left and middle concepts (labeled with *on top* of and *o'er*), is labeled with *above.* Together these labels make up the set of shared entries, or synonyms, of *over* found in Roget's Thesaurus Category 227, Covering, Paragraph 40, Sense 1.

The four senses labeling the bottom node contain no other entries (besides *over*) that are found in more than one sense of *over.* The top node is unlabeled as there is no sense which contains all of the words.

To find the senses of a particular word the lattice is read from the bottom up. So for example the word *over*, which is found in all senses, labels the lowest concept--all of the senses of *over* can be found by tracing the lines up (and conversely, all of the senses can be seen to contain the word *over* by tracing the lines down from them).

Above has six senses shared with *over*, {36.13.1; 206.24.2; 206.27.4; 227.40.1; 661:27:1; 40:10:1}, three of which are adverbial, one of which is prepositional, and two of which are adjectival. These can be identified and read off the lattice by tracing the lines up from the concept that is labeled with *above.*

The *scope* of the concept labeled with *above*, reading the lattice upwards, is the set of six senses of *above*; while the scope of the same concept, reading downwards, is the set of words that are contained as synonyms in the two senses that label that concept (*over* and *above*). In Formal Concept Analysis [Wille, 1989] the set of objects (the set of words) is called the *extent* of a concept; and the set of attributes (the set of senses), the *intent* of the concept.

It is not necessary to navigate the lattice expertly or understand the underlying mathematical formalism. Simply comparing adjacent concepts should convince the reader that this automatically-derived graphic has presented the senses of *over* in a coherent way—a way which supports Brugman and Lakoff's [1988] assertion that senses of a word are related and that there are gradual transitions, or *transformations*, as one navigates from closely to more distantly related senses. Similar lattices can be derived for any word in Roget's Thesaurus that has senses crossing part of speech boundaries. Figure 3 shows the concept lattice of *above*—also restricted to synonyms that occur in more than one sense. Six of the seven senses are shared with *over* (c.f. Figure 2). The seventh sense differentiates *above* from *over* in this lattice.

The automatically constructed lattices show that many closely related adjectives, adverbs and prepositions may be selected by focusing on a single word, and illustrate the overlap and blending among parts of speech, and among some words. These words are examples of the type described by Crystal [1987] as being at the "edges" of the word classes. They are the glue that ties the senses together, and incidentally, some of the most common (polysemous and high-frequency-usage) words in the thesaurus.

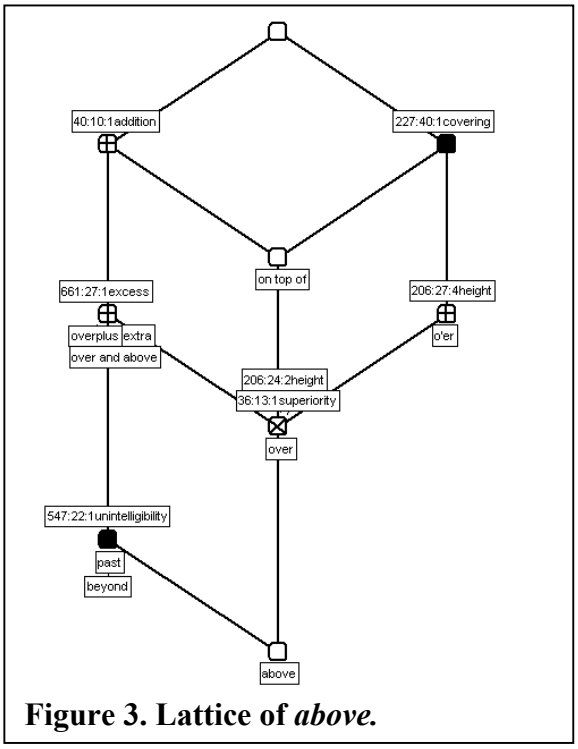

Figure 3. Lattice of *above.*

4. Genus and Differentiae

In contrast to Brugman and Lakoff's "radial category" of senses, there is no central sense evident in the lattice. None-the-less, the sense with index 40.10.1 from Category 40 *Addition,* an adverbial sense, shares words with many of the other senses. In the thesaurus it has 37 entries. Of the 37, 24 are words that have more than one part of speech, and 31 are polysemous. Of those with more than one part of speech, 14 double as adjectives, 12 double as prepositions, 4 as nouns[5], and 3 as verbs. Of the remaining "idiosyncratic" words (single-instance words, omitted from this lattice), *additionally, moreover,* and *furthermore* occur in the thesaurus only in this sense—they characterize it, differentiating it from other senses. They are the stripes that separate this tiger from other big cats—they distinguish this sense from other senses.

This sense, along with its idiosyncratic words, and relationships to other senses via those shared words, hints at what is at the core of prepositional semantics, it illustrates the concept of *genus and differentiae* used to construct sense-definitions in dictionaries. A simplified dictionary example would be: "A cup is a type of container (genus) that has a handle (differentia number one) and is used for drinking liquids (differentia number two)." As stated earlier, Figure 2 includes only those words that share more than one sense with *over*. The words that do not share more than one sense with *over* include the differentiating entries in each of its senses. So the lattice is a kind of "genus" topology, only. The missing words are what facilitate the discrimination of senses from one another in the same way that distinguishing features allow us to recognize and differentiate individual people, and living things are differentiated amongst in biological taxonomies.

Moreover, there is a symmetric organization among the words. In the same way that senses can be read down the lattice (their constituent words identified), and words can be read up the lattice (their various senses can be identified), some senses act as differentiators for words and some words act in a "genus" capacity, gluing the senses together.

Perhaps this "genus-differentiae" facet of word-sense organization has implications for the conceptual organization of the brain, but it is beyond the scope of this paper to enlarge on that. It is sufficient to say that the organization seen in the lattice emerges naturally from the data—from the semantic relationships between synonyms, and from the transitional or transformational connections between senses of polysemous words. This organization provides a natural way to arrange information in a fairly optimal fashion--so that the pieces of information become neither isolated, nor too densely packed.

5. Conclusion

A preposition is a word (or phrase). But in Roget's Thesaurus that specific word may be represented by many entries under separate prepositional senses. The same word, or string of characters (excluding homographs), may also have one or more entries classified under other, non-prepositional parts of speech. So, to say that

[5] In uses such as: the *more* the merrier; a blast from the *past*; a movie *extra;* a real *plus.*

over is a preposition is not to exclude it from being any other part of speech. Also, to say that *over* is a synonym of *above* is not to say that it is a synonym of *above* in all senses or, for that matter, for all parts of speech. To say a word "means" something, or "is" a preposition, is misleading. Outside of usage (spoken or written context), the meaning of a word can only be understood in the context of the semantics of all of its senses, synonyms, and parts of speech, together. Despite this apparently overwhelming complexity, senses of words, in context, can be disambiguated[6] almost instantaneously by native speakers. It may not be "despite of," but "because of" this complexity that we are able to do it.

References

Brugman C. and Lakoff, G., (1988). Cognitive Topology and Lexical Networks. In Small, S. I., Cottrell, G. W., Tanenhaus, M. K. (Eds.), *Lexical Ambiguity Resolution*, Morgan Kaufmann.

Crystal, David (1987). *The Cambridge Encyclopedia of Language,* Cambridge University Press, Cambridge.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., and Tengi, R. (1993). Five Papers on WordNet, *Technical Report*, Princeton University, Princeton, N.J.

Roget's International Thesaurus, 3rd Edition, Berry, L., (Ed.) Thomas Crowell Co., New York, 1962.

Webster's Third New International Dictionary (unabridged), Gove, P. B. (Ed.), G & C Merriam, Publishers, Springfield, 1965.

Wille, R., (1982). Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. In I. Rival, (Ed.), *Ordered Sets,* Reidel, Dordrecht-Boston, pp. 445-470.

Wille, R. (1989). Geometric Representation of Concept Lattices. In Opitz, O. (Ed.), *Conceptual and Numerical Analysis of Data,* Springer-Verlag, Berlin-Heidelberg.

[6] And if not immediately disambiguated, at least identified as congruent with the current context.