

Translations as Semantic Mirrors: From Parallel Corpus to Wordnet¹

Helge Dyvik

Section for Linguistic Studies
University of Bergen

Abstract

The paper reports from the project «From Parallel Corpus to Wordnet» at the University of Bergen (2001 - 2004), which explores a method for deriving wordnet relations such as synonymy and hyponymy from data extracted from parallel corpora. Assumptions behind the method are that semantically closely related words ought to have strongly overlapping sets of translations, and words with wide meanings ought to have a higher number of translations than words with narrow meanings. Furthermore, if a word *a* is a hyponym of a word *b* (such as *tasty* of *good*, for example), then the possible translations of *a* ought to be a subset of the possible translations of *b*.

Based on assumptions like these a set of definitions are formulated, defining semantic concepts like, e.g., 'synonymy', 'hyponymy', 'ambiguity' and 'semantic field' in translational terms. The definitions are implemented in a computer program which takes words with their sets of translations from the corpus as input and performs the following calculations: (1) On the basis of the input different senses of each word are identified. (2) The senses are grouped in semantic fields based on overlapping sets of translations, such overlap being assumed to indicate semantic relatedness. (3) On the basis of the structure of a semantic field a set of features is assigned to each individual sense in it, coding its relations to other senses in the field. (4) Based on intersections and inclusions among these feature sets a semilattice is calculated with the senses as nodes. According to our hypothesis, hyponymy/hyperonymy, near-synonymy and other semantic relations among the senses now appear through dominance and other relations among the nodes in the semilattice. Thus, the semilattice is supposed to contain some of the semantic information we want to represent in wordnets. (5) In accordance with this assumption, thesaurus-like entries for words are generated from the information in the semilattice.

In the project these assumptions are tested against data from the English-Norwegian parallel corpus ENPC (Johansson (1997)).

1. Introduction

1.1 Translations as semantic data

Parallel corpora, in which original texts are aligned with their translations into another language, are a rich source of semantic information. Translations come about when translators evaluate the degree of interpretational equivalence between linguistic expressions in specific contexts. In many ways such evaluations, made without any theoretical concerns in mind, seem more reliable as sources of semantic information than the careful paraphrases of the semanticist or the meaning descriptions of the lexicographer. Assuming that this is the case, can we then retrieve some of the semantic properties of expressions by going «backwards» from the network of translational relations in situated texts? Can we reconstruct semantic properties from the translational properties manifested in a parallel corpus?

The idea that semantic information can be gleaned from multilingual data has been explored by others. Resnik and Yarowsky (1997), discussing word sense disambiguation, suggest that in distinguishing between senses it may be fruitful to restrict attention to such distinctions as are lexicalised differently in other languages. Nancy Ide has explored the connections between semantics and translation in several papers; in Ide & al. (2002) the authors study versions of the same novel in seven languages and attempt to identify subsenses of words by considering how the translations of a given word cluster in the six other texts.

1.2 Wordnets and Thesauri

The output of the method presented here is an informational structure containing some of the information which we find in wordnets. A wordnet is a semantically structured lexical database. The Princeton WordNet (Fellbaum 1998), which has been built manually, distinguishes

between the senses of words and groups senses across words into ‘synsets’ according to near-synonymy. Pointers between such synsets express semantic relations like hypero- and hyponymy, antonymy, and holo- and meronymy. Wordnets for various European languages were developed within the project Eurowordnet (<http://www.illc.uva.nl/EuroWordNet/>).

Wordnets are important resources for many applications within language technology. They can be used in meaning-based information retrieval (searching for concepts rather than specific word forms), in logical inference (if a document mentions dogs, a wordnet allows the inference that it is about animals), in word sense disambiguation (providing the search space of alternative meanings), etc.

A related kind of semantic resource is the thesaurus. As an example we may consider the entry for the adjective *conspicuous* in the Merriam-Webster Collegiate Thesaurus (<http://www.m-w.com/home.htm>), where two senses are distinguished, each with its own sets of synonyms, antonyms etc.:

Entry Word: **conspicuous**
 Function: *adjective*
 Text: 1
Synonyms CLEAR 5, apparent, distinct, evident, manifest, obvious, open-and-shut, openhanded, patent, plain
 2
Synonyms NOTICEABLE, arresting, arrestive, marked, outstanding, pointed, prominent, remarkable, salient, striking
Related Word celebrated, eminent, illustrious; showy
Contrasted Words common, everyday, ordinary; covert, secret; concealed, hidden
Antonyms inconspicuous

We may compare this with the thesaurus-like entry for *conspicuous* below, which has been generated automatically from parallel corpus data by the method to be described in this paper:

conspicuous
Sense 1
 (Norwegian: avstikkende.)
Sense 2
Hyperonyms: great, hard, large.
Subsense (i) (Norwegian: synlig, tydelig.)
Near-synonyms:
 clear, conclusive, definite, distinct, distinctive, obvious, plain, substantial, unmistakable, vivid.
Hyponyms: apparent, evident, pervasive, visible.
Subsense (ii) (Norwegian: fremtredende, kraftig, sterk, stor.)
Near-synonyms: outstanding, primary.
Subsense (iii) (Norwegian: oppsiktsvekkende.)
Near-synonyms: amazing, spectacular, startling, surprising, unusual.

Antonyms and contrasted words are not included in the latter entry, since the method only allows the derivation of relations of semantic similarity (synonymy, hyperonymy and hyponymy) from the parallel corpus data. The entry displays a major division into two senses (of which the first one in this case has no information associated with it apart from a Norwegian translation), and furthermore a division into subsenses within the more informative second sense. «Sense 1» in this example is probably a spurious consequence of sparsity of data in the corpus. A better example of a major division into senses – although even there we would have liked sense 1 to have been merged with sense 4 – is provided by the following automatically derived entry for the Norwegian noun *rett*, which is contrastively ambiguous between a number of senses, among which we find ‘course in a meal’ and ‘court of law’. Some of the related words listed in this entry are surprising, while most of them are to the point:

rett N
Sense 1
 (English: course.)
Sense 2
 (English: court, justification.)
Near-synonyms:
 argument, begrunnelse, berettigelse, domstolsbehandling, gård,

gårds plass, plass, sak, ting.

Sense 3

Subsense (i) (English: option.)

Hyponyms: tilbud.

Subsense (ii) (English: rightN.)

Hyponyms: adgang, rettighet.

Subsense (iii) (English: order.)

Near-synonyms:

bestemmelse, klasse, krav, lov, løsning, måte, orden, regel, regelverk, stand, system, vedtak.

Sense 4

(English: dish, food, supper.)

Near-synonyms:

aftens, aftensmat, fat, føde, gryte, kar, kopp, kosthold, kveldsmat, lunsj, mat, matvare, middag, måltid, næring, skål, tallerken.

1.3 Semantic lattices

The thesaurus entries above are generated from *semantic lattices*, which in their turn are derived automatically from the translational data. Figure 1 below is an example of such a lattice, representing the semantic field associated with sense 4 of ‘rett’ in the entry above (labelled *rettN2* in the lattice):

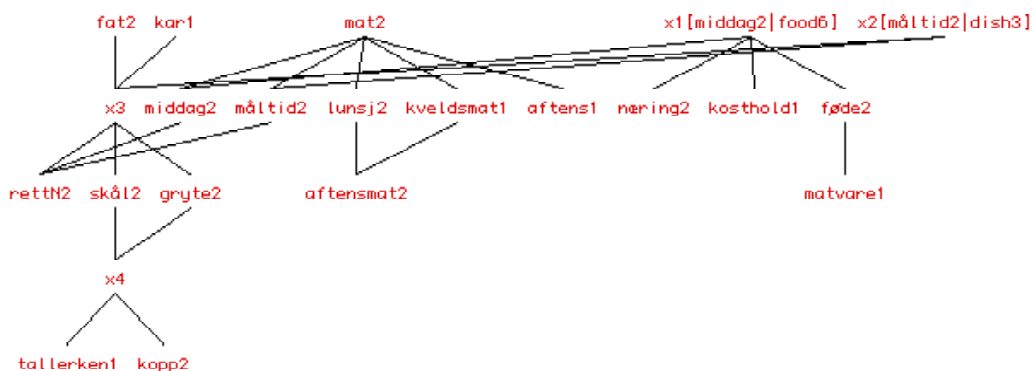


Fig 1 A semantic lattice

According to the hypothesis behind the method, senses on dominating nodes are hyperonyms of senses on dominated nodes. Thus, a sense of *mat* ‘food’ dominates senses of *rett* ‘dish’, *middag* ‘dinner’, *måltid* ‘meal’, *lunsj* ‘lunch’, *kveldsmat* ‘supper’, *aftensmat* ‘supper’, and *aftens* ‘supper’, all of which are plausible hyponyms of *mat*. Less convincingly, *lunsj* also dominates *aftensmat*.

Formally the lattice expresses inclusion and overlap relations among sets of translationally derived features, as described in section 2.3 below.

1.4 The Parallel Corpus

The English-Norwegian Parallel Corpus (ENPC), from which the above results are derived, comprises approximately 2.6 million words, originals and translations included. The corpus contains fiction as well as non-fiction and English originals translated into Norwegian as well as the other way around. The corpus is aligned at sentence level (Johansson & al. 1996), while it is a part of our present project to align the ENPC at word level, in order to be able to extract the sets of translations of a given word automatically. Our present data has been derived from the sentence-aligned corpus, however, which means that the translational data for each word in our data set has been extracted manually.

For example, searching for the Norwegian word form *bemerkelsesverdige* returns the sentences containing *bemerkelsesverdige* coupled with the corresponding English sentences in the parallel text (translation or original). Based on a set of heuristic criteria to decide whether a

word can be said to «correspond» to a given word in the translation or not, the set of translations of *bemerkelsesverdig* is extracted by the human analyser:

(bemerkelsesverdig (amazing notable remarkable spectacular surprising))

Sets of such lemmas with their associated sets of translations from the corpus constitute the input to the procedure deriving semantic lattices and thesaurus entries, by principles which we now proceed to describe.

2. The Method ‘Semantic Mirrors’

2.1 Separation of senses

We assume that contrastive ambiguity, such as the ambiguity between the two unrelated senses of the English noun *bank* – ‘money institution’ and ‘riverside’ – tends to be a historically accidental and idiosyncratic property of individual words. That is, we don't expect to find instances of the same contrastive ambiguity replicated by other words in the language or by words in other languages. Furthermore, we don't expect words with unrelated meanings to share translations into another language, except in cases where the shared word is contrastively ambiguous between the two meanings. By the first assumption there should then be at most one such shared word.

Given these assumptions contrastive ambiguity should be discoverable in the patterns of translational relations. We may consider the Norwegian noun *tak*, contrastively ambiguous between the meanings ‘roof’ and ‘grip’. Figure 2 shows the first *t*-image of *tak* in the right-hand box, and the first *t*-images of each of those English words again in the left-hand box. We refer to the last-mentioned set of sets as the inverse *t*-image of *tak*.

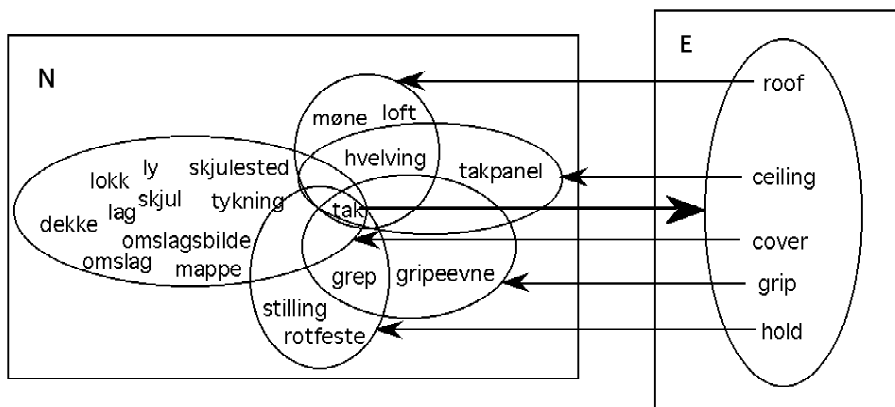


Fig 2 The first and inverse *t*-images of *tak*.

The point worth noticing is that the images of *roof* and *ceiling* overlap in *hvelving* in addition to *tak*, while the images of *grip* and *hold* overlap in *grep* in addition to *tak*. This indicates that *roof* and *ceiling* are semantically related, and similarly *grip* and *hold*, while no overlap (apart from *tak*) unites *grip/hold* and *roof/ceiling*. *Grip/hold* and *roof/ceiling* hence seem to represent unrelated meanings, and the conclusion is that *tak* is ambiguous.

The overlap patterns are necessarily preserved within the first *t*-image of *tak* when we make our third movement and find all the first *t*-images in English of the words in the inverse *t*-image. We refer to this set of sets as the second *t*-image of *tak*:

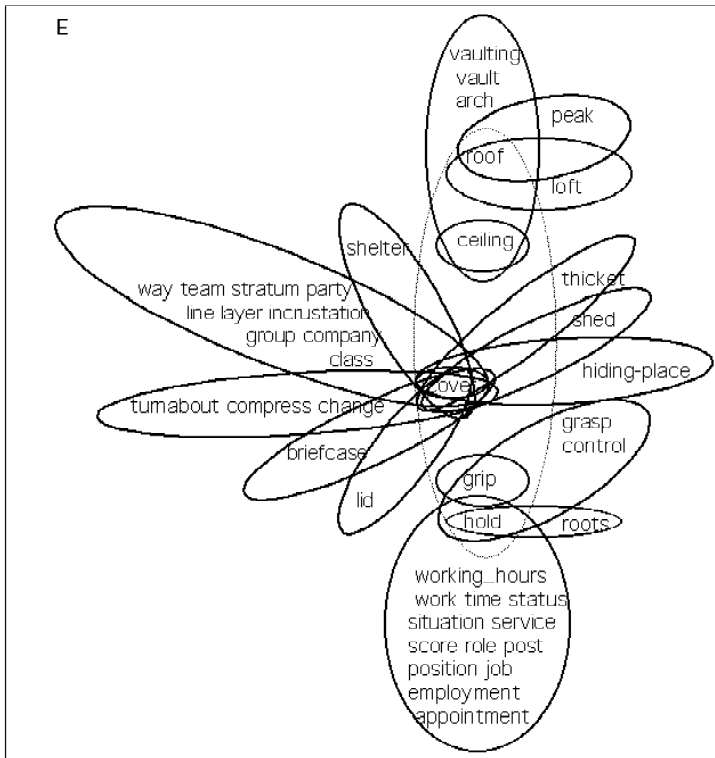


Fig 3 The second *t*-image of *tak*

As shown in figure 3, the second *t*-image can be divided into three clusters or groups of sets, each group being held together by overlap relations (we only consider overlaps in the restriction of the second *t*-image to the members of the first *t*-image). On the basis of these groups the first *t*-image of *tak* can be partitioned into three ‘sense partitions’:

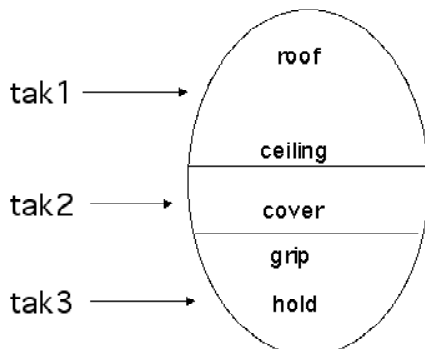


Fig 4 The sense partitions of *tak*'s first *t*-image

By this method the main senses of lemmas are individuated.

The size limitations of the corpus are a source of error: a translation *t* of *a* occurring only once in the corpus, or only occurring translationally related to *a*, will give rise to a separate sense partition only containing *t*, and hence give rise to a potentially spurious sense of *a* (cp. the doubtful «sense 1» of the examples *conspicuous* and *rett* in Section 1.2). A larger corpus might display more alternative translations of *t*, and thereby include *t* in one of the other sense partitions. A frequency filter excluding *hapax legomena* from consideration might reduce this problem.

2.2 Semantic fields

Once senses are individuated in the manner described, they can be grouped into *semantic fields*. Traditionally, a semantic field is a set of senses that are directly or indirectly related to each other by a relation of semantic closeness.

In our translational approach, the semantic fields are isolated on the basis of overlaps among the first *t*-images of the senses. Since we treat translational correspondence as a symmetric relation (disregarding the direction of translation), we get paired semantic fields in the two languages involved, each field assigning a subset structure to the other. Figure 5 gives a rough illustration of the principle (arrows indicate the *t*-image of each sense – for simplicity, the indicated sets are just suggested and in no way reflect the corpus data accurately):

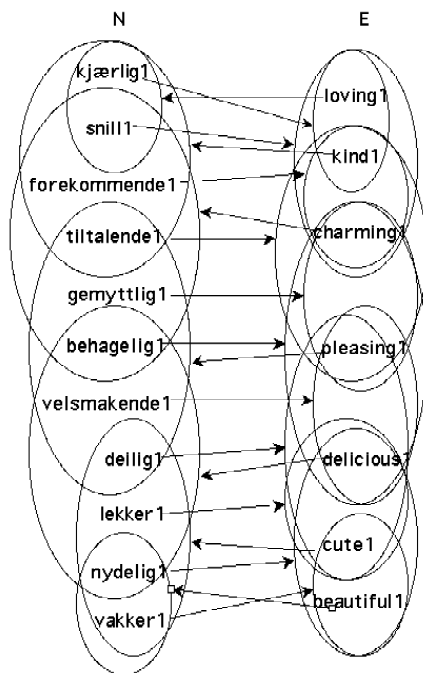


Fig 5 Paired semantic fields (simplified illustration)

The subset structure of a semantic field, assigned by its partner field in the other language, contains rich information about the semantic relations among its members. For example, senses with a wide meaning (such as *good*) will in general have a higher number of alternative translations than words with a narrower meaning (such as *tasty*). The number of translations is of course directly reflected in the number of subsets of which the sense is a member. Thus the senses at the «peaks» in the semantic fields will have the widest meanings.

We may illustrate this by means of a constructed and artificially simple example. Assume that we find the translational pattern illustrated in figure 6, where *hingst* 'stallion' is found translated into *animal*, *horse* and *stallion*, while *dyr* 'animal' is translated into *animal*, *horse*, *stallion*, *mare* and *dog*, etc.

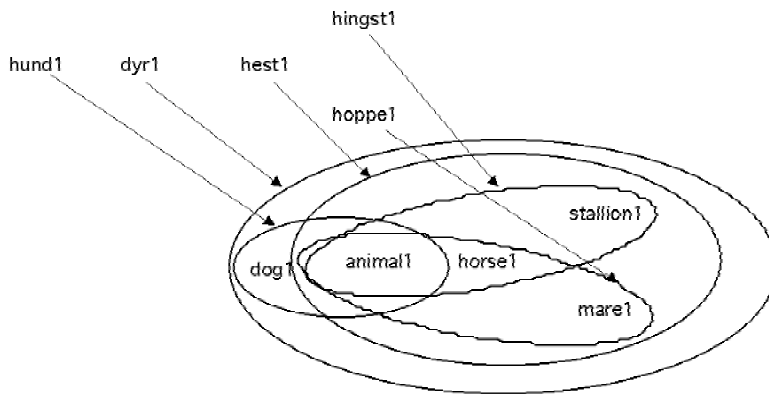


Fig 6 A constructed example

Since *animal1* is translationally related to every member of the Norwegian field, *animal1* becomes the «peak» of the English field, being a member of all the subsets, with *horse1* ranked immediately below it, etc. By symmetry, the Norwegian field gets a corresponding subset structure (cf. figure 7).

2.3 Feature assignment

The next step is to encode, for each sense, its position within the semantic field, along with its translational relations to the members of the other field. This is done by means of *feature sets*, automatically derived from the set structure. In accordance with traditional semantic componential analysis, the intention is that wide senses should have few features, while more specific senses should have more features, some of which are inherited from wider, superordinate senses. This is achieved by starting from the «tops» in two paired fields – i.e., the sense pair which is both translationally interrelated and whose members belong to the highest number of subsets – which in figure 7 gives us the pair *dyr1* and *animal1*. A feature [*dyr1|animal1*] is constructed from this pair and assigned to both its members *dyr1* and *animal1*. Then the feature is inherited (non-transitively) by «lower» senses according to the following principle: all senses in the first *t*-image of *animal1* and ranked lower than *dyr1* (i.e., belonging to fewer subsets than *dyr1*) inherit the feature, and conversely, all senses in the first *t*-image of *dyr1* and ranked lower than *animal1* inherit the feature. Then the procedure moves to the next highest, translationally interrelated, peaks *hest1* and *horse1*, constructs a feature from that pair, and assigns it according to the same principle. The result is shown in figure 7:

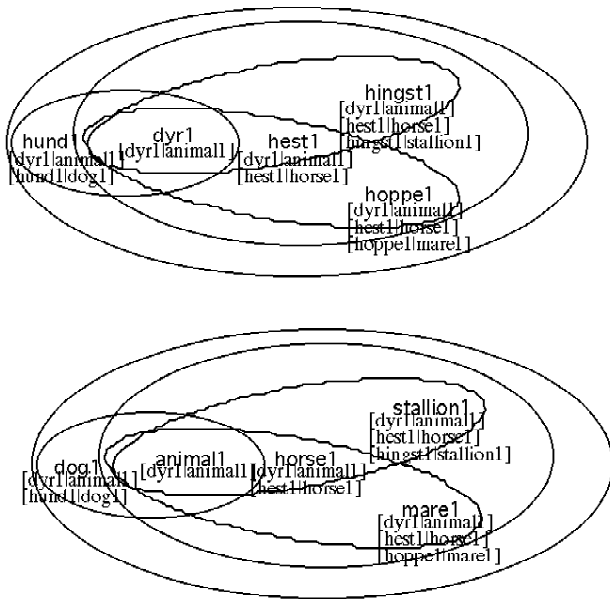


Fig 7 Feature assignment in semantic fields

The feature sets in figure 7 define a lattice based on inclusion relations among them, as shown in figure 8:

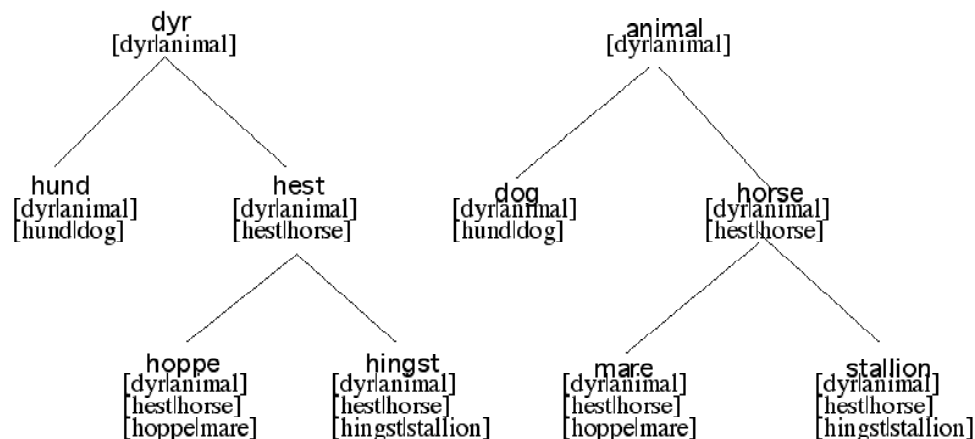


Fig 8 Lattices defined by the feature sets

In figure 8 the daughters of a node N have supersets of the feature set associated with N . In this constructed example the lattices evidently also reflect hyperonym / hyponym relations among the senses.

The lattices in figure 8 are simple trees, while actual derived lattices tend to be more complex. In the first place, senses may inherit features from more than one «peak» in the semantic field, which gives rise to multiple mothers in the lattice. In the second place, nodes may have intersecting feature sets without either of the sets including the other, so that there is no mother/daughter relationship between the nodes in question. When no actual sense is associated with the intersection, x-nodes (cp. figure 1) are introduced, carrying the intersection of the feature sets of their daughters. Thus the x-nodes can intuitively be seen as «virtual hyperonyms» of their daughters. It is the presence of x-nodes which guarantees that the structure is a semilattice (i.e., all nodes with intersecting feature sets are guaranteed to be dominated by a node carrying the intersection). In the semilattice, two senses are assumed to be more closely related the more of their features they share, i.e., the shorter the distance is to their common dominating node.

Returning now to the actual corpus-based lattice in figure 1, it is defined by the feature sets on the nodes according to the principles just described. For instance, *mat2* is associated with the singleton feature set {[*mat2*|*supper3*]}, *kveldsmat1* with {[*mat2*|*supper3*], [*kveldsmat1*|*meal1*]}, and *aftensmat2* with {[*mat2*|*supper3*], [*kveldsmat1*|*meal1*], [*lunsj2*|*meal1*], [*aftensmat2*]}. In figure 1, x-nodes with only one feature (such as *x1*) are displayed with the feature beside them.

2.4 Derivation of Thesaurus Entries

Derivation of thesaurus entries involves determining subsenses, hyperonyms, near-synonyms and hyponyms of each sense on the basis of the information in the semilattices. The semilattices are in some cases extremely complex, showing intricate networks of connections between the word senses. Much of this complexity should probably be considered as ‘noise’ resulting from accidental biases and gaps in the corpus. In the transition to a wordnet database or a thesaurus we therefore want to abstract away from much detail in the lattices, and this can obviously be done in more than one way. We presently use two parameters to regulate the generation of thesaurus entries: *OverlapThreshold* and *SynsetLimit*.

The value of the parameter *OverlapThreshold* decides the granularity of the division into subsenses in the thesaurus entry. This does not concern the division into main *senses* described above (*tak1*, *tak2*, *tak3* etc.) – those senses usually end up in different semantic fields and hence in different lattices. Division into *subsenses* is a further subdivision of each sense into related shades of meaning. We assume that there is no final and universal answer to the question of how many related subsenses a word sense has (cf. Kilgarriff 1997). By means of the parameter *OverlapThreshold* we may attune that kind of semantic granularity to our purposes.

We may illustrate the procedure by means of an example: the adjective *sweet*. Figure 9 shows a small sublattice of the large lattice including the sense *sweet1*:

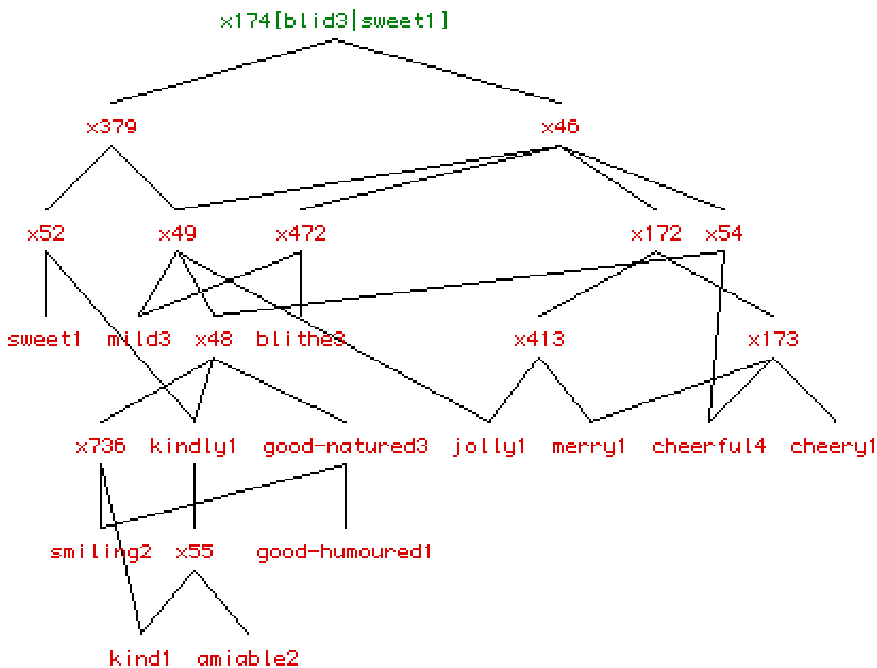


Fig 9 A sublattice containing *sweet1*

Sweet1 is also dominated by several nodes outside this sublattice; size limitations prevent displaying a more complete graph. The node *sweet1* is associated with the following feature set: {[*god3*|*good1*], [*fin2*|*nice2*], [*pen1*|*gentle3*], [*vakker1*|*soft2*], [*snill1*|*pleasant1*], [*deilig1*|*splendid3*], [*frisk4*|*sweet1*], [*blid3*|*sweet1*]}. Finding hyperonyms, near-synonyms and hyponyms of *sweet1* now first involves considering what other senses in the lattice share features with *sweet1*. The features in question are assigned to the following senses in the complete semilattice (we will refer to the sets of senses as the *denotations* of the features):

[god3|good1]:

(able1 accurate1 adept1 adequate2 affectionate1 all_right2 amiable2 appropriate5 attractive4 beautiful2 beneficial1 benign3 bright2 burning3 charming2 clean1 clear1 close3 comfortable2 comforting3 competent2 confident2 correct1 cozy2 cute1 decent2 delicious1 delightful2 detailed3 dishy1 easy1 efficient2 elegant3 excellent2 fair2 fancy1 favourable1 fine1 firmA1 first-class3 first-rate2 fit3 fortunate1 fresh3 friendly2 full2 genuine2 good1 handsome2 happy3 healthy2 high3 hot2 joyful2 kind1 kindly1 long3 lovely2 lucky2 magnificent3 marvellous1 neat2 nice2 okay1 peaceful1 perfect3 placid2 pleasant1 pleased2 pleasing1 pleasurable1 plentiful1 plenty1 polite2 positive1 pretty2 proficient1 quite_certain1 real2 reassuring2 respectable3 right2 ripe1 safe2 satisfactory1 satisfying1 secure2 sizeable1 smart2 smooth3 soft2 solid2 sound2 spectacular2 steady1 strong3 successful2 suited1 superb2 superior5 sure1 sweet1 talented2 thorough1 tidy1 well2 whole2 wholesome1 wonderful3 worthy2)

[fin2|nice2]:

(attractive4 beautiful2 breathtaking2 charming2 comfortable2 cute1 delicate3 dishy1 easy1 elegant3 enchanting1 excellent2 fancy1 fine1 first-class3 gentle3 glorious4 graceful2 handsome2 impressive2 lovely2 magnificent3 marvellous1 neat2 nice2 okay1 perfect3 pleasurable1 polite2 pretty2 pure2 slight3 smart2 soft2 splendid3 sweet1 thin2 wonderful3)

[pen1|gentle3]:

(attractive4 beautiful2 charming2 clean1 cute1 dishy1 elegant3 enchanting1 fancy1 fine1 first-class3 formal1 gentle3 graceful2 handsome2 lovely2 neat2 pleasant1 polite2 pretty2 soft2 sweet1 tidy1)

[vakker1|soft2]:

(attractive4 charming2 cute1 delightful2 dishy1 enchanting1 fair2 fancy1 graceful2 handsome2 lovely2 magnificent3 mild2 ornate2 pleasant1 pleasurable1 pretty2 soft2 sweet1)

[snill1|pleasant1]:

(all_right2 amiable2 benign3 friendly2 good-humoured1 good-natured3 jolly1 kind1 kindly1 mild3 pleasant1 pleasing1 polite2 smiling2 sweet1)

[deilig1|splendid3]:

(beautiful2 charming2 cute1 enchanting1 delicious1 delightful2 pleasureable1 splendid3 sweet1)

[frisk4|sweet1]:

(all_right2 brisk5 eager2 fit3 fresh3 healthy2 new1 pert2 sweet1 well2)

[blid3|sweet1]:

(amiable2 blithe3 cheerful4 cheery1 good-humoured1 good-natured3 jolly1 kind1 kindly1 merry1 mild3 smiling2 sweet1)

The most general features, [god3|good1], [fin2|nice2] and [pen1|gentle3], denote a high number of senses each – especially [god3|good1]. This reflects the fact that they are constructed from wide senses such as *god3* and *good1*. As a result, many of the senses carrying those features are not sufficiently close to *sweet1* to be called «near-synonyms». Therefore we do not want to consider all the senses sharing such general features as near-synonyms of each other. The value of the parameter *SynsetLimit* defines the maximal size which the set denoted by a feature can have in order to be included among the near-synonyms. With *SynsetLimit* = 20, the sets of senses denoted by [god3|good1], [fin2|nice2] and [pen1|gentle3] are not included among the near-synonyms of *sweet1* (unless they are denoted by other features as well). On the other hand, *good1*, *nice2* and *gentle3* – the English senses from which the wide features were constructed – are recorded as hyperonyms of *sweet1*.

Intuitively, the features represent different ‘aspects’ of the sense *sweet1*, and the question now is whether those ‘aspects’ are sufficiently different from each other to be considered different subsenses. Their distinctness can be measured in terms of the degree of overlap among the sets of senses they denote. If the set of features denote strongly overlapping sets of senses, the favoured conclusion is that there is no division into subsenses. On the other hand, the less the denotations of the features overlap, the more a division into subsenses is motivated. The degree of overlap in a set of sets can be measured as a value between 0 and 1, with 0 indicating no overlap and 1 full overlap (full overlap meaning that for each set *s*, every set either includes *s* or is included in *s*). In calculating the overlap degree among feature denotations we disregard the sense *sweet1* itself, since it is necessarily a member of all the feature denotations.

The value of the parameter *OverlapThreshold* is a number between 0 and 1. A feature belongs to subsense *n* if the overlap between its denotation and the denotation of at least one other feature in subsense *n* is equal to or greater than *OverlapThreshold*. Hence, the higher the *OverlapThreshold*, the more subsenses tend to be distinguished.

The two last features in the set above are constructed from *sweet1* itself, and we assume that senses sharing this feature are hyponyms of *sweet1*: They have inherited the feature from *sweet1* and must have been ranked lower in the semantic field.

Setting the parameter values with *SynsetLimit* = 20 and *OverlapThreshold* = 0.05, we consequently generate the following entry for *sweet*:

OverlapThreshold = 0.05:

sweet
Hyperonyms: gentle, good, nice.
Subsense (i) (Norwegian: frisk.)
Hyponyms: all_right, brisk, crisp, eager, fit, fresh, healthy, new, pert, well.
Subsense (ii) (Norwegian: blid, deilig, fin, god, pen, snill, s t, vakker.)
Near-synonyms:
 amiable, amused, attractive, beautiful, benign, blithe, charming, cheerful, cheery, cute, delicious, delightful, dishy, easygoing, enchanting, fair, fancy, friendly, good-humoured, good-natured, graceful, handsome, jolly, kind, kindly, lovely, magnificent, merry, mild, ornate, picturesque, pleasant, pleasing, pleasurable, polite, pretty, smiling, soft.
Hyponyms: all_right.

Subsense (ii) includes near-synonyms referring to personal character (e.g., *amiable*) as well as synonyms referring to appearance (e.g., *beautiful*). Raising the *OverlapThreshold* to 0.1 leads to the separation of those two kinds of near-synonyms:

OverlapThreshold = 0.1:

sweet
Hyperonyms: gentle, good, nice.
Subsense (i) (Norwegian: frisk.)
Hyponyms: all_right, brisk, crisp, eager, fit, fresh, healthy, new, pert, well.
Subsense (ii) (Norwegian: deilig, fin, god, pen, s t, vakker.)
Near-synonyms:
 attractive, beautiful, charming, cute, delicious, delightful, dishy, enchanting, fair, fancy, graceful, handsome, lovely, magnificent, ornate, picturesque, pleasant, pleasurable, pretty, soft.
Subsense (iii) (Norwegian: blid, snill.)
Near-synonyms:
 amiable, amused, benign, blithe, cheerful, cheery, easygoing, friendly, good-humoured, good-natured, jolly, kind, kindly, merry, mild, pleasant, pleasing, polite, smiling.
Hyponyms: all_right.

3. Conclusion

We have given an illustration of the method employed in the project «From Parallel Corpus to Wordnet». The method is implemented in a computer program taking words with their sets of translations from the parallel corpus as input and returning semantic lattices and thesaurus entries as output. The presentation has been based on examples of the results obtained on the basis of manually extracted data from the parallel corpus ENPC.

The examples have only served as illustrations and have not been subjected to a critical analysis in this paper. An important task within the project is the evaluation of the results, part of which involves comparisons with existing sources like the Princeton Wordnet and Merriam-Webster's Thesaurus. Another task is the alignment of the corpus ENPC at the word level, which will make it possible to extract lemmas with their sets of translations automatically.

Based on our results so far we feel able to conclude that the method merits further exploration.

Notes

¹ The analyses in this paper are based on corpus data resulting from work by Martha Thunes, Gunn Inger Lyse and the author. The software producing the semantic analyses has been developed by the author and reimplemented and improved by Paul Meurer.

References

- Aijmer, Karin, Bengt Altenberg, and Mats Johansson (eds.). 1996. *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies in Lund, 4-5 March 1994*, 73-85. Lund: Lund University Press.
- Diab, Mona & Philip Resnik (2002): An Unsupervised Method for Word Sense Tagging using Parallel Corpora. *40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, July, 2002.
- Dyvik, Helge (1998a): A translational basis for semantics. In: Stig Johansson and Signe Oksefjell (eds.) 1998, pp. 51-86.
- Dyvik, Helge (1998b): Translations as semantic mirrors. In *Proceedings of Workshop W13: Multilinguality in the lexicon II*, pp. 24.44, Brighton, UK. The 13th biennial European Conference on Artificial Intelligence ECAI 98.
- Fellbaum, Christiane (ed.) (1998): *WordNet. An Electronic Lexical Database*. Cambridge: The MIT Press.
- Grefenstette, Gregory (1994): *Explorations in Automatic Thesaurus Discovery*, Boston/Dordrecht/London: Kluwer.
- Hearst, Marti A. (1998): Automated Discovery of WordNet Relations. In Fellbaum (1998) pp. 131 - 151.
- Ide, Nancy (1999): Word sense disambiguation using cross-lingual information. In: *Proceedings of ACH-ALLC '99 International Humanities Computing Conference*, Charlottesville, Virginia. <http://jefferson.village.virginia.edu/ach-allc.99/proceedings>
- Ide, Nancy (1999): Parallel translations as sense discriminators. In: *SIGLEX99: Standardizing Lexical Resources, ACL99 Workshop*, College Park, Maryland, pp. 52-61.
- Ide, Nancy, Tomas Erjavec & Dan Tufis (2002): Sense Discrimination with Parallel Corpora. *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, 54-60.
- Johansson, Stig (1997): Using the English-Norwegian Parallel Corpus – a corpus for contrastive analysis and translation studies. In B. Lewandowska-Tomaszczyk and P.J. Melia (eds.), *Practical applications in language corpora*. Lodz: Lodz University. 282-296.
- Johansson, Stig, Jarle Ebeling, and Knut Hofland (1996): Coding and aligning the English-Norwegian Parallel Corpus. In: K. Aijmer, B. Altenberg, and M. Johansson (1996), 87-112.
- Johansson, Stig and Signe Oksefjell (eds.) (1998): *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi.
- Kilgarriff, Adam (1997): I don't believe in word senses. In: *Computers and the Humanities* 31 (2), pp. 91-113.

- Resnik, Philip Stuart & David Yarowsky (1997): A perspective on word sense disambiguation methods and their evaluation, position paper presented at the ACL SIGLEX Workshop on *Tagging Text with Lexical Semantics: Why, What, and How?*, held April 4-5, 1997 in Washington, D.C., USA in conjunction with ANLP-97.
- Turcato, Davide (1998): Automatically Creating Bilingual Lexicons for Machine Translation from Bilingual Text. In: *Proceedings of the 17th International Conference on Computational Linguistics (COLING-98) and of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-98)*, Montreal.