Semantic Clustering of Adjectives and Verbs Based on Syntactic Patterns

Costanza Navarretta
Center for Language Technology
Njalsgade 80 - Copenhagen
costanza@cst.ku.dk

Abstract

In this paper we show that some of the syntactic patterns in an NLP lexicon can be used to identify semantically "similar" adjectives and verbs. We define semantic similarity on the basis of parameters used in the literature to classify adjectives and verbs semantically. The semantic clusters obtained from the syntactic encodings in the lexicon are evaluated by comparing them with semantic groups in existing taxonomies. The relation between adjectival syntactic patterns and their meaning is particularly interesting, because it has not been explored in the literature as much as it is the case for the relation between verbal complements and arguments. The identification of semantic groups on the basis of the syntactic encodings in the considered NLP lexicon can also be extended to other word classes and, maybe, to other languages for which the same type of lexicon exists.

1 Introduction

The idea that the syntactic behaviour of words is connected with their meaning has been the assumption behind research in different fields such as lexical semantics and automatic clustering of words based on statistical methods. In particular much work has been done to describe the relation between the semantic characteristics of verbs and their syntactic patterns, among many Fillmore (1970) and Levin (1993), and to identify semantically similar words from large text corpora on the basis of their linguistic and distributional properties, i.a. Brown, della Pietra, de Souza, Lai & Mercer (1992), Pereira, Tishby & Lee (1993). Some research has also been done to extract the semantic meaning of adjectives on the basis of their co-occurrence with nouns, (Justeson & Katz 1993, Justeson & Katz 1995, Hatzivassiloglou & McKeown 1993, Hatzivassiloglou & McKeown 1997).

Justeson & Katz (1993) describe a method for disambiguating adjective senses by the nouns or the noun phrases they modify, using co-occurrences in large text corpora. They use statistical inference methods for organizing and analyzing the collected material. Their disambiguation method is based on the observation that certain nouns are strongly associated with some of the adjectives that modify them. For example the adjective *old* means "not-young" when combined with the noun "man",

but has the sense of "not-new" if occurring with the noun "house". Justeson and Katz disambiguate five common adjectives, *hard, old, light, right, short*, on the basis of their co-occurrence with sense-specific antonyms referring to opposite values of the same attribute (e.g. *old-new*, *old-young*).

Justeson & Katz (1995) investigate the semantic characteristics of the nouns which they used to disambiguate the five adjectives (Justeson & Katz 1993). Justeson and Katz find out that a few general semantic features such as *+/- animate*, *+/- concrete* are sufficient to characterize the disambiguating nouns. In the case of the adjective *hard* they also consider a syntactic construction in which the adjective does not modify a nominal, i.e. *it is hard/easy to do something.*

Hatzivassiloglou & McKeown (1993) describe a method for clustering adjectives semi-automatically according to their meaning in a parsed corpus as a first step towards the identification of adjectival scales. Their hypothesis is that adjectives describing the same property often modify the same set of nouns. The clustering method defined combines statistical techniques and linguistic information and relies on two similarity modules. Hatzivassiloglou and McKeown define *similarity* in terms of the distributional similarity of the adjectives in relation to the nouns they modify.

Hatzivassiloglou & McKeown (1997) identify constraints on the semantic orientation[1] of conjoined adjectives extracted from a large corpus. They combine statistical methods with morphological knowledge.

We follow the assumption that there is a connection between the syntactic behaviour and the meaning of words. Although we agree with Levin (1993) that "verb meaning is a key to verb behaviour", in this paper we go the other way round, i.e. from the syntactic behaviour of words we derive some of their semantic characteristics. In particular we have investigated to what extent it is possible to use the syntactic encodings of a corpus-based NLP lexicon to extract clusters of semantic related verbs and adjectives. Extracting semantic information from machine readable dictionaries has been the object of much research, i.a. (Vossen, Meijs & den Broeder 1989), (Wilks, Fass, Guo, McDonald, Plate & Slator 1989). Because we use an NLP lexicon, the data is already encoded in a structured way, making the extraction process straightforward. We have extracted adjectives and verbs sharing the same syntactic pattern in a corpus-based Danish NLP lexicon, the LE-PAROLE lexicon, and we have investigated to which extent the obtained clusters contained semantically "similar" elements. Because some syntactic constructions are common to a great number of adjectives and verbs, such as the simple attributive and/or predicative adjectival construction and the divalent verbal construction, these patterns cannot be used to cluster them. Instead we have extracted adjectives and verbs sharing more seldom patterns, such as adjectives subcategorizing for prepositional complements or taking expletives patterns.

Because the connection between verbal complements and verbal meaning has been widely studied, i.a. (Brent 1991, Levin 1993), the obtained clusters can be compared with semantic groups identified in the literature. Less studied is the connection between adjectival complementation and adjectival meaning[2].

In section 2 we give a definition of semantic similarity for adjectives and verbs, in 3 we

briefly introduce the LE-PAROLE Danish lexicon. In section 4 we present some examples of verbal semantic clusters extracted from the LE-PAROLE syntactic lexicon, while in 5 we give a few examples of the extracted clusters for adjectives. Finally in section 6 we propose a first evaluation of the obtained results and we make some concluding remarks.

2 A Definition of Semantic Similarity for Adjectives and Verbs

We define "similarity" of meaning for adjectives and verbs by parameters identified in the literature.

Adjectives have "similar" meaning if they are synonymous or antonymous (Miller, Beckwith, Fellbaum, Gross, Miller & Tengi 1993 (1990)) and if they belong to a linguistic scale (Hatzivassiloglou & McKeown 1993). Linguistic scales, according to the definition provided by Levinson (1983)[133], are "sets of linguistic alternates, or contrastive expressions of the same grammatical category, which can be arranged in a linear order by degree of informativeness or semantic strength". We relate adjectives belonging to the same scale, independently of their orientation, to a common "super-ordinate" concept.

Verbal linguistic scales exist, but they are not so frequent as adjectival scales, and only few verbs have "real" opposites. Thus we have extended the definition of similarity of verbs to include the *troponymy* relation. According to Miller et al. (1993 (1990)) verbs are troponyms if they are connected to a super-ordinate along more semantic dimensions. One of the most common relations holding among linguistic scales (Levinson 1983) and among many verbal troponyms (Miller et al. 1993 (1990)) is the entailment relation. In conclusion we consider verbs to be "similar" if they belong to a linguistic scale, are opposites, synonyms or troponyms.

3 The LE-PAROLE Lexicon

We have extracted adjectives and verbs using the syntactic encodings in the Danish LE-PAROLE lexicon which was produced in the EU-funded MLAP project LE-PAROLE. The Danish lexicon is one of 12 general language, monolingual electronic lexica for European languages encoded in SGML format according to a common model, the so-called PAROLE model[3]. This model guides the construction of generic NLP lexica, i.e. lexica which can be used in different applications and systems. The LE-PAROLE lexica are mainly encoded on the basis of the corpora collected by the LE-PAROLE corpus groups and the encodings in existing dictionaries.

The PAROLE model distinguishes three separate levels of description: morphology, syntax and semantics. At present the morphology and the syntax for 20,000 entries have been encoded[4]. A description of the PAROLE morphological and syn-

tactic levels can be found in (Guimier, Ogonowski & Partners 1998*a*) and (Guimier, Ogonowski & Partners. 1998*b*).

A simplified picture of the morphological and syntactic layers of the LE-PAROLE lexica can be seen in Figure 1.

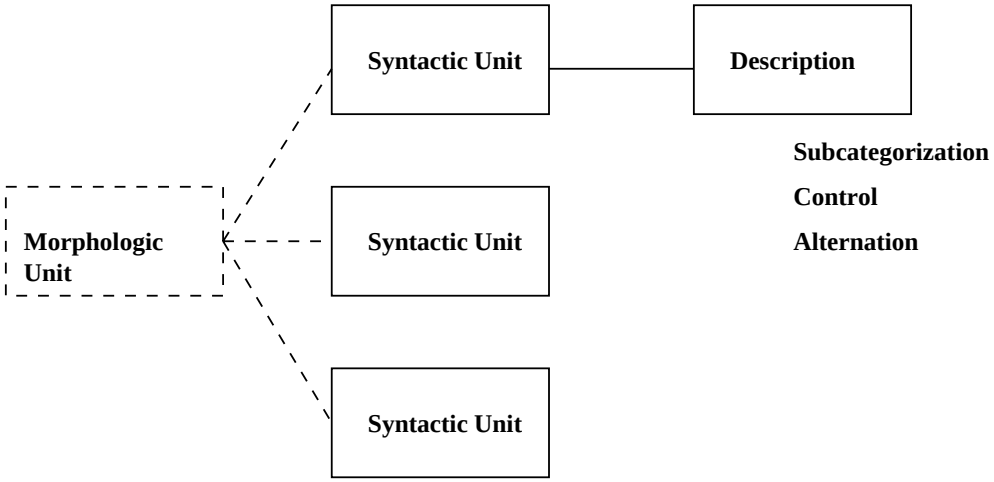

Figure 1: LE-PAROLE lexicon

The main entities of the morphological layer are Morphological Units (**MuS**) containing basic information on orthography, inflection and morphosyntactic features. One or more Syntactic Units (**SynUs**) are linked to each (**MuS**) and correspond to the syntactic patterns in which a morphological unit can occur. **SynUs** contain information about the syntactic behaviour of lexical units, such as sub-categorization, characteristics of the lexical unit when associated with a specific sub-categorization frame, control, diathesis alternations, linear order constraints. These information is encoded in the so-called **Description**. The Danish LE-PAROLE lexicon contains 20,000 morphological units. Of these units 2,816 are adjectival entries with their corresponding 3,304 syntactic units and 3,223 are verbal entries with corresponding 5,020 syntactic units.

4 Verbs

To verify the hypothesis that the syntactic encodings in an NLP lexicon can be used to extract semantically related verbs, we have looked at the syntactic patterns of verbs which belong to semantic groups recognized in the literature, in particular in (Levin 1993) and in WordNet (Miller et al. 1993 (1990)). Our study has shown that the elements of most of these groups share the same syntactic patterns (Description). Examples of verbal semantic clusters found by looking at the syntactic encodings in the Danish LE-PAROLE lexicon and the corresponding groups in other classifications are the following:

- Competition verbs (Miller et al. 1993 (1990)): *kæmpe* (battle), *fægte* (fence), *slås* (struggle), *stride* (fight), *konkurrere* (compete), *spille* (play) etc.
- Weather verbs (Miller et al. 1993 (1990)) (Levin 1993): *sne* (snow), *hagle* (hail), *regne* (rain), *blæse* (be windy) etc.
- Emotion verbs (Miller et al. 1993 (1990)) (Levin 1993): *genere* (bother), *pine* (torment), *fryde* (delight), "*bevæge*" ("move") etc.
- Verbs of Change of Possession (Levin 1993): *give* (give), *skænke* (donate), *forære* (present), *overdrage* (hand over), *testamentere* (leave by will) etc.

The verbs in each group share the same syntactic pattern, with the exception of the verbs of change of possessions which were obtained collecting verbs sharing three different descriptions. However, these descriptions are related and indicate the presence or absence of dative alternation and particular passive patterns where the second or the third complement (or both) can occur as subjects.

Emotion verbs share both a simple divalent pattern and a pattern with an expletive subject, an object and a clause as in the following examples:

Myggene generer mig
(The mosquitoes bother me)
Det generer mig at der er så mange myg
(It bothers me that there are so many mosquitoes)
Smerten piner hende
(The pain torments her)
Det piner mig at han ikke elsker mig mere
(It torments me that he does not love me any more)

Both patterns are also common to the motion verb *bevæge* used metaphorically as emotion verb:

Det bevægede ham at Maria havde husket hans fødselsdag
(It moved him that Maria had remembered his birthday)
Filmen bevægede ham dybt
(The film moved him deeply)

5 Adjectives

The patterns we have used to extract semantically related adjectives are predicative patterns where the adjectives subcategorize for prepositional phrases with nominal and clausal complements or raising constructions. The obtained groups have been checked manually and adjectives which were not semantically similar to the others have been removed. To validate the clusters we have also looked for corresponding synonyms and antonyms in WordNet. Finally we have identified common superordinates for each semantic cluster. In the following some of the obtained groups are given:

- "being afraid/not being afraid (in various degrees) of (doing) something": *bange* (afraid), *ræd* (scared), *angst* (fearful), *bekymret* (worried), *ubekymret* (carefree) ...
- "being easy/not easy (for somebody) to do something": *let* (easy), *nem* (simple), *besværlig* (troublesome), *vanskelig* (hard), *svær* (difficult)...
- "being irritated (in various degrees) at somebody": *gal* (mad), *vred* (angry), *sur* (irritated), *rasende* (raging), *forbitret* (furious)...
- "being happy/unhappy about something": *lykkelig* (happy), *ulykkelig* (unhappy)...
- "being friendly/not friendly (in various degrees) with somebody_1": *god* (kind), *sød* (nice), *venlig* (friendly), *flink* (nice), *streng* (strict), *styg* (nasty), *hård* (harsh), *modbydelig* (disgusting), *voldelig* (violent), *grusom* (cruel), *ond* (evil)...
- "being friendly/not friendly (in various degrees) with somebody_2": *god* (kind), *sød* (nice), *venlig* (friendly), *streng* (strict), *flink* (nice), *voldelig* (violent), *grusom* (cruel), *styg* (nasty), *modbydelig* (disgusting), *hård* (harsh), *ond* (evil) ...
- "being or not being capable (in various degrees) of doing something": *god* (good in the sense of capable), *snar* (quick), *fin* (good), *egnet* (fit), *flittig* (diligent), *flink* (good), *skrap* (sharp), *fortræffelig* (excellent), *enestående* (exceptional), *dygtig* (very good) *sød* (nice), *effektiv* (efficient), *langsom* (slow), *slem* (bad) ...

Although many of the elements in each group are also related by relations of synonymy, antonymy or hyponomy in WordNet, we have found more synonyms than in WordNet.

Some of the obtained groups had to be splitted up in more groups, such as the two groups "being angry (in various degrees) against somebody" and "being happy/unhappy about something" which share the same syntactic pattern. Some groups contained both related and unrelated adjectives. In the two groups "being friendly/not friendly (in various degrees) with somebody" the adjectives subcategorize for two different prepositions (*mod* (against) and *ved* (at)). We kept them separate because some Danes recognize a little semantic difference between the meaning of the adjectives in the two groups. It must be noted that the adjective *god* subcategorizing for the preposition *mod* can have two meanings depending on whether the prepositional nominal complement is animate or inanimate. In the former case the adjective belongs to the group we have identified, while in the latter case it means "effective" against something. Of course, we were not able to recognize this difference on the basis of the LE-PAROLE syntactic patterns.

6 Evaluation and Concluding Remarks

Before we evaluate the obtained results we must notice that the Danish LE-PAROLE lexicon only contains approximately 3,200 verbs and 2,800 adjectives and that only

some of the corresponding syntactic patterns have presently been encoded. The results obtained are based on this still incomplete lexicon. Although the Danish lexicon follows the common PAROLE model, the granularity chosen to identify syntactic patterns also depends on the lexicographic design chosen by the encoders. the results we have got also depend on these design choices.

our analysis of the extracted data has shown that all the groups of verbs and adjectives extracted from the LE-PAROLE Danish lexicon contain similar words. in the case of verbs all, or nearly all, the elements in the considered groups were semantically related. in few cases more "syntactic" groups formed a semantic cluster. the adjectival groups contained in some cases a few semantically unrelated elements besides the related ones and some of the adjectival syntactic groups had to be split up in different semantic clusters. The difference between verbal and adjectival behaviour is not surprising, because verbs have much richer, and thus more specialized, valency patterns than adjectives.

Although only unusual patterns, i.e. patterns which are shared by few words, can be used to identify semantically related words, and although the groups must be manually checked, we believe that the obtained results are quite interesting especially for adjectives, where the relation between syntactic pattern and meaning has not been exploited as much as it is the case for verbs. Another positive result is that we found more synonyms and antonyms than in WordNet for both verbs and adjectives.

In our opinion, semantic classifications of words must combine top-down with bottom-up strategies. Clustering words on the basis of their distributional behaviour in large corpora or their syntactic patterns in NLP corpus-based lexica is a valuable way to complement the top-down classification process. We believe also that the results obtained in our study, show that lexica with rich and well defined information as the lexica which follow the PAROLE model can be used to identify semantical related clusters and help in exploiting regularities/irregularities in the use of language.

Future work consists in extracting more groups of adjectives and verbs from the LE-PAROLE lexicon and analyzing them.The study should also be extended to complement-taking nouns and to adverbs. Because LE-PAROLE lexica, and/or NLP lexica containing the same type of syntactic information as these, exist for other European languages, the correspondence between syntactic behaviour and semantic meaning in more languages can also be investigated. The standardized encodings of the LE-PAROLE lexica offer new possibilities of analyzing alternations and other phenomena and of comparing them across different languages.

Footnotes

[1]Semantic orientation is also called polarity in the literature.

[2]Most of the proposed taxonomies for adjectives are not related to their syntactic behaviour. An exception is the taxonomy proposed in (Vendler 1963). A review of existing studies on the meaning of adjectives can be found in (Raskin & Nirenburg 1995).

[3]For a general description of the PAROLE model the reader is referred to (Calzolari 1996).

[4]The on-going European-funded project SIMPLE is in charge of encoding part of the semantic

level, i.a. (Pedersen & Keson 1999). The Danish STO project (Braasch, Christensen, Olsen & Pedersen 1998) is extending the vocabulary of the Danish LE-PAROLE lexicon to cover domain-specific words. However in this paper we exclusively work with the syntactic encodings in the LE-PAROLE lexicon.

References

Braasch, A., Christensen, A. B., Olsen, S. & Pedersen, B. (1998). A Large-scale Lexicon for Danish in the Information Society. *Proceedings from the First International Conference on Language Resources and Evaluation*, Granada, pp. 249–254.

Brent, M. (1991). Semantic Classification of Verbs from their Syntactic Contexts: An Implemented Case Study of Stativity. *Proceedings of the 5th European ACL Conference*, pp. 222–226.

Brown, P. F., della Pietra, V. J., de Souza, P. V., Lai, J. C. & Mercer, R. L. (1992). Class-based N-gram Models of Natural Language. *Computational Linguistics* **18**(4), 467–479.

Calzolari, N. (1996). PAROLE Linguistic Resources: Technical Specifications Overview. MLAP PAROLE 4, Pisa: CNR.

Fillmore, C. J. (1970). The grammar of *Hitting* and *Breaking*. R. Jacobs & P. Rosenbaum, eds. *Readings in English Transformational Grammar*. Ginn, Waltham, MA.

Guimier, E., Ogonowski, A. & Partners, P. (1998*a*). Report on the Morphological Layer. LE-PAROLE Report P-WP1.1-MEMO-ERLI-32, ELRI.

Guimier, E., Ogonowski, A. & Partners., P. (1998*b*). Report on the Syntactic Layer. LE-PAROLE Report P-WP1.1-MEMO-ERLI-33, ELRI.

Hatzivassiloglou, V. & McKeown, K. (1993). Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives according to their Meaning. *ACL Proceeding, 31st Conference*, pp. 172–182. Columbus, Ohio, USA.

Hatzivassiloglou, V. & McKeown, K. (1997). Predicting the Semantic Orientation of Adjectives. *EACL Proceeding, 8th Conference*, pp. 174–181. Madrid, Spain.

Justeson, J. & Katz, S. (1993). Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns. *Making Sense of Words, Proceedings of the 9th Conference of the UW Centre for the New OED and Text Research*, pp. 57–73. Oxford England.

Justeson, J. & Katz, S. (1995). Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns. *Computational Linguistics* **21**(1), 1–27.

Levin, B. (1993). *English Verb Classes and Alternations*. Chicago: The University of Chicago Press.

Levinson, S. C. (1983). *Pragmatics*. Cambridge, England: Cambridge University Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. & Tengi, R. (1993 (1990)). 5 Papers on Wordnet. CSL report 43, Cognitive Science Laboratory, Princeton University.

Pedersen, B.S. & Britt Keson (1999). 'SIMPLE - Semantic Information for Multifunctional Plurilingual Lexica: Some Danish Examples on Concrete Nouns'. *SIGLEX99: Standardizing Lexical Resources, Association of Computational Linguistics*. ACL99 Workshop, Maryland.

Pereira, F., Tishby, N. & Lee, L. (1993). Distributional Clustering of English Words. *Proceedings of the 31st Annual Meeting of the ACL*. pp. 183–190. Columbus, Ohio.

Raskin, V. & Nirenburg, S. (1995). *Lexical semantics of adjectives*. Technical Report MCCS-95-288, Computing Research Laboratory - New Mexico State University.

Vendler, Z. (1963). 'The Grammar of Goodness'. *The Philosophical Review*, pp. 446–465.

Vossen, P., Meijs, W. & den Broeder, M. (1989). *Computational Lexicography for Natural Language Processing*. chapter: 'Meaning and Structure in Dictionary Definitions', pp. 171–192. UK: Longman.

Wilks, Y., Fass, D., Guo, C.-M., McDonald, J., Plate, T. & Slator, B. (1989). *Computational Lexicography for Natural Language Processing*. chapter: 'A Tractable Machine Dictionary as a Resource for Computational Semantics', pp. 193–228. UK: Longman.