

An outline of Norwegian Lexical Database (LDB) and its classification of adjectives

Ruth VATVEDT FJELD, Oslo, Norway

Abstract

In the sections 1-4 this paper presents a Norwegian lexical project trying to catch up with some shortcomings in relation to modern lexicography in adding a systematic semantic component to existing bases of morphology. The project also aims to reuse consisting lexicographical products and as a spin off to improve and systematise them. The project is still in its beginning, and have so far just started the definition of adjectives. Section 5 presents the adjectiv classification in WordNet and discuss some of the questions raised in WordNet presentations of adjectives. The section concludes with an alternative adjective classification built on Norwegian and German studies, which seems to answer some of these questions.

1 Background

Historical events have given Norway two standards of written language, Bokmål and Nynorsk. As Nynorsk is the variant developed the last 150 years on the basis of the inherited Old Norse language and the modern Norwegian dialects, it has naturally been given much attention in language research and documentation of the Norwegian language.

Bokmål is an adapted version of Danish, which was the administrative language under the Danish union from the Middle Ages to 1814, and consequently it has been somewhat neglected as source and object for Norwegian linguistics, also in lexicographical description. The main resources in Norwegian lexicography are concentrated in the project Norsk Ordbok, which aims to document the Nynorsk standard. This is a big task for a small linguistic society, and maybe an explanation to the deficiencies of modern lexicographical description of the main standard Bokmål, the standard which is mostly used in administration and education in the society, and also used by approximately 80 % of the Norwegians in written language.

International projects developed in modern lexical descriptions have been very expensive. The Norwegian lexicographical community can benefit from what these projects have gained and build on their achievements. In adapting them to Norwegian we will hopefully be able to improve or refine the systems in some respects¹.

2 Goals

Norwegian Lexical Database (LDB) is a project which aims to describe the central lexicon in modern Norwegian bokmål in a systematic and partly machine readable way. This task can be formulated in the three main goals of the project:

- a) To give a systematic and formalised semantic description of 20 000 lexical units;

- b) To establish a database which can serve as a source for both traditional dictionaries; (monolingual and bilingual) and for language technology products
- c) To give a description of the Norwegian bokmål morphology actually in use.

Point a) means that the LDB-project implies a revision of existing lexical resources. It is advantageous to combine the traditional lexicographical handicraft with modern technology, and so the first step will be an analysis of the descriptions in *Bokmålsordboka*, one of the most trustworthy dictionaries of modern Norwegian. The relevant lemma information in *Bokmålsordboka* will be reinterpreted and reorganised in a way that makes it possible to be implemented in the lexical database. The lexical description shall be supplemented with explicit markings of polysemy, synonymy, antonymy, hyponymy and meronymy. In this work the Swedish dictionary *Nationalencyklopedins Ordbok* and the *Cobuild Dictionary of modern English* will be used as well. At the same time this work will be a systematic and critical revision of the existing dictionary, which has to be done anyway.

Point b) is necessary of economical reasons. The descriptions in *Bokmålsordboka* are to be automatically inserted in the field for free definitions, and these descriptions are supposed to be revised when the semantic or syntactic analysis give new insights. The examples will then also be easier to choose or change, according to the systematic description. For instance, the examples are supposed to show all the possible types of argument structure for a given lexical item, which is not systematically done in the existing version of the dictionary.

Point c) is important in a description of Norwegian, because the standard norm has several optional forms. No one knows to what extent the different variants actually are in use, and by whom. A thorough analysis of the realisation of the standard norms, and also of the forms not included in the standard norms, will provide an important tool for Norwegian language planning.

3 The dictionary basis

On the basis of the lemmas in *Bokmålsordboka* the IBM-company has made lexical full form lists, which have been further developed through the so called Documentation project as full fledged morphological bases of the standard forms in Norwegian bokmål. Our Text Laboratory has developed a morphological analysator and automatic tagger which by means of disambiguating analysis can give automatic tagging of parts of speech. The project NorKompLex at the University of Trondheim has formalised the argument structure of the verbs in *Bokmålsordboka*, and this analysis will be given a quality control and the results incorporated in the LDB-base. The results from these projects are to be included in the LDB-base.

One major deficiency in Norwegian lexicography is the lack of a good electronic lexicographical corpus. A very important basis for the LDB-project is to have a corpus consisting of modern texts. This work is under succesful progress, there has been some obstacles concerning copy-rights and other juridical and technical problems, which now seem to be solved. The corpus shall consist of 15 million words from all kinds of text types, mostly newspapers and literature, but also oral texts, unnormed texts, film texts and weekly magazines are included.

4 Lemma selection

The LDB-lexicon shall consist of 20 000 lemmas, which are considered to cover the central lexicon in modern Bokmål. The items will be chosen according to the Danish SIMPLE-lexicon of 10 000 lemmas, extended after frequency analyses of our lexical corpus. In addition, the items' number of relations to other lemmas will be considered. For instance will an adjective of low frequency be included if it is the antonym of an adjective with high frequency. The closed parts of speech, pronouns, prepositions and determiners, are to be selected in full.

5 Lexical description of adjectives

My first effort in the work with LDB will be a description of the adjectives. Adjectives are a lesser investigated part of speech in lexicography, but it is obvious that they demand a different type of definition than nouns and verbs. Adjectives are difficult to place in a semantic hierarchy, which again is necessary for a classical analytical definition build on hyponymy. Subclassifications of adjectives is also not so firmly established as for other parts of speech. The WordNet description of adjectives presented in [Fellbaum *et al.* (1993)] and [Miller (1998)] is built mainly on word associations.

The presented WN classification of adjectives might be schematized as follows:

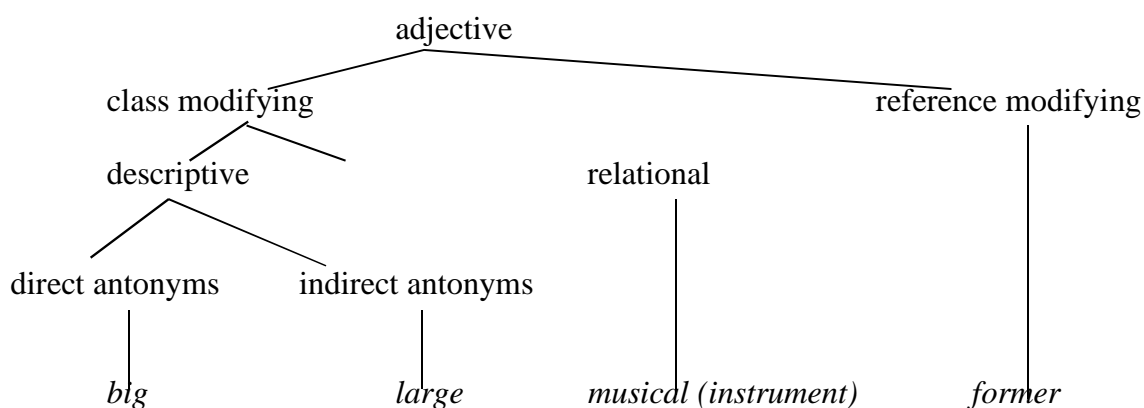


Figure 1: Adjective classification in WordNet

Class-modifying adjectives are the normal or typical adjectives, modifying the class of referents. Referent modifying adjectives point out the referents epistemological or temporal status. They are not a big class and not typical adjectives, and will give few problems in the LDB-project. Descriptive adjectives ascribe a value of an attribute to a noun. Relational adjectives are designing a class of modifiers derived from nouns, such as *dental* in *dental hygiene*, which is related to *tooth*. In Norwegian this kind of modification is usually lexicalised as *composita*, and will not be considered in our classification.

The descriptive adjectives in WordNet are divided into direct and indirect antonyms. This classification is made on the basis of word association tests, and seems disputable to me, since

there are no other criteria for how to sort the adjectives in their respective classes than individual assessments and intuitions. In accordance with my own studies of adjectives in Norwegian [Fjeld (1998)], and also according to [Bierwisch (1987)] and [Pinkal (1985)], I find this problematic. The problems are partly due to differences in lexicalisation in German/Norwegian and English, but also to deviant definitions of antonymy. In addition the WN presentation often uses vague terminology, as "closely similar in meaning", "semantic similarity", "appropriate sense", "satisfactory antonym", "related antonym" and "direct/indirect antonymy", which makes difficult to understand what is really meant.

The theoretical fundament in WN is psycholinguistics. [Fellbaum et al. 1993:27] and [Miller (1998)] claim that nearly all attributes are bipolar. The only evidence they present for this, is the following: "When a probe is a familiar adjective, the response commonly given by adult speakers is its antonym. For example, to the probe *good*, the common response is *bad*; to *bad*, the response is *good*. This mutuality of association is a salient feature of the data for adjectives." As a consequence they claim that this goes for all adjectives. But this claim gives them severe difficulties in answering the following questions:

- (1) When two adjectives have closely similar meanings, why do they not have the same antonym? For example, why do *heavy* and *weighty*, which are closely similar in meaning, have different antonyms, *light* and *weightless*, respectively?
- (2) If antonymy is so important, why do many descriptive adjectives seem to have no antonym? For example, continuing with WEIGHT, what is the antonym of *ponderous*? To the suggestion that *light* is the antonym of *ponderous*, the reply must be that the antonym of *light* (in the appropriate sense) is *heavy*. Is some different semantic relation (other than antonymy) involved in the subjective organisation of the rest of the adjectives?

They give no answers to these questions in the article from 1993, and exactly the same questions are repeated in 1998. I will therefore try to indicate some answers here and give some arguments for an alternative classification.

The first question is indeed very interesting, and has been described in several studies of adjective semantics. These studies show that there is an important difference between dimensional and evaluative adjectives, a difference which is ignored in the WN classification. Adjectives like *short* or *long* belong to a relatively closed class of adjectives. In order to be interpreted they only require a measurable norm or an average of the property that the adjective is assigning to its class. Evaluative adjectives like *good* or *reasonable*, however, demand a quality parameter recoverable from the context to be interpreted. Dimensional adjectives constitute a small and quite closed class (Bierwisch counts 17 for German, I think Norwegian has approximately 20), evaluative adjectives are innumerable.

Dimensional and evaluative adjectives have different gradability. Gradable adjectives are to be interpreted not only according to the nominal they modify (die Bezugsgrösse), but also to a class of comparison (die Vergleichsklasse). Dimensional adjectives are to be interpreted according to external properties of the nominal and a metric scale of comparison. Evaluative adjectives are subjective and unmeasurable, referring to internal and typical properties, and must be interpreted according to an unpredictable and subjective scale of measure related to its class of

comparison. These features make up what is to be known to reconstruct their quality parameter. The properties which evaluative adjectives denote are considered to be prototypical, with no inherent scales or norms for interpreting their dimensions. Their interpretation must be done relative to an ideal for the class of comparison. Recoverability of the class of comparison therefore plays a crucial role in the interpretation of evaluative adjectives. Dimensional adjectives denote attributes with physical properties. Evaluative adjectives denote attributes without physical measures, and are the expression of subjective assessments, and therefore they cannot have any established mean value or standard.

When it is difficult to find a clear antonym for adjectives like *weighty*, it is because it is an evaluative adjective, its apparent dimensionality is false. In modern Norwegian (and I think in English) this adjective is only used metaphorically, and this indicates that it denotes an evaluation, not a physical measure. This is also made clear in the semantic description in Collins Cobuild English Language Dictionary, which puts up the antonym *light* for *heavy*, but only annotates *weighty* as quality adjective, the same goes for *angry*.

If the semantics of adjectives is such as [Fellbaum *et al.* (1993)] and [Miller (1998)] seem to presuppose, the language has several adjectives to express the same dimension or quality, in other words they claim that full synonymy is normal for adjectives.

To solve their problem, the WN creators introduce a "similarity pointer" defined as follows: "The term *similar* as used here typically indicates a kind of specialization." This builds on [Gross, Fisher and Miller (1989)], who distinguish direct antonyms like *heavy/light*, which are conceptual opposites that are also lexical pairs, from indirect antonyms, like *heavy/airy*, which are conceptual opposites that are not lexically paired. Following from this, they have to operate with *head synsets* and *satellite synsets*, which represent senses that are similar to the sense of the head adjective. [Miller (1998)] further claims that many antonyms are derived through morphological rules by means of *un-*, *in-* and others. But these derivations are not antonyms, they just denote the exclusion of the quality which the underived adjective is denoting. But what an adjective does not include, is everything else than its inclusion, not only its opposition.

Question (2) follows naturally from question (1), and the answer is that similarity is a semantic relation involved in the subjective organisation of adjectives. The WN description takes it for granted that synonymy is common in adjectives, and that all adjectives has an antonym. According to [Fjeld (1998)], this is rather the exception than the rule. Normal adjectives can be divided into two classes: gradable and non-gradable. All gradable adjectives are indefinite, either because of their relativity or their "borderline indefiniteness", which means that it is impossible to decide exactly where the borderline between its plus- or minusarea goes. The most typical borderline indefinite adjectives are the colour adjectives. Such adjectives describe physical areas which are not naturally to be understood in terms of antonymy. [Fellbaum *et al.* (1993)] claim that well trained persons are able to identify the opposition of a colour, but this is also not true, *green* is not the opposition of *red*.

The simple answer to this second question is, that antonymy is not so important as the WN-classification presupposes. At least not in the classification of adjectives. According to [Lyons (1977)] only the adjectives which are gradable, have real antonyms. Lexical antonymy demands a property of bipolarity where both poles have lexical expressions. This is rarely so, and antonymy is therefore no sufficient or important criterium in the general classification of

adjectives. But yes, there is a different semantic relation involved in the organisation of adjectives, namely the difference between expressions for objective measurability and subjective assessment.

The alternative classification is solely built on semantic criteria and we find this more useful in the description of the semantics of adjectives and their abilities as nominal modifiers:

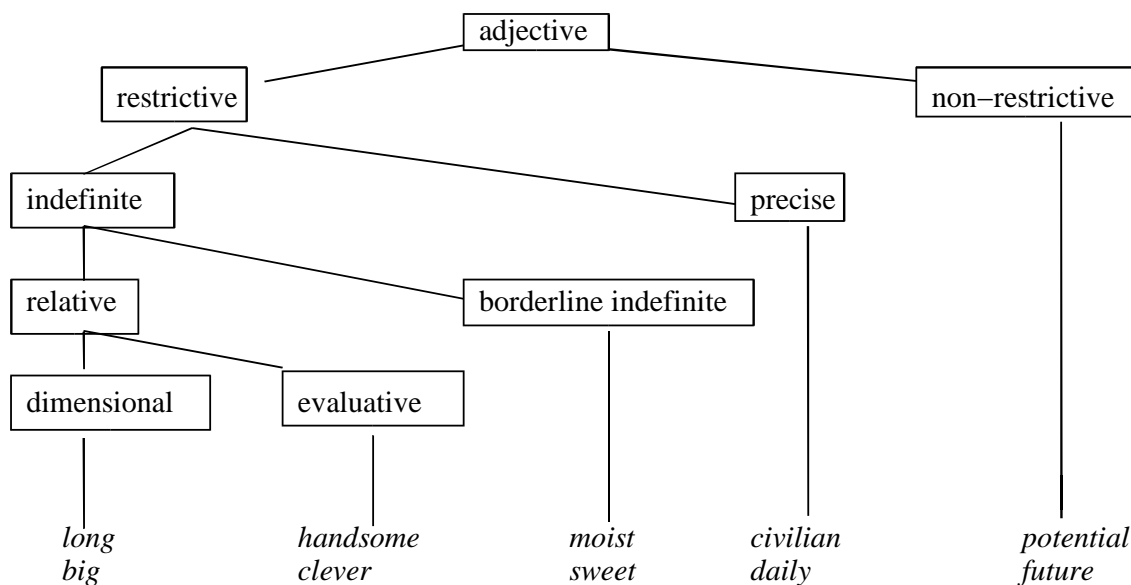


Figure 2: Alternative classification of adjectives

6 Conclusions

In this paper I have tried to answer some of the questions raised by the authors of two articles about the adjectives in WordNet. An alternative classification makes it possible to account for the lack of antonyms in the main gross of class modifying adjectives, and will also give a better fundament to account for gradability and the difference in objectivity and subjectivity in adjectives. These are important properties in the adjectives' ability to modify a noun, and is regarded as necessary in a lexical description. The Norwegian Lexical Database which is under development, will therefore use this alternative classification.

Notes

¹As Norway is not a member of the European Union, we have of course not been natural partners in the EU-initiated language technology projects. However, systematic lexical descriptions have already been developed for Swedish and Danish, and thus we have an excellent starting point for our Norwegian Lexical Database.

There are only two lexicographers working with this project, Boye Wangensteen and Ruth V. Fjeld, assisted by Anne Engø in administrating the base, and Ålov Runde in building the corpus. Associated to the project are three students writing theses about related subjects.

References

- [Bierwisch (1987)] Bierwisch, M. 1987a. "Dimensionsadjektive als strukturierender Ausschnitt des Sprachverhaltens". In: Bierwisch, M. & E. Lang. 1987. *Grammatische und konzeptuelle Aspekte von Dimensionsadjektiven*. Berlin
- [Fellbaum *et al.* (1993)] Fellbaum, C., Derek Gross and Katherine Miller. (Revised Aug. 1993). *Adjectives in WordNet*
- [Fjeld (1998)] Fjeld, R.V. (1998), *Rimelig ut fra sakens art. Om tolkning av ubestemte adjektiv i regel-givende språk*. PhD-Thesis, University of Oslo, Det historisk-filosofiske fakultet.
- [Gross, Fisher and Miller (1989)] Gross, D., Fischer, U., and Miller, G.A. 1989. "The organization of adjectival meanings." In: *Journal of Memory and Language*, 28.
- [Lyons (1977)] Lyons, J. 1977. *Semantics*. Cambridge, UK
- [Miller (1998)] Miller, K.J. 1998. "Modifiers in WordNet". In: Fellbaum, C. (ed.): *WORDNET: An electronic lexical database*. The MIT Press, Cambridge, Mass., US
- [Pinkal (1985)] Pinkal, M. 1985. *Kontext und Bedeutung*. Berlin. New York

