Feature Detection - A Tool for Unifying Dictionary Definitions

Jørg Asmussen
Det Danske Sprog- og Litteraturselskab, DSL
Society for Danish Language and Literature, DSL
Dept. for Digital Dictionaries and Text Corpora
Christians Brygge 1
DK-1219 Copenhagen K
DENMARK
ja@dsl.dk

Abstract
Based on the development of a web-based digital dictionary of contemporary Danish with integrated corpus access, *ORDNET*, this paper outlines some of the problems connected with semantically oriented search in digital dictionaries; it then presents a statistical method to extract significant indicators of differentiating features from dictionary definitions, *Feature Detection*, and discusses its potential as a tool for unifying semantically related dictionary definitions, to determine salient differentiating features, and to perform semantic grouping of related words.

1. Problem: Semantic search in digital dictionaries

On the basis of Korpus 2000[1] as well as other corpora of Danish[2] compiled by the Society for Danish Language and Literature, DSL, and on the basis of The Danish Dictionary, DDO,which is currently being published,[3] DSL has recently started developing a web-based digital dictionary of contemporary Danish with integrated corpus access, ORDNET. One of the goals of this project is to exploit the functional potential of the computer to a higher extent than is the case in many other digital versions of printed dictionaries - which are often very close to the printed source, both in appearence and functionality. As a consequence, one of the major issues of the ORDNET project is to strengthen the onomasiological potential of a dictionary significantly by implementing facilities for semantic queries.

In this paper, we examine the way word senses are defined in the printed version of the DDO and the implications of this on the possibilities for advanced semantic queries on a digital version of it. The type of semantic search we have in mind here is based upon a certain genus and the presence (or absence) of certain differentiating features, e.g. *find all words in the dictionary that denote rodents* (in general), *animals that can be kept as pets, to move quickly, young men, attractive women*, or *extremely big*.

Some digital versions of printed dictionaries provide a functionality that resembles these types of queries, e.g. *Dictionary Search* in the Longman Dictionary of Contemporary English, LDOCE, or *SmartSearch* in the Macmillan English Dictionary. A closer look at this functionality, however, reveals that the 'semantic' search that can be performed in the electronic versions of these dictionaries is not significantly more advanced than one can expect from a simple full text search where the scope is restricted to the content of the definitions only. Searching on *yellow fruit* in LDOCE gives among many examples of words denoting

fruits that are yellow also the verb *bear*, whereas the query *instrument NOT string* in Macmillan among many correct non-string instruments also gives you *balalaika* and *banjo*.

The examples reveal that there is a discrepancy between a human user's need for short, intelligible definitions avoiding redundancy and the need for logically exact definitions based on a well-defined stringent practice in a digital context. A closer investigation of definitions in a traditional dictionary often reveals that definition practice changes from one editor to another, or from one sense to another, and it may even be subject to a certain extent of stylistic idiosyncracy. This may be appealing to human users of the dictionary, but it runs counter to the wish to use the contents of the dictionary in a digital context.

2. Definitions in the DDO

Like the English dictionaries mentioned above, the digital version of the DDO, ORDNET, will provide semantically oriented query facilities based on the contents of the definitions. And also in the DDO, we find the same inconsistencies in definitions as indicated in the examples above. Compared to its predecessor, the 28-volume Dictionary of the Danish Language, ODS, which mixes at least six different types of definitions, even in one single entry,[4] the DDO is considerably more consistent regarding the way word senses are defined: the vast majority of definitions in the DDO follows the 'classical' genus-differentia principle, giving the closest hyperonym and specifying it by the necessary number of differentiating features.

2.1. Genuses in DDO definitions

Definitions were not written according to a pre-established ontological hierarchy of genuses; instead, the genus of every single definition was chosen by the responsible editor and explicitly marked up in order to be able to extract them automatically and thus make their underlying hierarchical relations explicit. Errors in the choice of appropriate genuses and inconsistencies in the emanating semantic hierarchy show that some editors did not have a clear understanding of the importance of selecting an appropriate genus and probably did not pay sufficient attention to them as they were meant to remain invisible in the printed version of the dictionary. However, the vast majority of definitions have useful genuses which can be used for extracting semantically related words in the digital version of the dictionary. Figure 1 shows a small - and fairly simple - branch of the underlying semantic hierarchy in the dictionary. The figure also shows some editorial inconsistencies (indicated by circles), e.g. *hare* ('hare') is subordinate to *mammal*, not *rodent* as one would expect, whereas *muldyr* ('mule') is directly subordinate to *dyr* ('animal'). The most superordinate concept *organisme* in the branch shown is defined as *dyr, plante eller anden levende enhed* ('animal, plant, or other living entity') where *dyr* ('animal') is marked as genus. This causes a conceptual circularity that could have been avoided at this point if *enhed* ('entity') had been marked instead. Furthermore, this extensional definition does not fully comply with the common definition practice of the DDO. A more appropriate, intensional version would have been *levende enhed som fx et dyr eller en plante* ("living entity such as an animal or a plant") which emphasizes the genus more clearly.

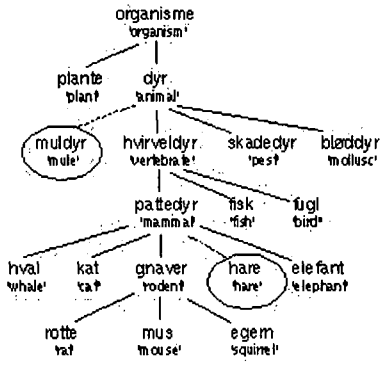

Figure 1: Part of the implicit conceptual hierarchy in the DDO

The advantage of making explicit (parts of) the underlying conceptual hierarchy of the dictionary is that it can be considered an 'empirically' established sketch of an ontology, a sketch that may highlight some of the conceptual problems concerning the design of ontologies in general, and furthermore could be used as a starting point for the elaboration of a more consistent ontology which can be used to improve the definitions in the dictionary, particularly a digital version of it.

2.2. Differentiating features in DDO definitions: The Feature Detection Method

Another issue in streamlining definitions for digital exploitation are the number and quality of differentiating features used in the definitions. Only 23 words in the DDO denote a rodent, so it would be quite straightforward to extract all definitions of rodents from the dictionary and harmonize them according to a certain definition principle.

For other genuses one will probably find typical features as well. One could run through the definitions manually and group them according to the presence or absence of certain features. But it would mean a tremendous amount of editorial work because many genuses are shared by quite large numbers of words, thus, e.g., 105 definitions use *institution*, 389 *plant*, 651 *place*, and 4382 *person* as genus. A more rational way is to get the job done automatically on a statistical basis. This is the way we consider the most consistent as it is not biased by individual human preferences.

Our method, *Feature Detection*, extracts typical, statistically significant, word forms from all definitions in a dictionary that share a certain genus. Consequently, genuses have to be marked-up throughout the dictionary, as is the case in the DDO. The extracted word forms can then be interpreted as *indicators* for significant differentiating features. The Feature Detection method basically regards all definitions in the dictionary as a corpus[5] D where the definitions based on a certain genus make up a subcorpus G. The relative frequency[6] for each type (or word form) in G, $f_r(t_G)$ is compared to that of the same type in D, $f_r(t_D)$ by computing the rounded quotient and express it as a score s: $s=round(f_r(t_G)/f_r(t_D))$. Thus, the method is conceptually closely related to the Mutual Information measure[7] which

may also be understood as comparing the frequency (or probability) of certain phenomena in two sets S1 and S_2 where S_1 is a subset of S_2. If we apply this method to the *rodent* definitions, we find among the types with highest scores the following nine: *bæverlignende* ('beaver like'), *fodsåler* ('foot soles'), *hårbeklædt* ('haircovered'), *nataktiv* ('active at night', 'nocturnal'), *nutria* ('nutria'), *prærieområder* ('prairie areas'), *Sibiriens* ('Siberia's'), *tundra* ('tundra'), *ungskov* ('young woods') - all of them quite obviously words that somehow can be semantically related to rodents. However, they seem extremely specific and not generally applicable for semantic searches in the dictionary. A closer look at their absolute frequencies reveals that the nine types quoted do not only occur just once in the *rodent* definitons but also just once in all definitions of the whole dictionary. Because of their specific character and their low frequency they do not seem useful as general semantic features in a digital version of the dictionary. It would make no sense to apply the feature *hårbeklædt* ('haircovered') or *nataktiv* ('nocturnal') to a semantic search throughout the dictionary as each of these features only matches one single word (*studsmus* ('microtine') and *hulepindsvin* ('porcupine') respectively) which as a matter of fact does not accord with the real zoological world where many animals are covered with hair or are nocturnal. Many of these top-scoring types represent editorial ad hoc compounds which could easily have been expressed in a less compressed style, e.g. *bæverlignende ('beaver-like')* > *som ligner en bæver* ('like a beaver'), *museagtig* (lit. 'mouseish') and *muselignende* ('mouse-like') > *som ligner en mus* ('like a mouse'), *hårbeklædt* ('haircovered') > *dækket af hår* ('covered with hair'), *prærieområder ('prairie areas')* > *prærier* ('prairies'). To some extent this may result in a more analytic syntactic definition style and thus in some cases somewhat longer definitions. However, the advantage is that the vocabulary and the style of the definitions become more homogenous and compounds which may be unintelligible for learners of Danish are avoided - this latter point is rather important as many of the compounds used in definitions cannot be looked up themselves in the dictionary, e.g. *ungskov*.[8] Thus, Feature Detection can be used to find utterly strange words in definitions that share a common genus.

The words with highest scores in the definitions are often those with the lowest frequencies - they are extremely specific. Like Mutual Information, Feature Detection overemphasizes the specific, low-frequency cases. But whereas we use Mutual Information to find typical collocates in corpus linguistics, we want to pinpoint these odd cases in order to get rid of them by streamlining our definitions in a digital dictionary context. However, the low-frequency cases can be suppressed by modifying Feature Detection with a frequency filter. Experimentally we have found $log_{10}(d)+1$ a useful threshold for $f_a(t_G)$ where d is the number of definitions with the genus in question. Table 1 shows all remaining types in *rodent* definitions after having applied $f_a(t_G)>log_{10}(d)+1$ to the Feature Detection method. It also shows the relative and absolute frequencies, f_a and f_r, for these types in G and D.

Type	literal equivalent	$f_r(t_G)$	$f_a(t_G)$	$f_r(t_D)$	$f_a(t_D)$	score
Ører	'ears'	12658	5	37	38	342
Hale	'tail'	30380	12	157	161	194
Pels	'fur'	17722	7	127	131	140
Gråbrun	'greyish brown'	7595	3	60	62	127
Korte	'short'	7595	3	137	141	55
Lever	'lives'	22758	9	427	439	53
Lang	'long'	40506	16	774	796	52
Forholdsvis	'relatively'	7595	3	316	325	24
Kort	'short'	10127	4	721	741	14
Lille	'little/small'	10127	4	1398	1438	7
Især	'especially'	7595	3	2146	2207	4

Table 2: Differentiating features in *rodent* definitions found by Feature Detection with filter

Many of these types can intuitively be recognized as indicators for features that in some way or other are typical for rodents, especially those located at the top of the table. At the bottom of the table, the picture gets a little more obscured with types like *især* ('especially') and *forholdsvis* ('relatively'). If we want the result of Feature Detection to focus on 'typical' feature indicators, we can introduce another filter that eliminates types with a low score. Experiments, again, indicate that for this purpose $2\cdot(\log_{10}(d)+1)$ may be a useful threshold, so if we apply the filter $s>2\cdot(\log_{10}(d)+1)$, the last type *især* ('especially') in table 1 will disappear from the output.

3. Applications of Feature Detection

3.1. Unifying definitions

Let us assume that Feature Detection applied with the mentioned filters isolates indicators of possible core features used in conjunction with a certain genus. Definitions based on the genus in question that do not use at least one of the isolated, 'typical', feature indicators may then deviate too much from an implicit definition standard that applies to the majority of the definitions with the genus in question. And the question is why these definitions deviate from the standard, and whether they can be redefined according to the implicit standard. E.g. the DDO entries *glædespige*, *bajadere*, *offentligt fruentimmer*, *massøse*, *gadepige*, *gadetøs*, *demimonde*, *kurtisane*, *hetære*, *callgirl* and *hore* all share the genus *kvinde* ('woman') and - according to Feature Detection - the high-score feature type *prostitueret* (the adjective 'prostitute') in at least one of their definitions, the recurrent definition pattern being *prostitueret kvinde som...* (lit. 'prostitute woman who...'). However, *ludder* ('bitch'), unexpectedly does not occur in this group, and does not even hold any statistically significant features after filtering has been applied to the Feature Detection method. A closer look at its definition *kvinde, som tilbyder samleje ... mod betaling* ('woman who offers sexual intercourse ... for payment') reveals that it deviates significantly from the definition pattern

used in the above examples - a case, where rewriting the definition in accordance with the tacitly established pattern should be considered.

3.2. Determining differentiating features

Feature Detection finds indicators of differentiating features that are typically used in conjunction with a certain genus. This can be used for making an inventory of differentiating features that can be used in definitions with a given genus. In a digital context, one could even make all core features obligatory. Thus, in the case of *rodent* definitions, for all rodents certain information concerning the shape and the relative size of the ears, the relative length of the tail, the quality and colour of the fur, and their habitat could be obligatorily mentioned in the definitions. The disadvantage of cramming all this info into a single definition may be that it becomes overloaded with redundancy: what if only one rodent had round ears and all others in the dictionary had pointed ears, should the default feature *pointed ears* be mentioned for every single rodent but one? An alternative way of applying these features is to put them into a more formalized version of the definition which could work as a digital counterpart to the classical human definition and could be used for semantic search. These formal definitions could probably even be used in certain NLP contexts and as a source for automatically generated 'human' definitions which could at least be used as templates for the human definition writer.

3.3. Semantic grouping of words

Finally, Feature Detection has been implemented in an ORDNET prototype as a feature for semantic grouping of words which share a certain genus. Even if the quality of this semantic grouping is dependent on the definitions used in the dictionary, i.e. to what extent they are unified and to what extent they are based upon a well-defined set of obligatory differentiating features, cf. the example given in 3.1, the results in many cases give valuable semantic information to the user.

Endnotes

[1]Cf. (Andersen et al. 2002) and (Asmussen forthcoming).

[2]Among other corpora compiled by DSL are the Corpus of The Danish Dictionary (Norling-Christensen and Asmussen 1998) and the Danish PAROLE Corpus (Keson 1998). An overview of Danish corpora is given in (Asmussen 2001).

[3]Cf. (Lorentzen 2004) and (Trap-Jensen 2004) in the proceedings.

[4]Cf. Asmussen 2003.

[5]In this context, a corpus is defined as a set of types (or word forms), each with a certain frequency.

[6]Measured in occurrences per one million tokens.

[7]Cf. Church and Hanks 1989.

[8]Even the author of this paper has no clear understanding of what this word exactly means.

References

Andersen, M.S., Asmussen, H., Asmussen, J. 2002: *The Project of Korpus 2000 Going Public.* In Braasch A. & Povlsen C. (eds.): Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen.

Asmussen, J. 2001: *Korpus 2000. Et overblik over projektets baggrund, fremgangsmåder og perspektiver.* NyS 30. Nydanske studier & almen kommunikationsteori, Copenhagen.

Asmussen, J. 2003: *Zur geplanten Retrodigitalisierung des Ordbog over det danske Sprog. Konzeption, Vorgehensweise, Perspektiven* in Burch, Fournier, Gärtner & Rapp (eds.): *Standards und Methoden der Volltextdigitalisierung. Beiträge des Internationalen Kolloquiums an der Universität Trier, 8./9. Oktober 2001.* Trier 2003 (Abhandlungen der Akademie der Wissenschaften und der Literatur, Mainz)

Asmussen, J. [forthcoming]: *Towards a methodology for corpus-based studies of linguistic change. Contrastive observations and their possible diachronic interpretations in the Korpus 2000 and Korpus 90 Corpora of Danish.* In Archer, Rayson, Wilson (eds.): *Corpus Linguistics Around the World.* Rodopi.

Church, K., Hanks, P. 1989: *Word association norms, mutual information and lexicography.* ACL Proceedings, 27th Annual Meeting, Vancouver.

Keson, B. 1998: *Documentation of The Danish Morphosyntactically Tagged PAROLE Corpus.* Society for Danish Language and Literature, DSL, Copenhagen, http://korpus.dsl.dk/e-resurser/parole-doc.rtf

Norling-Christensen O., Asmussen, J. 1998: *The Corpus of The Danish Dictionary.* In Lexikos 8, Afrilex Series 8:1998, Stellenbosch, pp. 223-242.