A Fuzzy Clustering Approach to Word Sense Discrimination

ERIK VELLDAL

This paper describes a novel approach to automatically categorize (i.e. cluster) a set of words in order to reflect their various senses and their relations of semantic similarity. We report on experiments on a set of Norwegian nouns that are represented by their co-occurrence profiles over various lexico-grammatical contexts extracted from corpora. With the purpose of capturing a notion of *typicality* the clusters themselves are construed as *fuzzy sets*, and the words are assigned varying degrees of membership with respect to the various classes. The membership functions are based on the distance between the context vectors that represent the words and the prototype vectors that represent the classes. The goal is to automatically uncover soft semantic classes, where the various memberships of a given word can be used to characterize its various senses.

Fuzzy clustering techniques have predominantly been used in such application areas of pattern recognition as *image processing* and *computer vision*. However, we argue that fuzzy clustering methods may be useful in modeling conceptual classes and word senses as well, by virtue of allowing for *multiple* and *graded* memberships (without being probabilistically constrained). This is in contrast to the *hard* classes and *crisp* memberships of conventional clustering methods that have often been used for deriving classes of (distributionally) similar words. It also contrasts with *probabilistic* approaches where the membership of a given word in a given class is relative and constrained with respect to its *other* membership values.

The categorization process has four main steps: *i*) Extracting local context features for words from corpora, *ii*) computing association scores for the word–context co-occurrences, *iii*) clustering the resulting association vectors to form a set of tight initial clusters (containing only a subset of all the words) and finally *iv*) assigning fuzzy membership values for words across the various clusters based on similarity towards class prototypes. In the next section we first briefly review the notion of fuzzy sets and describe how we will represent the semantic clusters. We then step through the different stages of the process before finally showing examples of what the resulting soft word classes look like.

Fuzzy Sets and Semantic Classes

We want the clusters that are formed to represent *meanings* in some sense, with words categorized according to their semantic content. As words are frequently seen to be homonymic, polysemous or vague, any attempt to pin down some aspect of word meaning should take these possibilities of ambiguity into account. For example, by the fact that words may have multiple meanings, our clustering model should allow objects to have *multiple memberships* across clusters. Moreover, different words can represent more or less *typical* instances of a given concept. Some words may represent clear-cut instances of a given category, while others represent peripheral or border-line cases. Correspondingly, the *boundaries* of conceptual categories are often fleeting and not precisely determined.

In order to represent the semantic categories and the associated memberships of words, we will adopt the notion of *fuzzy sets*. This construct was introduced by (Zadeh, 1965) for the purpose of describing classes that lack precisely defined criteria for membership. In contrast to classical sets, objects may "belong" to a fuzzy set with *varying degrees of membership*. We furthermore adopt a *similarity based interpretation* (Ruspini and Francesc, 1998) of fuzziness, where we let a membership value represent the degree of *typicality* or *compatibility* that a word holds toward the concept a class expresses.

A fuzzy set ζ on X is characterized by a membership function u_ζ that maps each $x_j \in X$ to a real number in the unit interval [0,1] (Zadeh, 1965). The value of $u_\zeta(x_j) = u_{\zeta j}$ represents the *grade of membership* that x_j holds in ζ, where unity corresponds to the highest degree of membership (Zadeh, 1965). By contrast, for an ordinary crisp set the two-valued characteristic function is restricted to either 1 or 0, corresponding to whether the object does or does not belong to the set. Note also that we do not impose the so-called Ruspini condition that would require the membership values for a given word to sum to 1.

To ease notation, we let u_ζ denote both the characteristic function and the set itself. Furthermore, a set of c fuzzy clusters on the set of k association vectors X is represented by a $c \times k$ partition matrix U where a component u_{ij} gives the strength of membership for x_j in the i-th class. A class u_i will itself actually be represented by a *prototype* vector v_i formed on the basis of a tight set of initial members. These prototypes may be seen to resemble the notion of *committee based centers* employed by (Pantel and Lin, 2002), and we will get back to the details of how these representations are construed later on.

The spatial metaphor underlying the vector space representations of the distributional profiles facilitates an intuitive approach to specifying the memberships as *a function of the distance* $d(x_j, v_i)$ between a word vector x_j and a

class prototype v_i. Many empirical and psychological studies of concept formation have also advocated that *similarity* should be modeled as an exponentially decaying function of distance in the representational space (see e.g. Gärdenfors, 2000). The membership function that we later describe is defined in line with this, computing the similarity-based fuzzy memberships based on distance towards cluster prototypes.

Local Context Features

Most work on distributional characterization of word similarity has been based on co-occurrences within *n*-grams, broad context windows, or even documents, without incorporating much linguistic information. However, the previous work of, among others, (Hindle, 1990), (Pereira et al., 1993), and (Pantel and Lin, 2002), clearly demonstrate the plausibility of deriving classes of semantically similar words on the basis of more "local" contextual information in the form of grammatical and syntactic relations.

The set of 3000 nouns that we analyze for this paper are characterized by way of their co-occurrences with other (lemmatized) words in various grammatical and syntactic constructions. These contexts are based on relations such as adjectival modification, prepositional modification, noun–noun modification, noun–noun conjunction, possessive modification, and various verb–argument relations. The features are extracted by an *ad hoc* shallow processing tool, Spartan[1] (Velldal, 2003), that works on top of the morpho-syntactical annotations of the Oslo-Bergen Tagger (Hagen et al., 2000). The tagged Norwegian texts that we use comprise 18.5 million words from The Oslo Corpus and 4 million words from a corpus which is still under development at the Section for Norwegian Lexicography and Dialectology at the University of Oslo (UiO). As an example of what the extracted contextual features may look like, the contexts recorded for the sentence in Example 1 are shown im Table 1 below.

Example 1: *Kunden bestilte den mest eksklusive vinen på menyen.*
(The customer ordered the most exclusive wine on the menu.)

Each noun $t_i \in T$ is represented by an *n*-dimensional feature vector $f_i = \langle f_{i1},\ldots,f_{in}\rangle$. We will use $F = \{f_1,\ldots,f_k\}$ to denote the set of *k* feature vectors that represent the nouns in T. The value of an element f_{ij} is the observed co-occurrence frequency for t_i and the *j*th contextual feature of a set $C = \{c_1,\ldots,c_n\}$, where C consists of the *n* most frequent contextual features resulting from the shallow processing step (with $n = 1000$ for the results reported

1. Shallow PARsing of TAgged Norwegian text (Velldal, 2003)

Target noun	Feature
kunde (customer)	SUBJ_OF bestille (order)
vin (wine)	OBJ_OF *bestille* (order)
vin (wine)	ADJ_MOD_BY *eksklusiv* (exclusive)
vin (wine)	PP_MOD_BY meny (menu)
meny (menu)	PP_MOD_OF vin (wine)

Table 1: Context features of nouns in Example 1.

here). We did not implement any additional feature selection other than this simple frequency-based approach, but this also seems like a more acute problem when using more broadly defined and crude contexts such as windows, than when dealing with these localized linguistic contexts. Intuitively, the property of occurring somewhere within a 100 words distance from the word-form *drink* is a lot less semantically focused than the property of occurring as the direct object of the verb with the same form

Association Weighting

Since raw frequency alone is not normally regarded a good indicator of relevance, co-occurrence counts are usually weighted with some measure of association strength, typically in the form of a statistical test of dependence. Let A be such a weighting function that maps each element f_{ij} of the feature vectors in F to a real value. We will use X to denote the resulting set of association vectors where each $x_i = \langle A(f_{i1}),\ldots, A(f_{in})\rangle$. In other words, the salience score of the contextual feature c_j for the noun represented by f_i is then given by $x_{ij} = A(f_{ij})$.

In this paper we take A to be based on the *log odds ratio*, $\log \theta$, as used in the semantic space experiments of (Lowe and McDonald, 2000). The odds ratio θ gives the ratio of the odds for some event to occur, where the odds themselves are also a ratio. Given a local context c, the odds of finding t rather than some other noun can be stated as $P(c,t) / P(c,\neg t)$ (where P is the maximum likelihood estimate based on the relative frequencies in F). Given any other context than c instead, the chance of seeing t rather than some other noun, is $P(\neg c,t) / P(\neg c,\neg t)$. Finally, the ratio of these two odds indicates how much the chance of seeing t increases in the event of c being present:

$$(1) \qquad \theta = \frac{P(c,t)/P(c,\neg t)}{P(\neg c,t)/P(\neg c,\neg t)} = \frac{F(c,t)F(\neg c,\neg t)}{F(c,\neg t)F(\neg c,t)}$$

Taking the natural logarithm of the odds ratio makes the score symmetric around 0, with 0 being the neutral value that indicates independence (Lowe and McDonald, 2000). If $\log \theta(c,t) > 0$, then the probability of seeing t increases when c is present. Note that we here assume all unobserved or negatively correlated co-occurrence pairs (c,t) to have zero association, both be-

cause this relation is of less interest for the task at hand and because of the problem of getting reliable estimates from the sparse corpus data.

Clustering the Nouns

The simple clustering scheme that we apply to the noun data consists of the following four stages: We first define a set of (hard) initial clusters B through a phase of standard *bottom-up* (*agglomerative) clustering* using the *within-groups average method* (WGAC), using what we call the *singletons ratio* to define a stopping condition. The resulting partition tree is then pruned before we compute a set of association weighted prototypes V' for the resulting set of hard clusters B defined on a subset of X. In the final pass we compute the partition matrix U where each u_{ij} is given by a function of the distance between the word vector x_j and the corresponding cluster center v'_i. This final step can be seen as a "fuzzified" version of a simple nearest prototype classifier. Instead of doing a crisp 1-NP classification, we let the soft classification of each $x_j \in X$ be a function of its distance to each class prototype $v'_j \in V'$.

Agglomerative Clustering

A pseudo-code outline of the general agglomerative algorithm is given in Table 2. One of the defining properties of different instances of this general algorithm is the way one chooses to compute the similarity between collections of objects (i.e. the clusters). When plugging in the WGAC method, the similarity of two clusters b_h and b_i is computed as the average pairwise similarities within their union. The within-group average similarity of a cluster $b_j \subseteq X$ is defined as

$$W(b_j) = \frac{1}{|b_j|(|b_j|-1)} \sum_{y \in b_j} \sum_{y \neq z \in b_j} s(y,z) \qquad (2)$$

With respect to the general procedure shown in Table 2, $sim(b_h, b_i)$ is computed as $W(b_h \cup b_i)$. We furthermore use the cosine measure to compute the similarity between the individual cluster members $s(y,z)$ (which for the length normalized association vectors in X simply is the dot product).

In order to secure good initial prototypes we apply the WGAC method to the *entire* noun set X, but with a cut-off for the *ratio* of objects merged, or, equivalently, a threshold for the number of singleton root nodes. With respect to the general outline in Table 2, we use a termination condition Λ that indicates if the ratio of singleton clusters in B is above a specified threshold $\varrho \in [0,1]$. Given a function *singletons* defined as

$\text{singletons}(B) = \{b \mid b \in B \wedge |b| = 1\}$, and a threshold ϱ, we define our stopping criterion Λ for the agglomerative algorithm in Table 2 as

$$(3) \qquad \Lambda(B) = \begin{cases} T, & \text{if } |\text{singletons}(B)| \geq k\rho \\ F, & \text{otherwise} \end{cases}$$

This means that we only need to perform a maximum of (k ϱ)-1 mergers and a minimum of (k ϱ)/2. The greedy WGAC method is guaranteed to produce a monotonic sequence of partitions, and the rationale behind the singleton ratio criterion is to ensure that only the objects that show the strongest degree of similarity are clustered. When forming the initial prototypes we thereby only rely on the most confident merging decisions. This is in contrast to the often-used Buckshot strategy that relies on a random sample of size $\sqrt{(ck)}$, clustered until $|B| = c$ (Cutting et al., 1992). Note also that the singleton ratio does not specify the number of classes c directly, as we do not know *a priori* the branching structure of the partition tree. This means that although a cut-off is employed, the number of clusters c is not specified in advance. The final number of clusters, however, is also determined by the pruning of the resulting partition tree. For the noun clustering we used ϱ = 0.5.

Parameters:
$X = \{x_1, \ldots, x_k\}$
sim: $P(X) \times P(X) \rightarrow \Re$
stopping criterion Λ
pruning function *prune*: $P(X) \rightarrow P(X)$

---

for $i = 1$ to k do
    $b_i \leftarrow \{x_i\}$
$B \leftarrow \{b_1, \ldots, b_k\}$
$j \leftarrow k + 1$

while $\Lambda(B)$ do
    $(b_h, b_i) \leftarrow \underset{(b_m, b_n) \in B \times B}{\arg\max} \{\text{sim}(b_m, b_n)\}$
    $b_j \leftarrow b_h \cup b_i$
    $B \leftarrow \left(B \setminus \{b_h, b_i\}\right) \cup \{b_j\}$
    $j \leftarrow j + 1$

$B \leftarrow prune(B)$

return B

Table 2: Agglomerative Clustering

Pruning the Partition Tree

In order to further secure the distinctiveness of the prototypes, a pruning procedure is applied to the noun clusters B resulting from the bottom-up run. After first discarding all singletons, the *prune* function then recursively merges all clusters that are *reciprocal nearest neighbors* (RNN) with a within-groups average similarity greater than a specified threshold δ. The purpose of this merging step is similar in spirit to the way

committees are defined in (Pantel and Lin, 2002), and is an attempt to ensure that the remaining clusters B are well scattered in the space, and to reduce the chances of discovering duplicate senses. In the final step of the pruning, we discard all remaining groups $b \in B$ for which $|b| < \sigma$, as these smallest clusters are less likely to yield good and representative prototypes. Note that the elements of the discarded groups are not reassigned to other clusters during the pruning, since this would dilute the final prototypes. After the initial clustering of the 3000 nouns, this merging and trimming (with $\delta = 0.35$ and $\sigma = 3$) leaves us with a set of $c = 167$ hard clusters $B = \{b_1,\ldots,b_c\}$ that includes roughly one third of the words in the initial data set.

Computing the Prototypes and the Partition Matrix

Based on the set of tight initial clusters obtained so far, we now compute the association-weighted prototypes in order to finally define the membership matrix U. As shown in Table 3, the first step of computing the class prototypes is to compute a set of class–context co-occurrence vectors V, analogous to the feature vectors in F for individual words.

Parameters:
Frequency vectors F
Association measure A
Clusters $B = \{b_1,\ldots,b_c\}$
Distance function d: $\Re^n \times \Re^n \rightarrow \Re$
Sensitivity weight w

---

for all $b_i \in B$ do
$\quad v_i \leftarrow \sum_{x_j \in b_i} f_j$
$V \leftarrow \{v_1,\ldots,v_c\}$
for all $v_i \in V$ do
$\quad v'_i \leftarrow \langle A(v_1),\cdots,A(v_n)\rangle$
$V' \leftarrow \{v'_1,\ldots,v'_c\}$
ensure $\|v'_i\| = 1 \quad \forall\, v'_i \in V'$

---

for all $u_{ij} \in U$ do
$\quad u_{ij} \leftarrow \exp\left(-\frac{d(x_j,v_i)^2}{w}\right)$
return (U, V')

Table 3: Fuzzy Prototype Classification

Each vector $v_i \in V$ is the sum of the frequency vectors $f_j \in F$ that correspond to the elements $x_j \in b_i$. This effectively means that each corpus occurrence of one of the clustered words is counted as an occurrence of the corresponding cluster type.

The next natural step then is to perform the same association weighting as we did for word vectors above. A set of association vectors V' is constructed by applying the weight function A (based on log θ) to each element of the class co-occurrence vectors. If desired we can now easily also check what it is that ties the members of a given cluster together, by simply inspecting the context features sorted according to their association scores (see (Velldal, 2003) for examples). Note also that, although the possible effects of differences in occurrence frequencies are reduced by the association-weighting, we also normalize the vectors to have unit length.

The final step of the categorization process is to assign the fuzzy membership values of all the words across the clusters. The memberships function itself is defined in Equation 4.

$$\mathrm{u}_i(x_j) = \exp\left(-\frac{\mathrm{d}(x_j, v_i)^2}{w}\right) \tag{4}$$

Decreasing the weight parameter w results in a more rapid decay of the function. Using this membership function we perform a single-pass assignment of word memberships computing a 167 × 3000 partition matrix U (see the procedural outline in Table 3).

Results and Discussion

We have described a fuzzy clustering approach to unsupervised acquisition of soft semantic classes with the purpose of modeling senses for a set of Norwegian nouns. Words and classes are represented on the basis of their lexical-syntactic environment in text, and a fuzzy clustering method assigns multiple and graded memberships to words across the constructed classes.

Some examples of clusters are shown in Tables 6 to 8 (with English translations in parenthesis). Each example shows a target noun and the four clusters in which it has its strongest degree of membership. The clusters themselves are represented by their ten most "typical" members (which might or might not include the target word) together with the associated membership values.

Many of the strongest clusters for the various target words seem very encouraging, and many of the classes themselves appear to be highly coherent. Unfortunately, however, we are not able to include any systematic quantitative evaluation in this paper. In order to assess the quality of automatically derived word classes, one needs to compare the results against some sort of gold standard, but no broad-coverage repository of semantic information for Norwegian exists as yet. Moreover, some important unresolved issues also remain, such as the possibility of delineating the number of senses for each word on an individual basis and by a more principled means than just relying on a globally specified similarity threshold. If our aim was a crisp clustering, we would simply assign every word uniquely to the class in which it holds the strongest degree of membership, thus obliterating the need for any threshold. When dealing with a fully fuzzy partition, on the other hand, we might need to determine (for some practical

purposes at least) to what degree a given word must be associated with a given class u_i, in order for u_i to be included among its senses. We soon run into trouble if we define a single such threshold to apply for all words and classes. To illustrate the problem, consider the two highest ranking clusters for the nouns *hest* (horse) and *gris* (pig) shown in Tables 4 and 5 below.

The two most salient clusters for *hest* (horse):

Cluster 154, membership: 0.5746		**Cluster 62, membership: 0.4791**	
bil (car)	0.9711	fugl (bird)	0.8558
bile (?, Def Sg/Pl = *bil*)	0.9611	hund (dog)	0.8330
buss (bus)	0.7988	katt (cat)	0.7990
busse (?, Def Sg/Pl = *buss*)	0.7617	katte (cat)	0.7660
båt (boat)	0.7248	slange (snake)	0.6261
tog (train)	0.6735	slang (slang, Def Sg = *slange*)	0.6039
drosje (taxi)	0.6212	mann (man)	0.5556
fly (airplane)	0.6152	dame (woman)	0.5293
hest (horse)	0.5746	dyr (animal)	0.4998
trikk (tram)	0.5635	gutt (boy)	0.4810

Table 4: Cluster memberships of *hest* (horse)

The two most salient clusters for *gris* (pig):

Cluster 62, membership: 0.2507		**Cluster 116, membership: 0.2433**	
fugl (bird)	0.8558	fisk (fish)	0.8008
hund (dog)	0.8330	brød (bread)	0.7990
katt (cat)	0.7990	kjøtt (meat)	0.7939
katte (cat)	0.7660	kak (?)	0.6599
slange (snake)	0.6261	kake (cake)	0.6429
slang (slang, Def Sg/Pl = *slange*)	0.6039	pølse (sausage)	0.5663
mann (man)	0.5556	bolle (bun, bread roll, bowl)	0.5413
dame (woman)	0.5293	melk (milk)	0.5153
dyr (animal)	0.4998	mat (food)	0.4821
gutt (boy)	0.4810	vin (wine)	0.4648

Table 5: Cluster memberships of *gris* (pig)

The two highest ranked sense classes for the noun *hest* (horse) (i.e. clusters 154 and 62), seem quite appropriate and can be seen to correspond to its *vehicle* and *animal* sense respectively. However, classes with a lower rank than these two, that have associated memberships less than $u_{(62)(horse)} = 0.48$, seem a lot less appropriate. A reasonable threshold in the case of *hest* (horse)

then might be 0.45, blocking every sense class with a membership value that falls below this limit. However, with this cut-off, none of the 2 nearest prototypes of the noun *gris* (pig) (clusters 62 and 116, see Table 5), would pass through, rendering the target "senseless", so to speak. Of course, lowering the threshold to, say, 0.2, in order to accomodate the *animal* and *food* senses for *gris* (pig), would mean that too many clusters are included for *hest* (horse). Instead of settling on some global criterion common to all words, the final sense assignments should be based on individually learned thresholds.

The four most salient clusters for *sjel* (soul):

Cluster 93, membership: 0.8428		**Cluster 55, membership: 0.3558**	
ånd (spirit)	0.9136	*hånd* (hand)	0.9350
sjel (soul)	0.8428	*hand* (hand)	0.8933
ånde (breath), Def Sg/Pl = ånd	0.8226	*ansikt* (face)	0.7918
gud (god)	0.4058	*arm* (arm)	0.7872
dyr (animal)	0.3934	*hode* (head)	0.7571
følelse (feeling)	0.3788	*finger* (finger)	0.7292
vesen (being)	0.3704	*skulder* (shoulder)	0.6846
kropp (body)	0.3697	*kropp* (body)	0.6817
menneske (human)	0.3686	*fot* (foot)	0.6689
natur (nature)	0.3489	*ben* (leg, bone)	0.6628
Cluster 10, membership: 0.3392		**Cluster 62, membership: 0.2907**	
tanke (thought)	0.8885	*fugl* (bird)	0.8558
tank (tank, Def Sg/Pl = tanke)	0.8806	*hund* (dog)	0.8330
følelse (feeling)	0.8378	*katt* (cat)	0.7990
tanker (? tanker, Pl = tanke)	0.7318	*katte* (cat)	0.7660
kjærlighet (love)	0.6250	*slange* (snake)	0.6261
opplevelse (experience)	0.6239	*slang* (slang, Def Sg = slange)	0.6039
glede (pleasure, happiness)	0.5888	*mann* (man)	0.5556
sorg (sorrow, grief)	0.5748	*dame* (woman)	0.5293
smerte (pain, ache)	0.5710	*dyr* (animal)	0.4998
lengsel (yearning, longing)	0.5476	*gutt* (boy)	0.4810

Table 6: Cluster memberships of *sjel* (soul)

One inherent limitation of the approach described in this paper is that it is only suitable for words of the higher frequency stratas for which we can observe sufficient syntactical co-occurence information. (Grefenstette, 1993) compares classical windowing techniques to methods using lexical-syntactic relations for the task of extracting similarity relations from corpora, and finds that local context information provides very precise sense indicators when

available, but that a window based approach seems more viable when dealing with infrequent and rare words. However, one type of representation does not exclude the other, and when using distributional data one might actually benefit from a division of labor between different types of contextual representations. In addition to the local context features that we used in this paper, distributional profiles based on more broadly defined "topical" contexts could also be associated with the words and the classes. The core members of the classes could consist of high-frequency words clustered on the basis of reliable features of the local context. When words of less frequent appearance are to be categorized or compared, one could then fall back on a representation of a broader contextual distribution.

(Velldal, 2003) also describes other variations over the hybrid approach presented in this paper, where the result of the bottom-up pass is used to initialize further clustering with the *fuzzy c-means* (Bezdek, 1981) and *possibilistic c-means* (Krishnapuram and Keller, 1993) methods. The literature on fuzzy computing contains a well of other clustering methods that can be applied in order to automatically elicit fuzzy membership functions directly from data. Undoubtedly, many of these methods might also profitably be applied to the task of inferring semantic word classes directly from distributional language data.

The four most salient clusters for *språk* (language):

Cluster 54, membership: 0.9157		**Cluster 132, membership: 0.4432**	
kultur (culture)	0.9332	*norsk* (Norwegian)	0.9761
språk (language)	0.9157	*engelsk* (English)	0.7895
tradisjon (tradition)	0.6337	*tysk* (German)	0.6423
litteratur (literature)	0.5628	*fransk* (French)	0.6351
religion (religion)	0.5507	*samisk* (Lapp)	0.4804
kunst (art)	0.5101	*språk* (language)	0.4432
identitet (identity)	0.4562	*morsmål* (mother tongue)	0.3445
samfunn (community, society)	0.4475	*matematikk* (mathematics)	0.3347
miljø (environment)	0.4153	*ord* (word)	0.3200
tenkning (thought, thinking)	0.3910	*fag* (subject)	0.3085

Cluster 86, membership: 0.2957		**Cluster 29, membership: 0.2403**	
ord (word)	0.9143	*uttrykk* (expression)	0.9127
ting (thing)	0.8142	*begrep* (notion, conception)	0.7510
navn (name)	0.7963	*setning* (sentence)	0.6710
sang (song)	0.5780	*ytring* (statement, utterance)	0.6690
musikk (music)	0.5779	*utsagn* (statement, assertion)	0.4715
lyd (sound)	0.4898	*ord* (word)	0.4498
vers (verse)	0.4785	*tekst* (text)	0.4084
melodi (melody)	0.4768	*fortelling* (story)	0.3868
tekst (text)	0.4598	*tegn* (sign)	0.3727
dikt (poem)	0.4138	*formulering* (formulation)	0.3614

Table 7: Cluster memberships of *språk* (language)

The four most salient clusters for *reaksjon* (reaction):

Cluster 10, membership: 0.3834		**Cluster 105, membership: 0.3427**	
tanke (thought)	0.8885	kritikk (criticism, review)	0.9933
tank (tank, Def Sg/Pl = tanke)	0.8806	anklage (accusation)	0.6776
følelse (feeling)	0.8378	beskyldning (accusation, charge)	0.6768
tanker (? tanker, Pl = tanke)	0.7318	angrep (attack, charge)	0.3921
kjærlighet (love)	0.6250	innvending (objection)	0.3904
opplevelse (experience)	0.6239	spark (kick)	0.3858
glede (pleasure, happiness)	0.5888	protest (protest)	0.3665
sorg (sorrow, grief)	0.5748	oppfordring (invitation, appeal)	0.3654
smerte (pain, ache)	0.5710	press (pressure, stress)	0.3600
lengsel (yearning, longing)	0.5476	reaksjon (reaction)	0.3427
Cluster 49, membership: 0.3181		**Cluster 152, membership: 0.3145**	
faktor (factor)	0.8494	virkning (effect)	0.8708
egenskap (quality, property)	0.8112	konsekvens (consequence)	0.8607
trekk (feature, move)	0.7797	betydning (meaning, consequence)	0.8236
element (element)	0.7651	effekt (effect)	0.7783
kjennetegn (mark, characteristic)	0.6440	utslag (outcome, result)	0.4903
aspekt (aspect)	0.5010	skadevirkning (harmful effect)	0.4700
forutsetning ((pre)	0.4325	sammenheng (connection)	0.4596
komponent (component)	0.4258	årsak (cause)	0.4340
svakhet (weakness)	0.4214	problem (problem)	0.3975
holdning (attitude)	0.4125	forskjell (difference)	0.3948

Table 8: Cluster memberships of *reaksjon* (reaction)

Bezdek, J. C; Pattern Recognition with Fuzzy Objective Function Algorithms, *Advanced Applications in Pattern Recognition*. Plenum Press, 1981.

Cutting, D.R.; Karger, D.; Pedersen, J. & Tukey, J. W.; Scatter/Gather: A cluster-based approach to browsing large document collections. *Proceedings of SIGIR-92*. Copenhagen, Denmark, 1992.

Gärdenfors, P.; *Conceptual spaces: the geometry of thought.* MIT Press, Cambridge, 2000.

Grefenstette, G.; Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches, *Proceedings of the ACL SIGLEX Workshop on Lexical Acquisition*. Columbus, Ohio, 1993.

Hagen, K.; J. B. Johannessen & A. Nøklestad; A constraint-based tagger for Norwegian, *Proceedings of the 17th Scandinavian Conference of Linguistics*. 2000.

Krishnapuram, R. & J. M. Keller; A possibilistic approach to clustering, *IEEE Transactions On Fuzzy Systems 1(2)*. 1993

Lowe, W. & S. McDonald; *The direct route: Mediated priming in semantic space* (Informatics Research Report EDI-INF-RR-0017). Division of Informatics, University of Edinburgh, 2000.

Pantel, P. & D. Lin; Discovering word senses from text, *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2002.

Pereira, F., N. Tishby, & L. Lee; Distributional clustering of english words, *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. 1993.

Ruspini, E. H. & E. Francesc; Interpretations of Fuzzy Sets, *Handbook of Fuzzy Computation* (Chapter B2.3). Institute of Physics Publishing, 1998.

Velldal, E.; *Modeling word senses with fuzzy clustering* (Cand.philol. thesis). University of Oslo, 2003.

Zadeh, L. A.; Fuzzy sets, *Information and Control 8*, pp. 338–353. 1965.