Characterizing Response Types and Revealing Noun Ambiguity in German Association Norms

Alissa Melinger and Sabine Schulte im Walde and Andrea Weber
Psycholinguistics and Computational Linguistics
Saarland University, Saarbrücken, Germany
`{melinger,schulte,aweber}@coli.uni-sb.de`

Abstract

This paper presents an analysis of semantic association norms for German nouns. In contrast to prior studies, we not only collected associations elicited by written representations of target objects but also by their pictorial representations. In a first analysis, we identified systematic differences in the type and distribution of associate responses for the two presentation forms. In a second analysis, we applied a soft cluster analysis to the collected target-response pairs. We subsequently used the clustering to predict noun ambiguity and to discriminate senses in our target nouns.

1 Introduction

Language is rife with ambiguity. Sentences can be structurally ambiguous, pronouns can be referentially ambiguous, and words can be polysemous. The human language faculty deals remarkably well with the omnipresent ambiguity, so well in fact that we are rarely aware of the multiple alternatives that are available. Despite our apparent lack of awareness, psycholinguistic research has shown that alternative meanings are nevertheless activated during processing. For example, in a seminal study of homograph recognition, Tanenhaus et al. (1979) demonstrated that multiple meanings of a homograph are initially activated even in highly constraining syntactic contexts, such as *They all rose* vs. *They bought a rose.* Likewise in speech production, Cutting and Ferreira (1999) showed that non-depicted senses of homophones are activated during picture naming. Thus, when either a homograph word or a picture with a homophone name are processed, multiple meanings are initially activated.

Intuitively, however, one might expect differences in the degree to which multiple meanings are activated depending on the presentation mode. To our knowledge no investigation has compared picture (top-down) and word (bottom-up) semantic processing. In this paper, we investigate differences in the semantic information, namely associations, elicited in these two presentation modes. We reason that, if multiple meanings of an ambiguous word are activated when the stimulus is processed, then the elicited associates should reflect the ambiguity. If the *degree* of activation differs with respect to the presentation mode, the associates should reflect this difference as well.

Manually linking associates to a particular word sense would be time intensive and subjective. Thus, we rely on computational methods that have the potential to automatically compare the associates provided for the two presentation modes and classify them into meaning-referring sets. These methods thus not only reveal differences in the associates elicited in the two presentation conditions but also, in the case of ambiguous nouns, identify which associates are related to which meaning of the word.

Our analyses are guided by the following two questions:

1. *Are there systematic differences in associate response types when target objects are presented in written form compared to when the written form is accompanied by a pictorial representation?* Predictions about which differences we expected in the response types are made, and the associate responses are analyzed accordingly (Section 4).

2. *Can we identify multiple senses of the nouns and discriminate between noun senses based on the associate responses?* We apply a clustering technique to the target-response pairs; the cluster analysis gathers semantically similar target nouns, based on overlapping sets of associate responses, and predicts the ambiguity of nouns and their senses (Section 5).

In Section 2, we provide an overview of the types of differences we anticipate; Section 3 describes the materials and procedure used for the association elicitation; in Sections 4 and 5, we explain how response types were characterized and noun senses identified.

2 Intuitions

A critical component of the current study was the presentation of target stimuli in two forms: Lexical stimuli consisted of the written name of target objects; pictorial stimuli consisted of the written names accompanied by black and white line drawings of the referred-to objects.

We assumed that, in some cases, associate responses elicited by written words would be different from associate responses elicited by pictures. Differences in responses might arise from a variety of sources: a) images might increase the salience of physical attributes of objects, b) images might show non-prototypical characteristics of objects that would not be evoked by words, c) when word forms have different shades of meaning, responses evoked by lexical stimuli might index any of the words' meanings while responses evoked by pictorial representations might be more biased towards the depicted sense.

To illustrate these points, consider the following example. The picture of a *Hexe* 'witch' from our study showed a witch riding on a broom, see Figure 1. This particular choice of activity, rather than, for example, a plausible alternative like stirring a cauldron or simply standing by herself, accentuated the relationship between *witch* and *broom*. Indeed, we found that this accentuation was reflected in the associate responses: 27 of the 50 participants (54%) who saw the picture of the *witch* produced *broom* as an associate while only 18 participants (36%) who read the word *witch* produced *broom*. Thus, the association strength of a response elicited by words does not necessarily generalize to picture stimuli, and vise versa.

To demonstrate the relevance of presentation mode for potentially ambiguous nouns, consider a second example. The German word for 'lock' is *Schloss*. *Schloss*, however, also means 'castle'. Associate responses such as *Schlüssel* 'key' and *Fahrrad* 'bicycle' might be elicited by the lock meaning of the word while responses such as *Prinzessin* 'princess' or *Burg* 'castle' would index the alternative meaning.

Figure 1: Example picture for *witch*.

3 Data Collection Method

This section introduces our elicitation procedure.

Materials: 409 German nouns referring to picturable objects were chosen as target stimuli. To ensure broad coverage, target objects represented a variety of semantic classes including animals, plants, professions, food, furniture, vehicles, and tools. Simple black and white line drawings of target stimuli were drawn from several sources, including Snodgrass and Vanderwart (1980) and the picture database from the Max Planck Institute for Psycholinguistics in the Netherlands.

Participants: 300 German participants, mostly students from Saarland University, received either course credit or monetary compensation for filling out a questionnaire.

Procedure: The 409 target stimuli were divided randomly into three separate questionnaires consisting of approximately 135 nouns each. Each questionnaire was printed in two formats: target objects were either presented as pictures together with their preferred name (to ensure that associate responses were provided for the desired lexical item) or the name of the target objects was presented without a representative picture accompanying it. Next to each target stimulus three lines were printed on which participants could write up to three semantic associate responses for the stimulus. The order of stimulus presentation was individually randomized for each participant. Participants were instructed to give one associate word per line, for a maximum of three responses per trial. No time limits were given for responding, though participants were told to work swiftly and without interruption. Each version of the questionnaire was filled out by 50 participants, resulting in a maximum of 300 data points for any given target stimulus (50 participants × 2 presentation modes × 3 responses).

Collected associate responses were entered into a database with the following additional infor-

mation: For each target stimulus we recorded a) whether it was presented as a picture or in written form, and b) whether the name was a homophone (and thus likely to elicit semantic associates for multiple meanings). For each response type provided by a participant, we coded a) the order of the response, i.e., first, second, third, b) the part-of-speech of the response, and c) the type of semantic relation between the target stimulus and the response (e.g., part-whole relations such as *car – wheel*, and categorical relationship such as hypernymy, hyponymy, and synonymy).

4 Analysis of Response Types

As described in Section 2, one might expect variation in the response types for the two presentation modes, because the associations provided in the 'picture+word' condition were biased towards the depicted sense of the target noun. Our first analysis evaluates what sorts of differences are in fact observed in the data, i.e., which intuitions are empirically supported, and which are not. To this end, this section is concerned with systematic differences in response types when target stimuli were presented in written form ('word only', subsequently *W* condition) or when the written form was accompanied by a picture ('picture+word', subsequently *PW* condition). We first give our predictions for the differences in response types and then continue with the corresponding analyses of response types.

4.1 Predictions

Based on our intuitions, we predicted the following differences.

1. The overall number of *response tokens* is unlikely to differ for the two presentation modes, since participants are limited to three associate responses per target stimulus in both presentation modes.

2. The overall number of *response types*, however, should differ: in the PW condition we expect a bias towards the depicted noun sense, resulting in a smaller number of response types than in the W condition.

3. The PW condition produces less *idiosyncratic response types* than the W condition, because pictures reinforce associations that are either depicted, or at least related to the depicted sense and its characteristics, resulting in less response diversity.

4. The PW condition receives more associations that show a *part-of relation* to the target stimulus than the W condition, because characteristics of the pictures can highlight specific parts of the whole.

5. The *type agreement*, i.e., the number of response types on which the PW and the W conditions agree is expected to differ with respect to the target noun. For target nouns that are highly ambiguous we expect low type agreement. Note that this prediction does not refer to a PW-W distinction, but instead uses the PW-W distinction to approach the issue of noun senses.

4.2 Response Type Distributions

The analyses to follow are based on stimulus-response frequency distributions: For each target stimulus and each response type, we calculated how often the response type was provided. The result was a frequency distribution for the 409 target nouns, providing frequencies for each response type. The frequency distributions were distinguished for the PW condition and the W condition. Table 1 provides an example of the most frequent response types and their frequencies for the homophone target noun *Schloss*, as described in Section 2; the 'lock' meaning was depicted, 'castle' is an alternative meaning. Hereafter, we will refer to an association provided in the PW condition as *association_PW*, and an association provided in the W condition as *association_W*, e.g., *Burg_PW* vs. *Burg_W*.

Association		POS	PW	W
Schlüssel	'key'	N	38	13
Tür	'door'	N	10	5
Prinzessin	'princess'	N	0	8
Burg	'castle'	N	0	8
sicher	'safe'	ADJ	7	0
Fahrrad	'bike'	N	7	0
schließen	'close'	V	6	1
Keller	'cellar'	N	7	0
König	'king'	N	0	7
Turm	'tower'	N	0	6
Sicherheit	'safety'	N	5	1
Tor	'gate'	N	2	4
zu	'shut'	ADV	4	1

Table 1: Response type frequencies for *Schloss*.

4.3 Results

Based on the frequency distributions in Section 4.2, we analyzed the response types according to our predictions in Section 4.1.

Number of response tokens: The number of response tokens was compared for each target stimulus in both presentation modes. The total number of response tokens was 58,642 (with mean $\mu = 143$) in the PW condition and 58,072 ($\mu = 142$) in the W condition. We had predicted that *Token(PW) ∼ Token(W)*. The analysis showed, however, that in 243 of 409 cases (59%) the number of response tokens was larger for PW than for W *(Token(PW) > Token(W))*; in 132 cases (32%) *Token(PW) < Token(W)*, and in 34 cases (8%) *Token(PW) = Token(W)*. The unpredicted difference between presentation modes was significant across items in a two-tailed t-test, $t(408) = 6.077, p < .001$. We take the result as an indication that pictures facilitate the production of associations. This is an interesting insight especially since the number of associate responses per target stimulus was limited while response time was not.

Number of response types: The number of response types was compared for each target stimulus in both presentation modes. The total number of response types in the PW condition was 19,800 ($\mu = 48$) compared with 20,332 ($\mu = 50$) in the W condition. We had predicted that *Type(W) > Type(PW)*. The results showed indeed that in 229 of the 409 cases (56%) the number of response types was larger for W than for PW *(Type(W) > Type(PW))*; in 152 cases (37%) *Type(PW) > Type(W)*, and in 28 cases (7%) *Type(PW) = Type(W)*. This predicted difference, although small, was significant, $t(408) = 3.63, p < .001$.

Idiosyncratic response types: The proportions of idiosyncratic response types (i.e., associate responses that were provided only once for a certain target stimulus) were compared for each target stimulus in both presentation modes. In total, 12,011 ($\mu = 29$) idiosyncratic responses were provided in the PW condition and 12,582 ($\mu = 31$) idiosyncratic responses in the W condition. We had predicted that *Idio(W) > Idio(PW)*. The analysis showed indeed that in 216 of the 409 cases (53%) the number of idiosyncratic responses was larger for W than for PW *(Idio(W) > Idio(PW))*; in 175 cases (43%) *Idio(PW) > Idio(W)*, and in 18 cases (4%) *Idio(PW) = Idio(W)*. The predicted difference was reliable across items, $t(408) = 3.76, p < .001$. This pattern of results is consistent with the notion of a restricted set of responses in the PW condition relative to the W condition.

Part-of response types: Based on the manual annotation of semantic relations between target nouns and responses, proportions of response types which stand in a part-of relation to the target nouns were determined. The total number of part-of response types was 876 ($\mu = 2.7$) in the PW condition, and 901 ($\mu = 2.8$) in the W condition. We predicted that *Part(PW) > Part(W)*. The analysis showed however that in only 94 of the 409 cases (29%) the number of part-of responses was larger for PW than for W *(Part(PW) > Part(W))*; in 114 cases (35%) *Part(W) > Part(PW)*, and in 115 cases (36%) *Part(W) = Part(PW)*. The difference between conditions was not significant across items, $t(322) = 1.42, p > .1$. The absence of a reliable difference in this analysis possibly suggests that our pictures did not regularly enhance a part-whole relationship.

Type agreement: The final analysis was based on response type agreement for PW and W. However, this analysis did not aim to distinguish between the two presentation modes but rather used the agreement proportions as a diagnostic of potential target noun ambiguity. Here we calculated the total amount of overlap between the PW and W conditions. For this calculation, we identified the number of response types that occur in both the PW and W conditions for a particular target stimulus and divided that number by the total number of response types produced for that target stimulus, irrespective of condition. In other words, if a *noun_PW* receives responses A and B and *noun_W* receives responses B and C, then the total number of shared response types is 1, namely response B, and the total number of response types across conditions is 3, namely A, B and C. Thus, the proportion of agreement is .33.

We reasoned that target nouns with low type agreement are likely to be ambiguous. To test this, we sorted the targets by their proportion of agreement, and compared the top and bottom 20 targets. In the manual annotation of our stimuli, cf. Section 3, we had recorded that 10% of our stimuli were homophones. Thus, a random distribution would predict two ambiguous items in a 20 item sample if the proportion of agreement is not an indicator of ambiguity. Instead, we found 11 ambiguous nouns in the set of 20 targets with lowest agreement proportions and 2 ambiguous nouns in the set of 20 targets with highest agreement proportions. A χ^2 test indicated that the number of

ambiguous nouns found in the two sets differed significantly, $\chi^2 = 7.29, p < .01$.

Summarizing this first set of analyses, we found that the associate responses for concrete German nouns differed significantly depending on the format under which they were elicited, namely the presentation mode. The fact that we found more response types in total and also more idiosyncratic responses when target nouns were presented in the 'word only' vs. the 'picture+word' condition suggests that alternative meanings were more active when participants were presented with written stimuli compared to depicted stimuli. It is also interesting to note that not all our intuitive predictions were born out. For example, despite our feeling that the picture should bias the inclusion of depicted part-of relations, such as the *broom* ∼ *witch* example discussed above, this intuition was not supported by the data. This fact highlights the importance of first analyzing the responses to ensure the necessary conditions are present for the identification of ambiguous words.

5 Analysis of Noun Senses

The second analysis in this paper addresses the distinction of noun senses on the basis of associations. Our goal is to identify the – potentially multiple – senses of target nouns, and to reveal differences in the noun senses with respect to the presentation modes. The analysis was done as follows.

1. The target-response pairs were clustered. The soft cluster analysis was expected to assign semantically similar noun senses into common clusters, as based on shared associate responses. (Section 5.1)

2. The clusters were used to predict the ambiguity of nouns and their respective senses. (Section 5.2)

3. The clusters and their predictability were evaluated by annotating noun senses with *Duden* dictionary definitions, and calculating interannotator agreement. (Section 5.3)

5.1 Latent Semantic Noun Clusters

Target nouns were clustered on the basis of their association frequencies, cf. Table 1. I.e., the clustering result was determined by joint frequencies of the target nouns and the respective associations. The targets themselves were described by

the noun-condition combination, e.g. *Schloss_PW*, and *Schloss_W*. We used noun-condition combinations as compared to nouns only, because the clustering result should not only distinguish senses of nouns in general, but in addition predict the noun senses with respect to the condition.

Various techniques have been exploited for word sense disambiguation. Closely related to our work, Schvaneveldt's *pathfinder networks* (Schvaneveldt, 1990) were based on word associations and were used to identify word senses. An enourmous number of approaches in computational linguistics can be found on the *SENSE-VAL* webpage (SENSEVAL,), which hosts a word sense disambiguation competition. We applied Latent Semantic Clusters *(LSC)* to our association data. The LSC algorithm is an instance of the Expectation-Maximisation algorithm (Baum, 1972) for unsupervised training based on unannotated data, and has been applied to model the selectional dependency between two sets of words participating in a grammatical relationship (Rooth, 1998; Rooth et al., 1999). The resulting cluster analysis defines two-dimensional soft clusters which are able to generalise over hidden data. LSC training learns three probability distributions, one for the probabilities of the clusters, and one for each tuple input item and each cluster (i.e., a probability distribution for the target nouns and each cluster, and one for the associations and each cluster), thus the two dimensions. We use an implementation of the LSC algorithm as provided by Helmut Schmid.

The LSC output depends not only on the distributional input, but also on the number of clusters to model. As a rule, the more clusters are modeled, the more skewed the resulting probability distributions for cluster membership are. Since the goal of this work was not to optimize the clustering parameters, but to judge the general predictability of such models, we concentrated on two clustering models, with 100 and 200 clusters, respectively.

Table 2 presents the most probable noun-condition combinations for a cluster from the 100-cluster analysis: The cluster probability is 0.01295 (probabilities ranged from 0.00530 to 0.02674). The most probable associations that were common to members of this cluster were *Ritter* 'knight', *Mittelalter* 'medieval times', *Rüstung* 'armour', *Burg* 'castle', *Kampf* 'fight', *kämpfen* 'fight', *Schwert* 'sword', *Waffe* 'weapon', *Schloss* 'castle',

scharf 'sharp'. This example shows that the associations provide a semantic description of the cluster, and the target nouns themselves appear in the cluster if one of their senses is related to the cluster description. In addition, we can see that, e.g., *Schloss* appears in this cluster only in the W condition. The reason for this is that the picture showed the 'lock' sense of *Schloss*, so the PW condition was less likely to elicit 'castle'-related responses. This example cluster illustrates nicely what we expect from the cluster analysis with respect to distinguishing noun senses.

Target Noun		Cond	Prob
Rüstung	'armour'	W	0.097
Schwert	'sword'	W	0.097
Burg	'castle'	W	0.096
Rüstung	'armour'	PW	0.096
Dolch	'dagger'	PW	0.095
Schwert	'sword'	PW	0.093
Burg	'castle'	PW	0.091
Dolch	'dagger'	W	0.089
Ritter	'knight'	PW	0.073
Ritter	'knight'	W	0.068
Schloss	'castle'	W	0.040
Turm	'tower'	PW	0.014

Table 2: Sample cluster, 100-cluster analysis.

5.2 Prediction of Noun Ambiguity and Noun Senses

The noun clusters were used to predict the ambiguity of nouns and their respective senses. The two-dimensional cluster probabilities, as introduced above, offer the following information:

- *Which associations are highly probable for a cluster?* The most probable associations are considered as defining the semantic content of the cluster.

- *Which target nouns are highly probable for a cluster and its semantic content, i.e. the associations?* Relating the target nouns in a cluster with the cluster associations defines the respective sense of the noun. To refer to the above example, finding *Schloss* in a cluster together with associations such as 'castle' and 'fight' relates this instance of *Schloss* to the 'castle' sense and not the 'lock' sense.

- *Which target nouns are in the same cluster and therefore refer to a common sense/aspect of the nouns?* This information is relevant for revealing sense differences of target nouns with respect to the conditions PW vs. W.

In order to predict whether a noun is in a cluster or not, we needed a cut-off value for the membership probability. We settled on 1%, i.e., a target noun with a probability of $\geq 1\%$ was considered a member of a cluster. Based on the 200-cluster information, we then performed the following analyses on noun ambiguity and noun senses.

Prediction of noun ambiguity: For each target noun, we predicted its ambiguity by the number of clusters it was a member of. For example, the highly ambiguous noun *Becken* 'basin, cymbal, pelvis' (among other senses), was a member of 8 clusters, as compared to the unambiguous *Bäcker* 'baker' which was a member of only one cluster. Membership in several clusters does not necessarily point to multiple noun senses (because different combinations of associations might define similar semantic contents), but nevertheless the clusters provide an indication of the degree of noun ambiguity. The total number of senses in the 200-cluster analysis was 735, which means an average of 1.8 senses for each target stimulus (across presentation condition).

Discrimination of noun senses: The most probable associations in the clusters were assumed to describe the semantic content of the clusters. They can be used to discriminate noun senses of polysemous nouns. Referring back to our example noun *Becken*, it appeared in one cluster with the most probable associations *Wasser* 'water', *Garten* 'garden', *Feuerwehr* 'fire brigade', *gießen* 'water', and *nass* 'wet', describing the 'basin' sense of the target noun; in a second cluster it appeared with *Musik* 'music', *laut* 'loud', *Instrument* 'instrument', *Orchester* 'orchestra', and *Jazz*, describing the music-related sense; and in a third cluster it appeared with *Hand* 'hand', *Bein* 'leg', *Ellenbogen* 'elbow', *Körper* 'body' and *Muskel* 'muscle', describing the body-related sense, etc.

Noun similarity: Those target nouns which were assigned to a common cluster were assumed to be semantically similar (with respect to the cluster content). Again, referring back to our example noun *Becken* and the three senses discriminated above, in the first cluster refering to the 'basin' sense we find other nouns such as *Eimer* 'bucket', *Fontäne* 'fountain', *Brunnen* 'fountain, well', *Weiher* 'pond', and *Vase* 'vase', all related to water and water container; in the second cluster referring to the music sense we find *Tuba* 'tuba', *Trompete* 'trumpet', *Saxophon* 'sax', and *Trommel* 'drum',

and in the third cluster referring to the body sense we find *Arm* 'arm', and *Knochen* 'bone'.

Discrimination of PW vs. W noun senses: Combining the previous two analyses allowed us to discriminate senses as provided by the two experimental conditions. Remember that the target nouns in the clusters included the specification of the condition. If we find a target noun in a certain cluster with both condition specifications, it means that some associations produced to both the PW and the W conditions referred to the same noun sense. If a target noun appears in a certain cluster only with one condition specified, it means that the associations captured the respective noun sense only in one condition. Thus, a target noun appearing in a cluster in only one condition was an indication for ambiguity. Going back to our example noun *Becken* and its three example clusters, we find the noun in both conditions only in one of the three clusters, namely the cluster for the music sense, and this happens to be the sense depicted in the PW condition. In the two other clusters, we only find *Becken* in the W condition. In total, *Becken* appears in both conditions only in 1 out of 8 clusters, in only the PW condition in 1 cluster, and in only the W condition in 6 clusters.

The four analyses demonstrate that and how the clusters can be used to predict and discriminate noun senses. Of course, the predictions are not perfect, but they apprximately correspond to our linguistic intuitions. Impressively, the clusters revealed not only blatantly polysemous words such as *Becken* but also distinct facets of a word. For example, the stimulus *Filter* 'filter' had associations to coffee-related senses as well as cigarette-related senses, both of which were then reflected in the clusters.

5.3 Evaluation of Noun Clusters

In order to perform a more independent evaluation of the clusters which is not only based on specific examples, we assessed the clusters by two annotators. 20 homophones were manually selected from the 409 target nouns. In addition, we relied on the indicators for ambiguity as defined in Section 4, and selected the 20 top and bottom nouns from the ordered list of type agreement for the two conditions. The manual list showed some overlap with the selection dependent on type agreement, resulting in a list of 51 target nouns.

For each of the selected target nouns, we looked

up the noun senses as defined by the *Duden*, a standard German dictionary. We primarily used the stylistic dictionary (Dudenredaktion, 2001), but used the foreign language dictionary (Dudenredaktion, 2005) if the noun was missing in the former. Each target noun was defined by its (short version) sense definitions. For example, *Schloss* was defined by the senses *Vorrichtung zum Verschließen* 'device for closing' and *Wohngebäude von Fürsten und Adeligen* 'residential building for princes and noblemen'.

As targets for the evaluation, we used the two cluster analyses as mentioned above, containing 100 and 200 clusters with membership probability cut-offs at 1%. Two annotators were then presented with two lists each: For each cluster analysis, they saw a list of the 51 selected target nouns, accompanied by the clusters they were members of, i.e., for which they showed a probability $\geq 1\%$, ignoring the condition of the target noun (PW vs. W). In total, the annotators were given 82/91 clusters which included any of the 51 selected nouns. For each cluster, the annotators saw the five most probable associations, and all cluster members. The annotators were asked to select a Duden sense for each cluster, if possible. The results of the annotation are presented in Table 3. Annotator 1 identified a *Duden* sense for 72/75% of the clusters, annotator 2 for 78/71%. Interannotator agreement on which of the *Duden* senses was appropriate for a cluster (if any) was 81/85%; $\kappa = .78/.75$.

Source	100 clusters		200 clusters	
No. of clusters	82		91	
Annotator 1	59	72%	68	75%
Annotator 2	64	78%	65	71%

Table 3: Clusters and identified *Duden* senses.

The evaluation of the clusters as carried out by the sense annotation demonstrates that the cluster senses correspond largely to *Duden* senses. This first kind of evaluation models the precision of the cluster analyses. A second kind of evaluation assessed how many different *Duden* senses we capture with the cluster analyses; this evaluation models the recall of the cluster analyses. *Duden* defines a total of 113 senses to our target nouns. Table 4 specifies the recall for the data sets and annotators.

The evaluations show that the precision is much larger than the recall. It might be worth applying the clustering with a different number of clusters

Source	100 clusters		200 clusters	
Annotator 1	46	41%	54	48%
Annotator 2	51	45%	52	46%

Table 4: Cluster recall of *Duden* senses.

and/or a different cut-off for the cluster membership probability, but that would lower the precision of the analyses. We believe that the evaluation numbers are quite impressive, especially considering that *Duden* not only specifies everyday vocabulary, but includes colloquial expressions (such as *Ballon* as 'human head'), out-dated senses (such as *Mond* as 'month'), and domain-specific senses (such as *Blatt* as 'shoulder of a hoofed game').

6   Conclusions

In this paper we evaluated differences in the types and strengths of semantic associations elicited under two conditions of presentation, 'picture+word' and 'word only'. Consistent with prior psycholinguistic research, we observed associations to different meanings of a word in both conditions, supporting the idea that multiple meanings of homonyms are active during both picture and word processing. However, our analyses of response types also showed that responses to pictures were less diverse and idiosyncratic than responses to words, suggesting that the degree to which alternative meanings are active in the two presentation modes may indeed be different. One further implication of the analyses is that semantic associations (and especially association strengths) from word-based norming studies do not necessarily generalize for the purpose of experiments using depicted materials. This insight should have an impact on psycholinguistic studies when selecting depicted vs. written stimuli.

Our predictions for the types of differences we expected were based on intuitive grounds. One might therefore question the value of the analyses presented in Section 4. It is interesting to note, however, that some of the predictions were in fact not born out. As the cluster analysis presented in Section 5 required differences between the two stimulus modes, it was critical that a proper evaluation of those differences be conducted, even if some of them seem trivially true.

The cluster analysis demonstrated that we can capitalize on the semantic associations and both identify and discriminate the various senses of the target nouns. Indeed, the clusters not only re-

vealed sense differences of target nouns with respect to their presentation modes, but also detected noun senses which had not been identified by the authors initially. This indicates that this method not only can discriminate between senses but it can also detect ambiguity. The cluster analysis allowed us to apply automatic methods of identifying which meaning of a word a particular associate refers to, which would otherwise be a time consuming and error-prone manual activity.

References

Leonard E. Baum. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, III:1–8.

J. Cooper Cutting and Victor S. Ferreira. 1999. Semantic and phonological information flow in the production lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(2):318–344.

Dudenredaktion, editor. 2001. *DUDEN – Das Stilwörterbuch*. Number 2 in 'Duden in zwölf Bänden'. Dudenverlag, Mannheim, 8th edition.

Dudenredaktion, editor. 2005. *DUDEN – Deutsch als Fremdsprache Standardwörterbuch*. Dudenverlag, Mannheim, 1st edition.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Mats Rooth. 1998. Two-dimensional clusters in grammatical relations. In *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3). Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Roger W. Schvaneveldt, editor. 1990. *Pathfinder Associative Networks*. Studies in Knowledge Organization. Ablex Publishing Corporation, Norwood, NJ.

SENSEVAL. Evaluation exercises for Word Sense Disambiguation. http://www.senseval.org/. Organized by ACL-SIGLEX.

Joan Gay Snodgrass and Mary Vanderwart. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6:174–215.

Michael K. Tanenhaus, James M. Leiman, and Mark S. Seidenberg. 1979. Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior*, 18:427–440.