Towards Robust Animacy Classification Using Morphosyntactic Distributional Features

Lilja Øvrelid
NLP-unit, Dept. of Swedish
Göteborg University
SE-40530 Göteborg, Sweden
lilja.ovrelid@svenska.gu.se

Abstract

This paper presents results from experiments in automatic classification of animacy for Norwegian nouns using decision-tree classifiers. The method makes use of relative frequency measures for linguistically motivated morphosyntactic features extracted from an automatically annotated corpus of Norwegian. The classifiers are evaluated using leave-one-out training and testing and the initial results are promising (approaching 90% accuracy) for high frequency nouns, however deteriorate gradually as lower frequency nouns are classified. Experiments attempting to empirically locate a frequency threshold for the classification method indicate that a subset of the chosen morphosyntactic features exhibit a notable resilience to data sparseness. Results will be presented which show that the classification accuracy obtained for high frequency nouns (with absolute frequencies $>$1000) can be maintained for nouns with considerably lower frequencies ($\sim$50) by backing off to a smaller set of features at classification.

1 Introduction

Animacy is a an inherent property of the referents of nouns which has been claimed to figure as an influencing factor in a range of different grammatical phenomena in various languages and it is correlated with central linguistic concepts such as agentivity and discourse salience. Knowledge about the animacy of a noun is therefore relevant for several different kinds of NLP problems ranging from coreference resolution to parsing and generation.

In recent years a range of linguistic studies have examined the influence of argument animacy in grammatical phenomena such as differential object marking (Aissen, 2003), the passive construction (Dingare, 2001), the dative alternation (Bresnan et al., 2005), etc. A variety of languages are sensitive to the dimension of animacy in the expression and interpretation of core syntactic arguments (Lee, 2002; Øvrelid, 2004). A key generalisation or tendency observed there is that prominent grammatical features tend to attract other prominent features;[1] subjects, for instance, will tend to be animate and agentive, whereas objects prototypically are inanimate and themes/patients. Exceptions to this generalisation express a more *marked* structure, a property which has consequences, for instance, for the distributional properties of the structure in question.

Even though knowledge about the animacy of a noun clearly has some interesting implications, little work has been done within the field of lexical acquisition in order to automatically acquire such knowledge. Orăsan and Evans (2001) make use of hyponym-relations taken from the Word Net resource (Fellbaum, 1998) in order to classify animate referents. However, such a method is clearly restricted to languages for which large scale lexical resources, such as the Word Net, are available. Merlo and Stevenson (2001) present a method for verb classification which relies only on distributional statistics taken from corpora in order to train a decision tree classifier to distinguish between three groups of intransitive verbs.

[1] The notion of prominence has been linked to several properties such as most likely as topic, agent, most available referent, etc.

This paper presents experiments in automatic classification of the animacy of unseen Norwegian common nouns, inspired by the method for verb classification presented in Merlo and Stevenson (2001). The learning task is, for a given common noun, to classify it as either belonging to the class *animate* or *inanimate*. Based on correlations between animacy and other linguistic dimensions, a set of morphosyntactic features is presented and shown to differentiate common nouns along the binary dimension of animacy with promising results. The method relies on aggregated relative frequencies for common noun lemmas, hence might be expected to seriously suffer from data sparseness. Experiments attempting to empirically locate a frequency threshold for the classification method will therefore be presented. It turns out that a subset of the chosen morphosyntactic approximators of animacy show a resilience to data sparseness which can be exploited in classification. By backing off to this smaller set of features, we show that we can maintain the same classification accuracy also for lower frequency nouns.

The rest of the paper is structured as follows. Section 2 identifies and motivates the set of chosen features for the classification task and describes how these features are approximated through feature extraction from an automatically annotated corpus of Norwegian. In section 3, a group of experiments testing the viability of the method and chosen features is presented. Section 4 goes on to investigate the effect of sparse data on the classification performance and present experiments which address possible remedies for the sparse data problem. Section 5 sums up the main findings of the previous sections and outlines a few suggestions for further research.

2 Features of animacy

As mentioned above, animacy is highly correlated with a number of other linguistic concepts, such as transitivity, agentivity, topicality and discourse salience. The expectation is that marked configurations along these dimensions, e.g. animate objects or inanimate agents, are less frequent in the data. However, these are complex notions to translate into extractable features from a corpus. In the following we will present some morphological and syntactic features which, in different ways, approximate the multi-faceted property of animacy:

Transitive subject and (direct) object As mentioned earlier, a prototypical transitive relation involves an animate subject and an inanimate object. In fact, a corpus study of animacy distribution in simple transitive sentences in Norwegian revealed that approximately 70% of the subjects of these types of sentences were animate, whereas as many as 90% of the objects were inanimate (Øvrelid, 2004). Although this corpus study involved all types of nominal arguments, including pronouns and proper nouns, it still seems that the frequency with which a certain noun occurs as a subject or an object of a transitive verb might be an indicator of its animacy.

Demoted agent in passive Agentivity is another related notion to that of animacy, animate beings are usually inherently sentient, capable of acting volitionally and causing an event to take place - all properties of the prototypical agent (Dowty, 1991). The passive construction, or rather the property of being expressed as the demoted agent in a passive construction, is a possible approximator of agentivity. It is well known that transitive constructions tend to passivize better (hence more frequently) if the demoted subject bears a prominent thematic role, preferably agent.

Anaphoric reference by personal pronoun Anaphoric reference is a phenomenon where the animacy of a referent is clearly expressed. The Norwegian personal pronouns distinguish their antecedents along the animacy dimension - animate *han/hun* 'he/she' vs. inanimate *den/det* 'it-MASC/NEUT'.

Anaphoric reference by reflexive pronoun Reflexive pronouns represent another form of anaphoric reference, and, may, in contrast to the personal pronouns locate their antecedent locally, i.e. within the same clause. In the prototypical reflexive construction the subject and the reflexive object are coreferent and it describes an action directed at oneself. Although the reflexive pronoun in Norwegian does not distinguish for animacy, the agentive semantics of the construction might still favour an animate subject.

Genitive -s There is no extensive case system for common nouns in Norwegian and the only distinction that is explicitly marked on the noun is the genitive case by addition of -s. The genitive construction typically describes possession, a relation which often involves an animate possessor.

2.1 Feature extraction

In order to train a classifier to distinguish between animate and inanimate nouns, training data consisting of distributional statistics on the above features were extracted from a corpus. For this end, a 15 million word version of the Oslo Corpus, a corpus of Norwegian texts of approximately 18.5 million words, was employed.[2] The corpus is morphosyntactically annotated and assigns an underspecified dependency-style analysis to each sentence.[3]

For each noun, relative frequencies for the different morphosyntactic features described above were computed from the corpus, i.e. the frequency of the feature relative to this noun is divided by the total frequency of the noun. For transitive subjects (SUBJ), we extracted the number of instances where the noun in question was unambiguously tagged as subject, followed by a finite verb and an unambiguously tagged object.[4] The frequency of direct objects (OBJ) for a given noun was approximated to the number of instances where the noun in question was unambiguously tagged as object. We here assume that an unambiguously tagged object implies an unambiguously tagged subject. However, by not explicitly demanding that the object is preceded by a subject, we also capture objects with a "missing" subject, such as objects occurring in relative clauses and infinitival clauses.

As mentioned earlier, another context where animate nouns might be predominant is in the by-phrase expressing the demoted agent of a passive verb (PASS). Norwegian has two ways of expressing the passive, a morphological passive (verb + *s*) and a periphrastic passive (*bli* + past participle). The counts for passive by-phrases allow for both types of passives to precede the by-phrase containing the noun in question.

With regard to the property of anaphoric reference by personal pronouns, the extraction was bound to be a bit more difficult. The anaphoric personal pronoun is never in the same clause as the antecedent, and often not even in the same sentence. Coreference resolution is a complex problem, and certainly not one that we shall attempt to solve in the present context. However, we might attempt to come up with a metric that approximates the coreference relation in a manner adequate for our purposes, that is, which captures the different coreference relation for animate as opposed to inanimate nouns. To this end, we make use of the common assumption that a personal pronoun usually refers to a discourse salient element which is fairly recent in the discourse. Now, if a sentence only contains one core argument (i.e. an intransitive subject) and it is followed by a sentence initiated by a personal pronoun, it seems reasonable to assume that these are coreferent (Hale and Charniak, 1998). For each of the nouns then, we count the number of times it occurs as a subject with no subsequent object and an immediately following sentence initiated by (i) an animate personal pronoun (ANAAN) and (ii) an inanimate personal pronouns (ANAIN).

The feature of reflexive coreference is easier to approximate, as this coreference takes place within the same clause. For each noun, the number of occurrences as a subject followed by a verb and the 3.person reflexive pronoun *seg* 'him-/her-/itself' are counted and its relative frequency recorded. The genitive feature (GEN) simply contains relative frequencies of the occurrence of each noun with genitive case marking, i.e. the suffix *-s*.

3 Method viability

In order to test the viability of the classification method for this task, and in particular, the chosen features, a set of forty highly frequent nouns were selected - twenty animate and twenty inanimate nouns. A frequency threshold of minimum one thousand occurrences ensured sufficient data for all the features, as shown in table 1, which reports the mean values along with the standard deviation for each class and feature. The total data points for each feature following the data collection are as follows: SUBJ: 16813, OBJ: 24128, GEN: 7830, PASS: 577, ANAANIM: 989, ANAINAN: 944, REFL: 558. As we can see, quite a few of the features express morphosyntactic cues that are

[2] The corpus is freely available for research purposes, see http://www.hf.uio.no/tekstlab for more information.

[3] The actual framework is that of Constraint Grammar (Karlsson et al., 1995), and the analysis is underspecified as the nodes are labelled only with their dependency function, e.g. subject or prepositional object, and their immediate heads are not uniquely determined.

[4] The tagger works in an eliminative fashion, so tokens may bear two or more tags when they have not been fully disambiguated.

	SUBJ		OBJ		GEN		PASS		ANAAN		ANAIN		REFL	
Class	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A	0.14	0.05	0.11	0.03	0.04	0.02	0.006	0.005	0.009	0.006	0.003	0.003	0.005	0.0008
I	0.07	0.03	0.23	0.10	0.02	0.03	0.002	0.002	0.003	0.002	0.006	0.003	0.001	0.0008

Table 1: Mean relative frequencies and standard deviation for each class (A(nimate) vs. I(nanimate)) from feature extraction (SUBJ=Transitive Subject, OBJ=Object, GEN=Genitive *-s*, PASS=Passive *by*-phrase, ANAAN=Anaphoric reference by animate pronoun, ANAIN=Anaphoric reference by inanimate pronoun, REFL=Anaphoric reference by reflexive pronoun).

Feature	% Accuracy
SUBJ	85.0
OBJ	72.5
GEN	72.5
PASS	62.5
ANAAN	67.5
ANAIN	50.0
REFL	82.5

Table 2: Accuracy for the individual features using leave-one-out training and testing

Features used	Feature Not Used	% Accuracy
1. SUBJ OBJ GEN PASS ANAAN ANAIN REFL		87.5
2. OBJ GEN PASS ANAAN ANAIN REFL	SUBJ	85.0
3. SUBJ GEN PASS ANAAN ANAIN REFL	OBJ	87.5
4. SUBJ OBJ PASS ANAAN ANAIN REFL	GEN	85.0
5. SUBJ OBJ GEN ANAAN ANAIN REFL	PASS	82.5
6. SUBJ OBJ GEN PASS ANAIN REFL	ANAAN	82.5
7. SUBJ OBJ GEN PASS ANAAN REFL	ANAIN	87.5
8. SUBJ OBJ GEN PASS ANAAN ANAIN	REFL	75.0
9. OBJ PASS ANAAN ANAIN	SUBJ GEN REFL	77.5

Table 3: Accuracy for all features and 'all minus one' using leave-one-out training and testing

rather rare. This is in particular true for the passive feature and the anaphoric features ANAAN, ANAIN and REFL. There is also quite a bit of variation in the data (represented by the standard deviation for each class-feature combination), a property which is to be expected as all the features represent approximations of animacy, gathered from an automatically annotated, possibly quite noisy, corpus. Even so, the features all express a difference between the two classes in terms of distributional properties; the difference between the mean feature values for the two classes range from double to five times the lowest class value.

3.1 Experiment 1

Based on the data collected on seven different features for our 40 nouns, a set of feature vectors are constructed for each noun. They contain the relative frequencies for each feature along with the name of the noun and its class (animate or inanimate). Note that the vectors do not contain the mean values presented in Table 1 above, but rather the individual relative frequencies for each noun.

The experimental methodology chosen for the classification experiments is similar to the one described in Merlo and Stevenson (2001) for verb classification. We also make use of leave-one-out training and testing of the classifiers and the same software package for decision tree learning, C5.0 (Quinlan, 1998), is employed. In addition, all our classifiers employ the *boosting* option for constructing classifiers (Quinlan, 1993). For calculation of the statistical significance of differences in the performance of classifiers tested on the same data set, McNemar's test is employed.

Table 2 shows the performance of each individual feature in the classification of animacy. As we can see, the performance of the features differ quite a bit, ranging from mere baseline performance (ANAIN) to a 70% improvement of the baseline (SUBJ). The first line of Table 3 shows the performance using all the seven features collectively where we achieve an accuracy of 87.5%, a 75% improvement of the baseline. The SUBJ, GEN and REFL features employed individually are the best performing individual features and their classification performance do not differ significantly from the performance of the combined classifier, whereas the rest of the individual features do (at the p<.05 level).

The subsequent lines (2-8) of Table 3 show the accuracy results for classification using all features except one at a time. This provides an indication of the contribution of each feature to the classification task. In general, the removal of a feature causes a 0% - 12.5% deterioration of results, however, only the difference in performance caused by the removal of the REFL feature is significant (at the p<0.05 level). Since this feature is one of the best performing features individually, it is not surprising that its removal causes a notable difference in performance. The removal of the

ANAIN feature, on the other hand, does not have any effect on accuracy whatsoever. This feature was the poorest performing feature with a baseline, or mere chance, performance. We also see, however, that the behaviour of the features in combination is not strictly predictable from their individual performance, as presented in table 2. The SUBJ, GEN and REFL features were the strongest features individually with a performance that did not differ significantly from that of the combined classifier. However, as line 9 in Table 3 shows, the classifier as a whole is not solely reliant on these three features. When they are removed from the feature pool, the performance (77.5% accuracy) does not differ significantly ($p<.05$) from that of the classifier employing all features collectively.

4 Data sparseness and back-off

The classification experiments reported above impose a frequency constraint (absolute frequencies >1000) on the nouns used for training and testing, in order to study the interaction of the different features without the effects of sparse data. In the light of the rather promising results from these experiments, however, it might be interesting to further test the performance of our features in classification as the frequency constraint is gradually relaxed.

To this end, three sets of common nouns each counting 40 nouns (20 animate and 20 inanimate nouns) were randomly selected from groups of nouns with approximately the same frequency in the corpus. The first set included nouns with an absolute frequency of 100 +/-20 ($\sim$100), the second of 50+/-5 ($\sim$50) and the third of 10+/-2 ($\sim$10). Feature extraction followed the same procedure as in experiment 1, relative frequencies for all seven features were computed and assembled into feature vectors, one for each noun.

4.1 Experiment 2: Effect of sparse data on classification

In order to establish how much of the generalizing power of the old classifier is lost when the frequency of the nouns is lowered, an experiment was conducted which tested the performance of the old classifier, i.e. a classifier trained on all the more frequent nouns, on the three groups of less frequent nouns. As we can see from the first column in Table 4, this resulted in a clear deterioration of results, from our earlier accuracy of 87.5% to new accuracies ranging from 70% to 52.5%, barely above the baseline. Not surprisingly, the results decline steadily as the absolute frequency of the classified noun is lowered.

Accuracy results provide an indication that the classification is problematic. However, it does not indicate what the damage is to each class as such. A confusion matrix is in this respect more informative. Confusion matrices for the classification of the three groups of nouns, $\sim$100, $\sim$50 and $\sim$10, are provided in table 5. These clearly indicate that it is the animate class which suffers when data becomes more sparse. The percentage of misclassified animate nouns drop drastically from 50% at $\sim$100 to 80% at $\sim$50 and finally 95% at $\sim$10. The classification of the inanimate class remains pretty stable throughout. The fact that a majority of our features (SUBJ, GEN, PASS, ANAAN and REFL) target animacy, in the sense that a higher proportion of animate than inanimate nouns exhibit the feature, gives a possible explanation for this. As data gets more limited, this differentiation becomes harder to make, and the animate feature profiles come to resemble the inanimate more and more. Because the inanimate nouns are expected to have low proportions (compared to the animate) for all these features, the data sparseness is not as damaging. In order to examine the effect on each individual feature of the lowering of the frequency threshold, we also ran classifiers trained on the high frequency nouns with only individual features on the three groups of new nouns. These results are depicted in Table 4. In our earlier experiment, the performance of a majority of the individual features (OBJ, PASS, ANAAN, ANAIN) was significantly worse (at the $p<0.05$ level) than the performance of the classifier including all the features. Three of the individual features (SUBJ, GEN, REFL) had a performance which did not differ significantly from that of the classifier employing all the features in combination.

As the frequency threshold is lowered, however, the performance of the classifiers employing all features and those trained only on individual features become more similar. For the $\sim$100 nouns, only the two anaphoric features ANAAN and the reflexive feature REFL, have a performance that differs significantly ($p<0.05$) from the classifier employing all features. For the $\sim$50 and $\sim$10 nouns, there are no significant differences between the classifiers employing individual fea-

Freq	All	SUBJ	OBJ	GEN	PASS	ANAAN	ANAIN	REFL
∼100	70.0	75.0	80.0	72.5	65.0	52.5	50.0	60.0
∼50	57.5	75.0	62.5	77.5	62.5	57.5	50.0	55.0
∼10	52.5	52.5	65.0	50.0	57.5	50.0	50.0	50.0

Table 4: Accuracy obtained when employing the old classifier on new lower-frequency nouns with leave-one-out training and testing: all and individual features

∼100 nouns			∼50 nouns			∼10 nouns		
(a)	(b)	← classified as	(a)	(b)	← classified as	(a)	(b)	← classified as
10	10	(a) class animate	4	16	(a) class animate	1	19	(a) class animate
2	18	(b) class inanimate	1	19	(b) class inanimate		20	(b) class inanimate

Table 5: Confusion matrices for classification of lower frequency nouns with old classifier

tures only and the classifiers trained on the feature set as a whole. This indicates that the combined classifiers no longer exhibit properties that are not predictable from the individual features alone and they do not generalize over the data based on the combinations of features.

In terms of accuracy, a few of the individual features even outperform the collective result. On average, the three most frequent features, the SUBJ, OBJ and GEN features, improve the performance by 9.5% for the ∼100 nouns and 24.6% for the ∼50 nouns. For the lowest frequency nouns (∼10) we see that the object feature alone improves the result by almost 24%, from 52.5% to 65 % accuracy. In fact, the object feature seems to be the most stable feature of all the features. When examining the means of the results extracted for the different features, the object feature is the feature which maintains the largest difference between the two classes as the frequency threshold is lowered. The second most stable feature in this respect is the subject feature.

The group of experiments reported above shows that the lowering of the frequency threshold for the classified nouns causes a clear deterioration of results in general, and most gravely when all the features are employed together.

4.2 Experiment 3: Back-off features

The three most frequent features, the SUBJ, OBJ and GEN features, were the most stable in the two experiments reported above and had a performance which did not differ significantly from the combined classifiers throughout. In light of this we ran some experiments where all possible combinations of these more frequent features were employed. The results for each of the three groups of nouns is presented in Table 6. The exclusion of the less frequent features has a clear positive effect on the accuracy results, as we can see in table 6. For the ∼100 and ∼50 nouns, the performance has improved compared to the classifier trained both on all the features and on the individual features. The classification performance for these nouns is now identical or only slightly worse than the performance for the high-frequency nouns in experiment 1. For the ∼10 group of nouns, the performance is, at best, the same as for all the features and at worse fluctuating around baseline.

In general, the best performing feature combinations are SUBJ&OBJ&GEN and SUBJ&OBJ . These two differ significantly (at the p<.05 level) from the results obtained by employing all the features collectively for both the ∼100 and the ∼50 nouns, hence indicate a clear improvement. The feature combinations both contain the two most stable features - one feature which targets the animate class (SUBJ) and another which target the inanimate class (OBJ), a property which facilitates differentiation even as the marginals between the two decrease.

It seems, then, that backing off to the most frequent features might constitute a partial remedy for the problems induced by data sparseness in the classification. The feature combinations SUBJ&OBJ&GEN and SUBJ&OBJ both significantly improve the classification performance and actually enable us to maintain the same accuracy for both the ∼100 and ∼50 nouns as for the higher frequency nouns, as reported in experiment 1.

Freq	SUBJ&OBJ&GEN	SUBJ&OBJ	SUBJ&GEN	OBJ&GEN
∼100	87.5	87.5	77.5	85.0
∼50	82.5	90.0	70.0	77.5
∼10	57.5	50.0	50.0	47.5

Table 6: Accuracy obtained when employing the old classifier on new lower-frequency nouns: combinations of the most frequent features

4.3 Experiment 4: Back-off classifiers

Another option, besides a back-off to more frequent features in classification, is to back off to another classifier, i.e. a classifier trained on nouns with a similar frequency. An approach of this kind will attempt to exploit any group similarities that these nouns may have in contrast to the mores frequent ones, hopefully resulting in a better classification.

In this experiment classifiers were trained and tested using leave-one-out cross-validation on the three groups of lower frequency nouns and employing individual, as well as various other feature combinations. The results for all features as well as individual features are summarized in Table 7. As we can see, the result for the classifier employing all the features has improved somewhat compared to the corresponding classifiers in experiment 3 (as reported above in Table 4) for all our three groups of nouns. This indicates that there is a certain group similarity for the nouns of similar frequency that is captured in the combination of the seven features. However, backing off to a classifier trained on nouns that are more similar frequency-wise does not cause an improvement in classification accuracy. Apart from the SUBJ feature for the ∼100 nouns, none of the other classifiers trained on individual or all features for the three different groups differ significantly ($p<.05$) from their counterparts in experiment 3.

As before, combinations of the most frequent features were employed in the new classifiers trained and tested on each of the three frequency-ordered groups of nouns. In the terminology employed above, this amounts to a backing off both classifier- and feature-wise. The accuracy measures obtained for these experiments are summarized in table 8. For these classifiers, the backed off feature combinations do not differ significantly (at the $p<.05$ level) from their counterparts in experiment 3, where the classifiers were trained on the more frequent nouns with feature back-off.

5 Conclusion

The above experiments have shown that the classification of animacy for Norwegian common nouns is achievable using distributional data from a morphosyntactically annotated corpus. The chosen morphosyntactic features of animacy have proven to differentiate well between the two classes. As we have seen, the transitive subject, direct object and morphological genitive provide stable features for animacy even when the data is sparse(r). Four groups of experiments have been reported above which indicate that a reasonable remedy for sparse data in animacy classification consists of backing off to a smaller feature set in classification. These experiments indicate that a classifier trained on highly frequent nouns (experiment 1) backed off to the most frequent features (experiment 3) sufficiently capture generalizations which pertain to nouns with absolute frequencies down to approximately fifty occurrences and enables an unchanged performance approaching 90% accuracy.

Even so, there are certainly still possibilities for improvement. As is well-known, singleton occurrences of nouns abound and the above classification method is based on data for lemmas, rather than individual instances or tokens. One possibility to be explored is token-based classification of animacy, possibly in combination with a lemma-based approach like the one outlined above.

Such an approach might also include a finer subdivision of the nouns. We have chosen to classify along a binary dimension, however, it might be argued that this is an artificial dichotomy. (Zaenen et al., 2004) describe an encoding scheme for the manual encoding of animacy information in part of the English Switchboard corpus. They make a three-way distinction between human, other animates, and inanimates, where the ‘other animates’ category describes a rather heterogeneous group of entities: organisations, animals, intelligent machines and vehicles. However, what these seem to have in common is that they may all be construed linguistically as ani-

Freq	All	SUBJ	OBJ	GEN	PASS	ANAAN	ANAIN	REFL
∼100	85.0	52.5	87.5	65.0	70.0	50.0	57.5	50.0
∼50	77.5	77.5	75.0	75.0	50.0	50.0	50.0	50.0
∼10	52.5	50.0	62.5	50.0	50.0	50.0	50.0	50.0

Table 7: Accuracy obtained when employing a new classifier on new lower-frequency nouns: all and individual features

Freq	SUBJ&OBJ&GEN	SUBJ&OBJ	SUBJ&GEN	OBJ&GEN
∼100	85.0	85.0	67.5	82.5
∼50	75.0	80.0	75.0	70.0
∼10	62.5	62.5	50.0	62.5

Table 8: Accuracy obtained when employing a new classifier on new lower-frequency nouns: combinations of the most frequent features

mate beings, even though they, in the real world, are not. Interestingly, the two misclassified inanimate nouns in experiment 1, were *bil* 'car' and *fly* 'air plane', both vehicles. A token-based approach to classification might better capture the context-dependent and dual nature of these types of nouns.

Automatic acquisition of animacy in itself is not necessarily the primary goal. By testing the use of acquired animacy information in various NLP applications such as parsing, generation or coreference resolution, we might obtain an extrinsic evaluation measure for the usefulness of animacy information. Since very frequent nouns are usually well described in other lexical resources, it is important that a method for animacy classification is fairly robust to data sparseness. This paper suggests that a method based on seven morphosyntactic features, in combination with feature back-off, can contribute towards such a classification.

References

Judith Aissen. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory*, 21:435–483.

Joan Bresnan, Anna Cueni, Tatiana Nikitina and Harald Baayen. 2005. Predicting the Dative Alternation. To appear in Royal Netherlands Academy of Science Workshop on Foundations of Interpretation proceedings.

Shipra Dingare. 2001. The effect of feature hierarchies on frequencies of passivization in English. M.A. Thesis, Stanford University.

David Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619.

John Hale and Eugene Charniak. 1998. Getting Useful Gender Statistics from English Text. Technical Report, Comp. Sci. Dept. at Brown University, Providence, Rhode Island.

Christiane Fellbaum, editor. 1998. *WordNet, an electronic lexical database*. MIT Press.

Fred Karlsson and Atro Voutilainen and Juha Heikkilä and Atro Anttila. 1995. *Constraint Grammar: A language-independent system for parsing unrestricted text.* Mouton de Gruyer.

Hanjung Lee. 2002. Prominence Mismatch and Markedness Reduction in Word Order. *Natural Language and Linguistic Theory*, 21(3):617–680.

Paola Merlo and Suzanne Stevenson. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–408.

Constantin Orăsan and Richard Evans. 2001. Learning to Identify Animate References. in *Proceedings of the Workshop on Computational Natural Language Learning*, ACL-2001.

Lilja Øvrelid. 2004. Disambiguation of syntactic functions in Norwegian: modeling variation in word order interpretations conditioned by animacy and definiteness. in Fred Karlsson (ed.): *Proceedings of the 20th Scandinavian Conference of Linguistics*, Helsinki.

J. Ross Quinlan. 1998. *C5.0: An Informal Tutorial.* http://www.rulequest.com/see5-unix.html.

J. Ross Quinlan. 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, Series in Machine Learning.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor and Tom Wasow. 2004. Animacy encoding in English: why and how. in D. Byron and B. Webber (eds.): *Proceedings of ACL Workshop on Discourse Annotation*, Barcelona.