Animacy classification based on morphosyntactic corpus frequencies: some experiments with Norwegian nouns

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

Abstract

This paper presents results from experiments in automatic classification of animacy for a set of Norwegian nouns using decision-tree classifiers. The method makes use of seven linguistically motivated syntactic and morphological features of these nouns extracted from an automatically annotated corpus of Norwegian. It classifies unseen nouns and achieves an accuracy of 90% under 10-fold cross-validation, as well as single hold-out training and testing.

1 Introduction

Animacy is 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. It is also 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 applications ranging from coreference resolution to parsing and generation. This article presents experiments on automatic classification of the animacy of unseen nouns, inspired by a method for verb classification, as presented in Merlo and Stevenson (2001). The experiments make use of statistical distributions of a set of linguistically motivated morphosyntactic cues for animacy, as gathered from an automatically annotated corpus of Norwegian.

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). Within the typological field of linguistics, the notion of markedness has been of great importance, underlying much of the cross-linguistic comparative work performed there. This notion is based on “asymmetrical or unequal grammatical properties of otherwise equal linguistic elements” (Croft, 2003), and is linked to, among others, the relative frequency of a given structure. An unmarked structure will typically be more frequent than its marked counterpart and relatedly, figure in a greater number of linguistic contexts (Croft, 2003). So-called prominence hierarchies figure frequently in typological descriptions. These hierarchies express the relative prominence of a structure, and incorporate the relativity of markedness into the theory. The notion of prominence has been linked to several properties such as most likely as topic, agent, most available referent etc. Among the hierarchies established in typological literature are those of syntactic functions, animacy and thematic role (Croft, 2003; Aissen, 2003):
Syntactic function: Subject > Object

Animacy: Human > Animate > Inanimate

Thematic Role: Agent > Patient

A key generalisation or tendency regarding these hierarchies is that features placed high on one hierarchy tend to attract other prominent or high-placed features; 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, in 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. Orasun and Evans (2001) make use of hyponym-relations taken from the WordNet resource in order to classify animate referents. However, such a method is clearly restricted to languages for which large scale lexical resources, such as the WordNet, 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. A key question in the following thus becomes whether a similar method may be applied to the task of animacy classification based on linguistically motivated cues extracted from a corpus.

2 Morphosyntactic features of animacy

The method for verb classification described in Merlo and Stevenson (2001) makes use of a training set consisting of relative frequency data for each verb in a certain class, which summarise its overall count for a certain feature. The task is thus a matter of classifying an unseen verb based on properties of all the instances of this verb (the lemma), rather than classifying individual instances by themselves.

What features, then, can be exploited as cues for the animacy of a noun? As mentioned above, animacy is highly correlated with a number of other linguistic concepts, such as agentivity, topicality and discourse salience. One would expect marked configurations along these dimensions, e.g. animate objects or agentive inanimates, to be 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. It is important, however, to stress that these features only provide approximations of animacy, which, hopefully, lead to observable distributional differences between nouns.

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, i.e. pronouns and proper nouns as well, 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.

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

---

1Simple transitive sentences are main sentences which include a simple transitive main verb, i.e. no auxiliaries or modals. Due to the fact that Norwegian is a V2-language which does not case mark nouns, and allows for both SVO and OVS word order, sentences like these are syntactically/functionally ambiguous (Øvrelid, 2004).
agent, according to Dowty (1991). However, if no additional information on argument structure for verbs is to be assumed, other ways of approximating the agentivity of a noun must be arrived at. One possibility is to use the passive construction, or rather the property of being expressed as the demoted agent in a passive construction. As is well known, transitive constructions tend to passivise better (hence more frequently) if the demoted subject bears a prominent thematic role, preferably agent, rather than a role less prominent on the thematic role hierarchy. A prediction to be tested is therefore whether the relative frequency with which a noun occurs in a passive by-phrase, is an indicator of its animacy.

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’. This is one reason why information regarding the animacy of a noun can be helpful in the task of coreference resolution. However, in this context it might be interesting to make use of an approximation of anaphoric reference in determining the animacy of a noun.

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\(^2\). The third person Norwegian reflexive pronoun seg ‘him/her/itself’ does not, however, differentiate its antecedent along the animacy dimension. 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\(^3\) of the construction might favour an animate subject.

Finally, when it comes to morphological properties, animate nouns are not marked specifically as such in Norwegian. Common nouns are marked for number and definiteness, as well as having an inherent gender (masculine, feminine or neuter). There is no extensive case system for common nouns 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. However, this is certainly not always the case, semantic relationships such as a whole-part relation as in bilens hjul ‘the car’s wheel’ or a quantificational meaning as in en times arbeide ‘an hour’s work’ etc. also commonly occur. An alternative construction to the s-genitive in Norwegian is constructed by inserting the possessive pronoun sin between the possessor and the possessed, as in mannen sin bil ‘the man’s car’. The sin-genitive is to be preferred when the relation is one of possession (Faarlund et al., 1997), hence often involving an animate possessor. Generally then, the frequency with which a noun occurs as a modifier might provide an indicator of the animacy of that noun.

3 Feature extraction

In order to train a classifier to distinguish between animate and inanimate nouns, training data consisting of distributional statistics have to be extracted from a corpus. Appropriate approximations of the linguistically motivated features described above also have to be constructed. For this end, a 15

\(^2\)Norwegian has two types of reflexive constructions - a simple reflexive seg and a complex reflexive seg selv. The difference between these two have traditionally been viewed as based on locality - the complex reflexive is bound locally, whereas the simple one is bound non-locally. However, as Lødrup (1999) shows, this represents an idealization which does not hold up against real data. In particular, the simple reflexive pronoun is far more versatile than previously assumed and may very well be bound locally. We will therefore include both the simple and complex reflexives in our study.

\(^3\)Reflexives in Norwegian do not necessarily express an agentive event, and may be employed, for instance in medial constructions. However, with regards to productivity one would assume the agentive reflexives to be predominant, hence implying an animate subject.
million word version of the Oslo Corpus, a corpus of Norwegian texts of approximately 18.5 million
words, was employed\(^4\). The corpus is morphosyntactically annotated and assigns an underspecified
dependency-style analysis to each sentence\(^5\).

As training data for the classifier, a set of forty nouns were chosen - twenty animate and twenty
inanimate nouns, exemplified in (1a) and (1b) respectively:

‘boy’, leder ‘leader’, lege ‘doctor’
‘property’, fly ‘airplane’

The corpus study of Norwegian simple transitives mentioned earlier, showed that nouns expressing
animate beings aside from humans (e.g. animals) are very infrequent (0.0025%) in the corpus (Øvre-
lid, 2004), and these were therefore not focused on in the following. Also, as some of the features
employed were assumed to be quite rare, e.g. anaphoric pronominal reference or passive by-phrases,
a cut-off point with regards to frequency was maintained throughout the study; all nouns had at least
one thousand occurrences in the corpus.

3.1 Feature approximation

For each noun, relative frequencies for the different morphosyntactic features described above were
computed from the corpus.

**Subjects and objects** For transitive subjects, we extracted the number of instances where the noun
in question was unambiguously tagged as subject and followed by a finite verb and an unambiguously
tagged object\(^6\). The frequency of direct objects 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 relative clauses and infinitival clauses.

**Passive** As we remember, another context where animate nouns might be predominant is in the by-
phrase expressing the demoted agent of a passive verb. 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.

**Anaphoric reference** With regards to the property of anaphoric reference by personal pronouns, the
extraction was bound to be a bit trickier. 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

\(^4\)The corpus is freely available for research purposes, see http://www.hf.uio.no/tekstlab for more information.
\(^5\)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 function, e.g. subject or prepositional object, and not its immediate head or dependent(s).
\(^6\)The tagger works in an eliminative fashion, so tokens may bear two or more tags when they have not been fully
disambiguated.
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). (2) below shows an authentic example from the results for the noun mann ‘man’ taken from the Oslo Corpus:

(2)   Mannen, ble pågrepet etter tre kvarters dramatisk biljakt. Han, var beruset . . .

The man, was apprehended after a three-quarter long car chase. He, was intoxicated . . .

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) the animate personal pronouns han ‘he’, hun ‘she’ or de ‘they’, and (ii) the inanimate personal pronouns den ‘it-MASC’ or det ‘it-NEUT’. Now, the 3. person plural de ‘they’ is not strictly an indicator of animacy as it may refer to both animate and inanimate referents, as in English. However, Merlo and Stevenson (2001) claim that, in English, this plural pronoun usually refers to animate entities and in a selection of 100 occurrences of this pronoun, they found that 76% of these had an animate antecedent 7. We therefore make the same assumption for Norwegian, although this is a possible source for mistakes in the counts, we assume that the general distribution of instances will still differentiate with regards to animacy. Another possible source for mistakes in the relative frequencies lies in the fact that we cannot assume to have knowledge regarding the natural gender of our training nouns. As this often does not coincide with the grammatical gender of a noun in Norwegian, we must therefore count all occurrences of the personal pronouns following a noun without controlling for agreement with respect to natural gender.

For the inanimate pronouns, the neuter form det ‘it-NEUT’ is problematic as this is also the expletive subject form. This pronoun therefore often initiates a sentence, but has a clearly non-referential function. However, as the distinction between expletive and pronominal subjects is not annotated for in the corpus, we will count all occurrences of this pronoun when it initiates a subsequent sentence. Another possibility would have been to exclude all occurrences of det ‘it’ from the counts, with the consequence that this test would be inapplicable for the set of neuter nouns in our training set (8 nouns).

Reflexive    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.

Genitive -s    This feature simply contains relative frequencies of the occurrence of each noun with genitive case marking, i.e. the suffix -s. As mentioned earlier, the Norwegian sin-genitive is usually preferred with animate possessors and might provide a useful feature of animacy. Unfortunately, however, this construction is far too rare and yielded zero occurrences for a large number of the nouns (both animate and inanimate), hence was abandoned. 8

7Merlo and Stevenson (2001) make use of personal pronouns as indicators of argument structure for a verb. If it often occurs with an animate pronominal subject, they assume that the verb distributes agentive role to its subject.
8The sin-genitive is generally a property of spoken rather than written Norwegian, although one can find examples in more informal writing (Faarlund et al., 1997).
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, ANAANIM=Anaphoric reference by animate pronoun, ANAINAN=Anaphoric reference by inanimate pronoun, REFL=Anaphoric reference by reflexive pronoun).

<table>
<thead>
<tr>
<th>Class</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.14</td>
<td>0.05</td>
<td>0.11</td>
<td>0.03</td>
<td>0.04</td>
<td>0.02</td>
<td>0.006</td>
<td>0.005</td>
<td>0.009</td>
<td>0.006</td>
<td>0.003</td>
<td>0.003</td>
<td>0.005</td>
<td>0.0008</td>
</tr>
<tr>
<td>I</td>
<td>0.07</td>
<td>0.03</td>
<td>0.23</td>
<td>0.10</td>
<td>0.02</td>
<td>0.03</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
<td>0.002</td>
<td>0.006</td>
<td>0.003</td>
<td>0.001</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

3.2 Results

The mean relative frequencies for each class - animate and inanimate - are presented in table 1. The standard deviation for each class and feature is provided alongside the mean. 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 rather rare. This is in particular true for the passive feature and the anaphoric features ANAANIM, ANAINAN and REFL. This is perhaps not so surprising, however, the question is whether these features express the relevant distinction although they are sparse. When examining the features in table 1. this certainly seems to be the case; the difference between the mean feature values for the two classes range from double to five times the lowest class value.

Another point is that the values for the features that one would expect to be quite frequent, e.g. SUBJ and OBJ only range from about 3% to 14% of all occurrences. The reason for this is that the search patterns designed to extract the counts require the subjects and objects in question to be unambiguously tagged. However, all subjects and objects of the simple transitive sentences mentioned earlier are tagged as being both subjects and objects on account of their functional ambiguity. This means that the transitive subjects and objects that are counted are only those that occur in a syntactic environment which clearly disambiguates them functionally.

3.3 Other features

The features and the mean values presented in table 1. are the features that were actually employed in the experiments. However, several other features were also extracted, which did not exhibit the required distinction. The indirect object of the ditransitive double object construction expresses the thematic role of recipient and is known for displaying an “animacy effect” (Bresnan et al., 2005). An approximation was therefore attempted for the feature of indirect objects in ditransitive constructions. This however, turned out to yield a result that was contrary to the expected results. The mean result for the animate class was 0.007%, whereas the inanimate class had the higher count of 0.008%. However, a quick look at some of the extracted sentences shows that the tagger’s automatic analysis of indirect objects contains a lot of errors. This feature was therefore abandoned and not included in the classification experiments.

Due to the mentioned correlation between animacy and discourse salience or topicality, the morpho-
logical definiteness of the animate vs. inanimate nouns was also recorded. One might assume that a topical element is also definite. However, this feature only yielded a mean 1% difference between the categories, hence was also abandoned. One possible reason for this is that morphological and semantic definiteness do not necessarily overlap, hence the crude measure of morphological definiteness might not be able to fully capture the semantic definiteness of a given noun.

4 Experiments

The experimental methodology chosen for the classification experiments for animacy is pretty much identical to the one described in Merlo and Stevenson (2001) for verb classification. The same software package for decision tree learning, C5.0 (Quinlan, 1993), has also been employed.

A decision tree is a classification model which relates a set of predefined classes with properties of the instances to be classified. In this case we wish to classify Norwegian common nouns along the binary dimension of animacy, i.e. animate vs. inanimate. The properties in question are the morphosyntactic features on which we have gathered data. Classification using a decision tree proceeds by means of a set of weighted, disjunctive tests which at each step (node) in the process assigns an appropriate test to an input, and which proceeds along one of its branches, representing possible outcomes of the test.

4.1 Training and testing methodology

Based on the data collected on seven different features for our 40 nouns, a set of feature vectors may be 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.

Merlo and Stevenson (2001) experiment with two different methodologies for training and testing the decision tree classifier(s) - 10-fold cross-validation and single hold-out. They have in common that the reported results from both are on unseen test data, i.e. data that are not part of the training set, however they are also different in the sense that their results contribute slightly different information (Merlo and Stevenson, 2001). 10-fold cross-validation has the advantage that it reports an average accuracy result for the entire data set, whereas single hold-out provides more specific results regarding which classes and nouns are misclassified, thus forming the base for further analysis. For our experiments in animacy classification both methods for training and testing were employed. As the task is a binary classification task, we assume a baseline accuracy of at best 50%.

4.2 Results

4.2.1 10-fold cross-validation

As the 10-fold cross validation method reports accuracy measures averaged over all runs, it facilitates the testing of different features and their individual contribution to the classification task.

Table 2. shows the performance of each individual feature in the classification of animacy. As we can see, the features perform quite well, ranging from mere baseline performance (ANAINAN) to a 65% improvement of the baseline (REFL).

---

10The C5.0 software package may be downloaded from http://www.rulequest.com/.
Table 2: Accuracy for the individual features using 10-fold cross validation

<table>
<thead>
<tr>
<th>Feature</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJ</td>
<td>77.5</td>
</tr>
<tr>
<td>OBJ</td>
<td>72.5</td>
</tr>
<tr>
<td>GEN</td>
<td>75.0</td>
</tr>
<tr>
<td>PASS</td>
<td>67.5</td>
</tr>
<tr>
<td>ANAANIM</td>
<td>70.0</td>
</tr>
<tr>
<td>ANAINAN</td>
<td>50.0</td>
</tr>
<tr>
<td>REFL</td>
<td>82.5</td>
</tr>
</tbody>
</table>

Table 3: Accuracy for all features and ‘all minus one’ using 10-fold cross validation

<table>
<thead>
<tr>
<th>Features used</th>
<th>Feature Not Used</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. SUBJ OBJ GEN PASS ANAANIM ANAINAN REFL</td>
<td></td>
<td>90.0</td>
</tr>
<tr>
<td>2. OBJ GEN PASS ANAANIM ANAINAN REFL</td>
<td>SUBJ</td>
<td>85.0</td>
</tr>
<tr>
<td>3. SUBJ GEN PASS ANAANIM ANAINAN REFL</td>
<td>OBJ</td>
<td>90.0</td>
</tr>
<tr>
<td>4. SUBJ OBJ PASS ANAANIM ANAINAN REFL</td>
<td>GEN</td>
<td>85.0</td>
</tr>
<tr>
<td>5. SUBJ OBJ GEN ANAANIM ANAINAN REFL</td>
<td>PASS</td>
<td>82.5</td>
</tr>
<tr>
<td>6. SUBJ OBJ GEN PASS ANAINAN REFL</td>
<td>ANAANIM</td>
<td>77.5</td>
</tr>
<tr>
<td>7. SUBJ OBJ GEN PASS ANAANIM REFL</td>
<td>ANAINAN</td>
<td>85.0</td>
</tr>
<tr>
<td>8. SUBJ OBJ GEN PASS ANAANIM ANAINAN REFL</td>
<td>REFL</td>
<td>77.5</td>
</tr>
</tbody>
</table>

The first line of table 3 shows the performance using all the seven features where we achieve an accuracy of 90%, an 80% improvement of the baseline. The subsequent lines 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. The removal of the transitive object feature in line 3 does not affect the accuracy of the classifier at all and this feature is therefore redundant. Removal of the transitive subject feature on the other hand, causes a 5% deterioration of accuracy. In general, the removal of a feature causes a 0% - 12.5% deterioration of results. We also see that the behaviour of the features in combination is not strictly predictable from their individual performance, as presented in table 2. For instance, the removal of the ‘anaphoric reference with animate pronoun’ feature (ANAINAN) has the most severe effect on the result, but is one of the poorest performing features on its own.

4.2.2 Single hold-out

As mentioned earlier, the single hold-out method has the advantage of providing results regarding the individual classes as well as individual nouns. Because it facilitates class-wise comparisons, a F score may also be computed, which relates true/false negatives and positives for each class\(^{11}\). In this case, the simple accuracy (number of correct classifications / all classifications) and the F score are identical when all the features are employed, as shown in line 1 in table 4. The number of misclassifications are symmetrical - two nouns are misclassified for each class (hence two were deemed false positives for the opposing class). As we see, then, the result for all the features combined is the same as for the 10-fold cross validation method - 90% accuracy. For the ‘all minus one’ feature sets the results are not completely identical to that of the 10-fold cross-validation method. The accuracy measures differ somewhat, showing that the learner is slightly sensitive to the exact makeup of the test sets. Also, the removal of the object feature here shows a 2.5% deterioration of results, caused by the misclassification of one extra noun, in contrast to the cross-validation results. It seems that the SUBJ and OBJ features are somewhat overlapping. A possible reason for this, is that the information contained in the subject feature actually implies a direct object, only nouns that were unambiguously tagged as subject and followed by an unambiguous object were counted. As we remember, the object feature, on the other hand, does not demand a realized subject.

The balanced F score provided for each class in table 4 provides us with a more detailed picture of

\(^{11}\)We make use of a balanced F score: \(2PR/P+R\) (P=precision: true positives / true positives + false positives, R=recall: true positives / true positives + false negatives) (Merlo and Stevenson, 2001)
Table 4: Accuracy and balanced F-score per class for all features and ‘all minus one’ using single hold-out method.

<table>
<thead>
<tr>
<th>Features Used</th>
<th>Not Used</th>
<th>% Acc</th>
<th>% F Anim</th>
<th>% F Inan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. SUBJ OBJ GEN PASS ANAANIM ANAinan REFL</td>
<td>90.0</td>
<td>90.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>2. OBJ GEN PASS ANAANIM ANAinan REFL</td>
<td>SUBJ</td>
<td>85.0</td>
<td>84.2</td>
<td>85.7</td>
</tr>
<tr>
<td>3. SUBJ GEN PASS ANAANIM ANAinan REFL</td>
<td>OBJ</td>
<td>87.5</td>
<td>87.8</td>
<td>87.2</td>
</tr>
<tr>
<td>4. SUBJ OBJ PASS ANAANIM ANAinan REFL</td>
<td>GEN</td>
<td>85.0</td>
<td>85.0</td>
<td>85.0</td>
</tr>
<tr>
<td>5. SUBJ OBJ GEN ANAANIM ANAinan REFL</td>
<td>PASS</td>
<td>82.5</td>
<td>83.7</td>
<td>81.1</td>
</tr>
<tr>
<td>6. SUBJ OBJ GEN PASS ANAANIM REFL</td>
<td>ANAANIM</td>
<td>82.5</td>
<td>82.0</td>
<td>83.0</td>
</tr>
<tr>
<td>7. SUBJ OBJ GEN PASS ANAANIM REFL</td>
<td>ANAinan</td>
<td>85.0</td>
<td>84.2</td>
<td>85.7</td>
</tr>
<tr>
<td>8. SUBJ OBJ GEN PASS ANAANIM ANAinan</td>
<td>REFL</td>
<td>72.5</td>
<td>73.2</td>
<td>71.8</td>
</tr>
</tbody>
</table>

The effect of each feature on each class, as measured by the removal of this feature from the feature set. We are also informed of whether the effects of the features are as we predicted earlier. This seems largely to be the case; the removal of a feature which targets a specific class causes a lower F score for this class. For instance, the removal of SUBJ causes a lower F score for the animate class than the inanimate, indicating that a higher number of misclassifications related to the animate class took place.

In general, it seems fair to say that more features perform better. The nouns that are misclassified following the removal of a feature are seldom the same ones, hence underlining the need for all the features. An idiosyncratic behaviour of a noun in the light of one specific feature, is attributed less importance when the evidence from all the features is weighted in.

5 Discussion

The above experiments have shown that the classification of animacy for common nouns is achievable using distributional data from a syntactically annotated corpus. The results of the experiments are encouraging, and due to the fact that the features are linguistically motivated, hopefully also generalisable to a larger set of nouns. However, several questions remain open for future work.

We have chosen to classify along a binary dimension (animate vs. inanimate) with a relatively small set of nouns. Two related objections may be put forward at this point. Firstly, it might be argued that a binary dimension such as this is artificial and that there should be a finer subdivision of nouns. 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 describe 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 animate beings, even though they, in the real world, are not. Interestingly, the two misclassified inanimate nouns in our experiments, were *bil* ‘car’ and *fly* ‘airplane’, both vehicles. They exhibited a more agentive pattern which showed up in the transitive subject feature, the passive feature and the reflexive feature, in particular. However, they did not pattern completely with the animate nouns, they had a high object count and behaved like the inanimate nouns when it came to anaphoric pronouns. Secondly and related to the above, the choice of nouns in the experiment might be considered too limited. Had we chosen to include, for instance, nouns that have a metonymic use e.g. organisations, the classification into only two classes might have been less successful. However, we chose to start out with a binary classification in order to test
the viability of the method and its suitability for the classification task. Further experiments should probably enlarge the set of training nouns and also include an intermediate category, as proposed in Zaenen et al. (2004).

One might also ask whether the chosen features represent sufficient information to base classification on. As mentioned several times, the features only provide approximations of animacy by relying on related linguistic dimensions such as syntactic functions and thematic roles. Now, one of the misclassified animate nouns was *venn* ‘friend’, a clearly animate noun. However, according to our seven chosen features, this noun largely patterns with the inanimate nouns. When considering it, this probably also makes sense, as we are basing our classification of a real world property only on our linguistic depiction of it. A friend is probably more like a physical object in the sense that it is someone one likes/hates/loves or otherwise reacts to, rather than being an agent that acts upon its surroundings. Also, it is neutral with regards to natural gender, hence, probably less likely to be followed by a gender-specific pronoun. The features for anaphoricity therefore point more in the direction of inanimate nouns, as well.

In the long run, the acquisition of animacy by itself is not necessarily the only goal. By testing the use of acquired animacy information in various applications, such as parsing, generation or coreference resolution, the generalisations from linguistic studies regarding animacy effects in human language may be made use of, or even tested. A common problem in studies that rely on corpus data, however, is data sparseness. As mentioned earlier, the nouns experimented with in this study are all rather frequent (more than a thousand occurrences) in the corpus data. However, if one wishes to scale up the current approach, this problem will have to be dealt with. As the cut-off point of one thousand occurrences was rather randomly selected, an experiment was performed where the cut-off point was drastically reduced to approximately one hundred occurrences (+/- 20). Ten nouns of each class were attempted classified by i) the classifier trained earlier on the more frequent nouns by single hold-out ii) a new classifier trained and tested only on the twenty infrequent nouns by single hold-out. Both the experiments showed that most of the features are affected negatively by sparse data, a result which is not at all surprising, given that quite a few of the features are rather rare. Both the experiments yielded an accuracy of 65% when all seven features were employed, an error analysis showed that the majority of mistakes made by the two classifiers were misclassifications of animate nouns as inanimate. Due to the fact that the majority of our features (SUBJ, GEN, PASS, ANAANIM, REFL) require a higher relative frequency for animates than inanimates, it seems obvious that these are the nouns which will suffer most from sparse data in classification. However, both the classifiers improved their performance quite drastically when tested with only the transitive subject feature SUBJ - a more frequent feature which targets animate nouns. The old classifier applied to the twenty low frequency nouns here achieved an accuracy of 80%, whereas the classifier trained only on the new nouns achieved an accuracy of 85%. It thus seems that backing off to more frequent features when classifying lower frequency nouns might be a strategy worth investigating further. Experiments should also be performed in order to locate an appropriate cut-off point, as well as investigating further the interaction of our features.

In conclusion then, we have seen that the method for verb classification described in Merlo and Stevenson (2001) yield promising results for classification of animacy when applied to Norwegian common nouns using a set of seven linguistically motivated features of animacy. The theoretical predictions that the relative markedness of a construction along so-called prominence hierarchies would influence its frequency turned out to provide useful clues for an automatic classification task.
References


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


