Empirical evaluations of animacy annotation

Lilja Øvrelid
Department of Linguistics
University of Potsdam
Germany
lilja@ling.uni-potsdam.de

Abstract

This article presents empirical evaluations of aspects of annotation for the linguistic property of animacy in Swedish, ranging from manual human annotation, automatic classification and, finally, an external evaluation in the task of syntactic parsing. We show that a treatment of animacy as a lexical semantic property of noun types enables generalization over distributional properties of these nouns which proves beneficial in automatic classification and furthermore gives significant improvements in terms of parsing accuracy for Swedish, compared to a state-of-the-art baseline parser with gold standard animacy information.

1 Introduction

The property of animacy influences linguistic phenomena in a range of different languages, such as case marking (Aissen, 2003) and argument realization (Bresnan et al., 2005; de Swart et al., 2008), and has been shown to constitute an important factor in the production and comprehension of syntactic structure (Branigan et al., 2008; Weckerly and Kutas, 1999).[1] In computational linguistic work, animacy has been shown to provide important information in anaphora resolution (Orăsan and Evans, 2007), argument disambiguation (Dell'Orletta et al., 2005) and syntactic parsing in general (Øvrelid and Nivre, 2007).

The dimension of animacy roughly distinguishes between entities which are alive and entities which are not, however, other distinctions are also relevant and the animacy dimension is often viewed as a continuum ranging from humans to inanimate objects. Following Silverstein (1976) several animacy hierarchies have been proposed in typological studies, focusing on the *linguistic* category of animacy, i.e., the distinctions which are relevant for linguistic phenomena. An example of an animacy hierarchy, taken from (Aissen, 2003), is provided in (1):

(1) Human $>$ Animate $>$ Inanimate

Clearly, non-human animates, like animals, are not less animate than humans in a biological sense, however, humans and animals show differing linguistic behaviour.

Empirical studies of animacy require human annotation efforts, and, in particular, a well-defined annotation task. However, annotation studies of animacy differ distinctly in their treatment of animacy as a type or token-level phenomenon, as well as in terms of granularity of categories. The use of the annotated data as a computational resource furthermore poses requirements on the annotation which do not necessarily agree with more theoretical considerations. Methods for the induction of animacy information for use in practical applications require the resolution of issues of level of representation, as well as granularity.

This article addresses these issues through empirical and experimental evaluation. We present an in-depth study of a manually annotated data set which indicates that animacy may be treated as a lexical semantic property at the type level. We then evaluate this proposal through supervised machine learning of animacy information and focus on an in-depth error analysis of the resulting classifier, addressing issues of granularity of the animacy dimension. Finally, the automatically an-

[1] Parts of the research reported in this paper has been supported by the *Deutsche Forschungsgemeinschaft* (DFG, *Sonderforschungsbereich 632*, project D4).

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 630–638,
Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics

notated data set is employed in order to train a syntactic parser and we investigate the effect of the animacy information and contrast the automatically acquired features with gold standard ones.

The rest of the article is structured as follows. In section 2, we briefly discuss annotation schemes for animacy, the annotation strategies and categories proposed there. We go on to describe annotation for the binary distinction of ‘human reference’ found in a Swedish dependency treebank in section 3 and we perform an evaluation of the consistency of the human annotation in terms of linguistic level. In section 4, we present experiments in lexical acquisition of animacy based on morphosyntactic features extracted from a considerably larger corpus. Section 5 presents experiments with the acquired animacy information applied in the data-driven dependency parsing of Swedish. Finally, section 6 concludes the article and provides some suggestions for future research.

2 Animacy annotation

Annotation for animacy is not a common component of corpora or treebanks. However, following from the theoretical interest in the property of animacy, there have been some initiatives directed at animacy annotation of corpus data.

Corpus studies of animacy (Yamamoto, 1999; Dahl and Fraurud, 1996) have made use of annotated data, however they differ in the extent to which the annotation has been explicitly formulated as an annotation scheme. The annotation study presented in Zaenen et. al. (2004) makes use of a coding manual designed for a project studying genitive modification (Garretson et al., 2004) and presents an explicit annotation scheme for animacy, illustrated by figure 1. The main class distinction for animacy is three-way, distinguishing Human, Other animate and Inanimate, with subclasses under two of the main classes. The ‘Other animate’ class further distinguishes Organizations and Animals. Within the group of inanimates, further distinctions are made between concrete and non-concrete inanimate, as well as time and place nominals.[2]

The annotation scheme described in Zaenen et. al. (2004) annotates the markables according to the animacy of their referent in the particular context. Animacy is thus treated as a token level property, however, has also been proposed as a lexical semantic property of nouns (Yamamoto, 1999). The indirect encoding of animacy in lexical resources, such as WordNet (Fellbaum, 1998) can also be seen as treating animacy as a type-level property. We may thus distinguish between a purely *type level* annotation strategy and a purely *token level* one. Type level properties hold for lexemes and are context-independent, i.e., independent of the particular linguistic context, whereas token-level properties are determined in context and hold for referring expressions, rather than lexemes.

3 Human reference in Swedish

Talbanken05 is a Swedish treebank which was created in the 1970’s and which has recently been converted to dependency format (Nivre et al., 2006b) and made freely available. The written sections of the treebank consist of professional prose and student essays and amount to 197,123 running tokens, spread over 11,431 sentences. Figure 2 shows the labeled dependency graph of example (2), taken from Talbanken05.

(2)	*Samma*	*erfarenhet*	*gjorde*	*engelsmännen*
	same	experience	made	englishmen-DEF

‘The same experience, the Englishmen had’

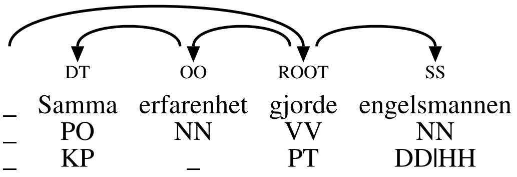

Figure 2: Dependency representation of example (2) from Talbanken05.

In addition to information on part-of-speech, dependency head and relation, and various morphosyntactic properties such as definiteness, the annotation expresses a distinction for nominal elements between reference to human and non-human. The annotation manual (Teleman, 1974) states that a markable should be tagged as human (HH) if it may be replaced by the interrogative pronoun *vem* ‘who’ and be referred to by the personal pronouns *han* ‘he’ or *hon* ‘she’.

There are clear similarities between the annotation for human reference found in Talbanken05 and the annotation scheme for animacy discussed

[2] The fact that the study focuses on genitival modification has clearly influenced the categories distinguished, as these are all distinctions which have been claimed to influence the choice of genitive construction. For instance, temporal nouns are frequent in genitive constructions, unlike the other inanimate nouns.

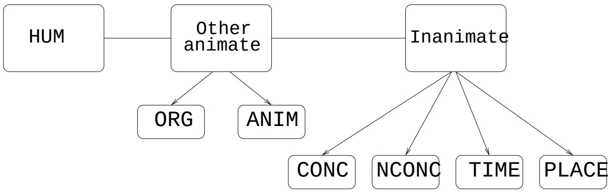

Figure 1: Animacy classification scheme (Zaenen et al., 2004).

above. The human/non-human contrast forms the central distinction in the animacy dimension and, in this respect, the annotation schemes do not conflict. If we compare the annotation found in Talbanken05 with the annotation proposed in Zaenen et. al. (2004), we find that the schemes differ primarily in the granularity of classes distinguished. The main source of variation in class distinctions consists in the annotation of collective nouns, including organizations, as well as animals.

3.1 Level of annotation

We distinguished above between type and token level annotation strategies, where a type level annotation strategy entails that an element consistently be assigned to only one class. A token level strategy, in contrast, does not impose this restriction on the annotation and class assignment may vary depending on the specific context. Garretson et. al (2004) propose a token level annotation strategy and state that "when coding for animacy [...] we are not considering the nominal per se (e.g., the word 'church'), but rather the entity that is the referent of that nominal (e.g. some particular thing in the real world)". This indicates that for all possible markables, a referent should be determinable.

The brief instruction with respect to annotation for human reference in the annotation manual for Talbanken05 (Teleman, 1974, 223) gives leeway for interpretation in the annotation and does not clearly state that it should be based on token level reference in context. It may thus be interesting to examine the extent to which this manual annotation is consistent across lexemes or whether we observe variation. We manually examine the intersection of the two classes of noun lemmas in the written sections of Talbanken, i.e., the set of nouns which have been assigned both classes by the annotators. It contains 82 noun lemmas, which corresponds to only 1.1% of the total number of noun lemmas in the treebank (7554 lemmas all together). After a manual inspection of the intersective elements along with their linguistic contexts, we may group the nouns which were assigned to both classes, into the following categories:that 'HH' is the tag for

Abstract nouns Nouns with underspecified or vague type level properties with respect to animacy, such as quantifying nouns, e.g. *hälft* 'half', *miljon* 'million', as well as nouns which may be employed with varying animacy, e.g. *element* 'element', *part* 'party', as in (3) and (4):

(3) ... *också den andra* **parten**$_{HH}$ *står utanför*
... also the other party-DEF stands outside
'... also the other party is left outside'

(4) *I ett förhållande är aldrig bägge* **parter_** *lika starka*
in a relationship are never both parties same strong
'In a relationship, both parties are never equally strong'

We also find that nouns which denote abstract concepts regarding humans show variable annotation, e.g. *individ* 'individual', *adressat* 'addressee', *medlem* 'member', *kandidat* 'candidate', *representant* 'representative', *auktoritet* 'authority'

Reference shifting contexts These are nouns whose type level animacy is clear but which are employed in a specific context which shifts their reference. Examples include metonymic usage of nouns, as in (5) and nouns occurring in dereferencing constructions, such as predicative constructions (6), titles (7) and idioms (8):

(5) ... **daghemmens**$_{HH}$ *otillräckliga resurser*
... kindergarten-DEF.GEN inadequate resources
'... the kindergarten's inadequate resources'

(6) ... *för att bli en bra* **soldat_**
... for to become a good soldier
'... in order to become a good soldier'

(7) ... *menar* **biskop_** *Hellsten*
... thinks bishop Hellsten
'thinks bishop Hellsten'

(8) *ta* **studenten_**
take student-DEF
'graduate from highschool (lit. take the student)'

It is interesting to note that the main variation in annotation stems precisely from difficulties in determining reference, either due to bleak type level properties such as for the abstract nouns, or due to properties of the context, as in the reference shifting constructions. The small amount of variation in the human annotation for animacy clearly supports a type-level approach to animacy, however, underline the influence of the linguistic context on the conception of animacy, as noted in the literature (Zaenen et al., 2004; Rosenbach, 2008).

4 Lexical acquisition of animacy

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 animacy information. Orăsan and Evans (2007) 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. The task of animacy classification bears some resemblance to the task of named entity recognition (NER) which usually makes reference to a 'person' class. However, whereas most NER systems make extensive use of orthographic, morphological or contextual clues (titles, suffixes) and gazetteers, animacy for nouns is not signaled overtly in the same way.

Following a strategy in line with work on verb classification (Merlo and Stevenson, 2001; Stevenson and Joanis, 2003), we set out to classify common noun *lemmas* based on their morphosyntactic distribution in a considerably larger corpus. This is thus equivalent to treatment of animacy as a lexical semantic property and the classification strategy is based on generalization of morphosyntactic behaviour of common nouns over large quantities of data. Due to the small size of the Talbanken05 treebank and the small amount of variation, this strategy was pursued for the acquisition of animacy information.

Class	Types	Tokens covered
Animate	644	6010
Inanimate	6910	34822
Total	7554	40832

Table 1: The animacy data set from Talbanken05; number of noun lemmas (Types) and tokens in each class.

In the animacy classification of common nouns we exploit well-documented correlations between morphosyntactic realization and semantic properties of nouns. For instance, animate nouns tend to be realized as agentive subjects, inanimate nouns do not (Dahl and Fraurud, 1996). Animate nouns make good 'possessors', whereas inanimate nouns are more likely 'possessees' (Rosenbach, 2008). Table 1 presents an overview of the animacy data for common nouns in Talbanken05. It is clear that the data is highly skewed towards the non-human class, which accounts for 91.5% of the type instances. For classification we organize the data into *accumulated frequency bins*, which include all nouns with frequencies above a certain threshold. We here approximate the class of 'animate' to 'human' and the class of 'inanimate' to 'nonhuman'. Intersective elements, see section 3.1, are assigned to their majority class.[3]

4.1 Features for animacy classification

We define a feature space, which makes use of distributional data regarding the general syntactic properties of a noun, as well as various morphological properties. It is clear that in order for a syntactic environment to be relevant for animacy classification it must be, at least potentially, nominal. We define the *nominal potential* of a dependency relation as the frequency with which it is realized by a nominal element (noun or pronoun) and determine empirically a threshold of .10. The syntactic and morphological features in the feature space are presented below:

Syntactic features A feature for each dependency relation with nominal potential: (transitive) subject (SUBJ), object (OBJ), prepositional complement (PA), root (ROOT)[4], apposition (APP), conjunct (CC), determiner (DET), predicative (PRD), complement of comparative subjunction (UK). We also include a feature for the head of a genitive modifier, the so-called 'possessee', (GENHD).

Morphological features A feature for each morphological distinction relevant for a noun

[3] When there is no majority class, i.e. in the case of ties, the noun is removed from the data set. 12 lemmas were consequently removed.

[4] Nominal elements may be assigned the root relation of the dependency graph in sentence fragments which do not contain a finite verb.

in Swedish: gender (NEU/UTR), number (SIN/PLU), definiteness (DEF/IND), case (NOM/GEN). Also, the part-of-speech tags distinguish dates (DAT) and quantifying nouns (SET), e.g. *del, rad* 'part, row', so these are also included as features.

For extraction of distributional data for the set of Swedish nouns we make use of the Swedish Parole corpus of 21.5M tokens.[5] To facilitate feature extraction, we part-of-speech tag the corpus and parse it with MaltParser[6], which assigns a dependency analysis.[7]

4.2 Experimental methodology

For machine learning, we make use of the Tilburg Memory-Based Learner (TiMBL) (Daelemans et al., 2004).[8] Memory-based learning is a supervised machine learning method characterized by a lazy learning algorithm which postpones learning until classification time, using the k-nearest neighbor algorithm for the classification of unseen instances. For animacy classification, the TiMBL parameters are optimized on a subset of the full data set.[9]

For training and testing of the classifiers, we make use of leave-one-out cross-validation. The baseline represents assignment of the majority class (inanimate) to all nouns in the data set. Due to the skewed distribution of classes, as noted above, the baseline accuracy is very high, usually around 90%.Clearly, however, the class-based measures of precision and recall, as well as the combined F-score measure are more informative for these results. The baseline F-score for the animate class is thus 0, and a main goal is to improve on the rate of true positives for animates, while limiting the trade-off in terms of performance for the majority class of inanimates, which start out with F-scores approaching 100. For calculation of the statistical significance of differences in the performance of classifiers tested on the same data set, McNemar's test (Dietterich, 1998) is employed.

Bin	Instances	Baseline	MBL	SVM
>1000	291	89.3	97.3	95.2
>500	597	88.9	97.3	97.1
>100	1668	90.5	96.8	96.9
>50	2278	90.6	96.1	96.0
>10	3786	90.8	95.4	95.1
>0	5481	91.3	93.9	93.7

Table 2: Accuracy for MBL and SVM classifiers on Talbanken05 nouns in accumulated frequency bins by Parole frequency.

4.3 Results

Column four (MBL) in table 2 shows the accuracy obtained with all features in the general feature space. We observe a clear improvement on all data sets (p<.0001), compared to the respective baselines. As we recall, the data sets are successively larger, hence it seems fair to conclude that the size of the data set partially counteracts the lower frequency of the test nouns. It is not surprising, however, that a method based on distributional features suffers when the absolute frequencies approach 1. We obtain results for animacy classification, ranging from 97.3% accuracy to 93.9% depending on the sparsity of the data. With an absolute frequency threshold of 10, we obtain an accuracy of 95.4%, which constitutes a 50% reduction of error rate.

Table 3 presents the experimental results relative to class. We find that classification of the inanimate class is quite stable throughout the experiments, whereas the classification of the minority class of animate nouns suffers from sparse data. It is an important point, however, that it is largely recall for the animate class which goes down with increased sparseness, whereas precision remains quite stable. All of these properties are clearly advantageous in the application to realistic data sets, where a more conservative classifier is to be preferred.

4.4 Error analysis

The human reference annotation of the Talbanken05 nouns distinguishes only the classes corresponding to 'human' and 'inanimate' along the

[5] Parole is freely available at http://spraakbanken.gu.se

[6] http://www.maltparser.org

[7] For part-of-speech tagging, we employ the MaltTagger – a HMM part-of-speech tagger for Swedish (Hall, 2003). For parsing, we employ MaltParser (Nivre et al., 2006a), a language-independent system for data-driven dependency parsing , with the pretrained model for Swedish, which has been trained on the tags output by the tagger.

[8] http://ilk.uvt.nl/software.html

[9] For parameter optimization we employ the paramsearch tool, supplied with TiMBL, see http://ilk.uvt.nl/software.html. Paramsearch implements a hill climbing search for the optimal settings on iteratively larger parts of the supplied data. We performed parameter optimization on 20% of the total data set, where we balanced the data with respect to frequency. The resulting settings are $k = 11$, GainRatio feature weighting and Inverse Linear (IL) class voting weights.

	Animate			Inanimate		
	Precision	Recall	Fscore	Precision	Recall	Fscore
>1000	89.7	83.9	86.7	98.1	98.8	98.5
>500	89.1	86.4	87.7	98.3	98.7	98.5
>100	87.7	76.6	81.8	97.6	98.9	98.2
>50	85.8	70.2	77.2	97.0	98.9	97.9
>10	81.9	64.0	71.8	96.4	98.6	97.5
>0	75.7	44.9	56.4	94.9	98.6	96.7

Table 3: Precision, recall and F-scores for the two classes in MBL-experiments with a general feature space.

>10 nouns		
(a)	(b)	← classified as
222	125	(a) class animate
49	3390	(b) class inanimate

Table 4: Confusion matrix for the MBL-classifier with a general feature space on the $>$10 data set on Talbanken05 nouns.

animacy dimension. An interesting question is whether the errors show evidence of the gradience in categories discussed earlier and explicitly expressed in the annotation scheme by Zaenen et.al. (2004) in figure 1. If so, we would expect erroneously classified inanimate nouns to contain nouns of intermediate animacy, such as animals and organizations.

The error analysis examines the performance of the MBL-classifier employing all features on the > 10 data set in order to abstract away from the most serious effects of data sparseness. Table 4 shows a confusion matrix for the classification of the nouns. If we examine the errors for the inanimate class we indeed find evidence of gradience within this category. The errors contain a group of nouns referring to animals and other living beings (bacteria, algae), as listed in (9), as well as one noun referring to an "intelligent machine", included in the intermediate animacy category in Zaenen et al. (2004). Collective nouns with human reference and organizations are also found among the errors, listed in (11). We also find some nouns among the errors with human denotation, listed in (12). These are nouns which typically occur in dereferencing contexts, such as titles, e.g. *herr* 'mister', *biskop* 'bishop' and which were annotated as non-human referring by the human annotators.[10] Finally, a group of abstract, human-denoting nouns are also found among the errors, as listed in (13). In summary, we find that nouns with gradient animacy properties account for 53.1% of the errors for the inanimate class.

(9) Animals/living beings:
alg 'algae', *apa* 'monkey', *bakterie* 'bacteria', *björn* 'bear', *djur* 'animal', *fågel* 'bird', *fladdermöss* 'bat', *myra* 'ant', *mås* 'seagull', *parasit* 'parasite'

(10) Intelligent machines:
robot 'robot'

(11) Collective nouns, organizations:
myndighet 'authority', *nation* 'nation', *företagsledning* 'corporate-board', *personal* 'personell', *stiftelse* 'foundation', *idrottsklubb* 'sport-club'

(12) Human-denoting nouns:
biskop 'bishop', *herr* 'mister', *nationalist* 'nationalist', *tolk* 'interpreter'

(13) Abstract, human nouns:
förlorare 'loser', *huvudpart* 'main-party', *konkurrent* 'competitor', *majoritet* 'majority', *värd* 'host'

It is interesting to note that both the human and automatic annotation showed difficulties in ascertaining class for a group of abstract, human-denoting nouns, like *individ* 'individual', *motståndare* 'opponent', *kandidat* 'candidate', *representant* 'representative'. These were all assigned to the animate majority class during extraction, but were misclassified as inanimate during classification.

4.5 SVM classifiers

In order to evaluate whether the classification method generalizes to a different machine learning algorithm, we design an identical set of experiments to the ones presented above, but where classification is performed with Support Vector Machines (SVMs) instead of MBL. We use the LIBSVM package (Chang and Lin, 2001) with a RBF kernel ($C = 8.0, \gamma = 0.5$).[11]

[10] In fact, both of these showed variable annotation in the treebank and were assigned their majority class – inanimate – in the extraction of training data.

[11] As in the MBL-experiment, parameter optimization, i.e., choice of kernel function, C and γ values, is performed on 20% of the total data set with the `easy.py` tool, supplied with LIBSVM.

As column 5 (SVM) in table 2 shows, the classification results are very similar to the results obtained with MBL.[12] We furthermore find a very similar set of errors, and in particular, we find that 51.0 % of the errors for the inanimate class are nouns with the gradient animacy properties presented in (9)-(13) above.

5 Parsing with animacy information

As an external evaluation of our animacy classifier, we apply the induced information to the task of syntactic parsing. Seeing that we have a treebank with gold standard syntactic information and gold standard as well as induced animacy information, it should be possible to study the direct effect of the added animacy information in the assignment of syntactic structure.

5.1 Experimental methodology

We use the freely available MaltParser system, which is a language-independent system for data-driven dependency parsing (Nivre, 2006; Nivre et al., 2006c). A set of parsers are trained on Talbanken05, both with and without additional animacy information, the origin of which is either the manual annotation described in section 3 or the automatic animacy classifier described in section 4.2- 4.4 (MBL). The common nouns in the treebank are classified for animacy using leave-one-out training and testing. This ensures that the training and test instances are disjoint at all times. Moreover, the fact that the distributional data is taken from a separate data set ensures non-circularity since we are not basing the classification on gold standard parses.

All parsing experiments are performed using 10-fold cross-validation for training and testing on the entire written part of Talbanken05. Overall parsing accuracy will be reported using the standard metrics of *labeled attachment score* (LAS) and *unlabeled attachment score* (UAS).[13] Statistical significance is checked using Dan Bikel's randomized parsing evaluation comparator.[14] As our baseline, we use the settings optimized for Swedish in the CoNLL-X shared task (Buchholz and Marsi, 2006), where this parser was the best performing parser for Swedish.

	Gold standard		Automatic	
	UAS	LAS	UAS	LAS
Baseline	89.87	84.92	89.87	84.92
Anim	89.81	84.94	89.87	**84.99**

Table 5: Overall results in experiments with automatic features compared to gold standard features, expressed as unlabeled and labeled attachment scores.

5.2 Results

The addition of automatically assigned animacy information for common nouns (Anim) causes a small, but significant improvement in overall results (p<.04) compared to the baseline, *as well as* the corresponding gold standard experiment (p<.04). In the gold standard experiment, the results are not significantly better than the baseline and the main, overall, improvement from the gold standard animacy information reported in Øvrelid and Nivre (2007) and Øvrelid (2008) stems largely from the animacy annotation of pronouns.[15] This indicates that the animacy information for common nouns, which has been automatically acquired from a considerably larger corpus, captures distributional distinctions which are important for the general effect of animacy and furthermore that the differences from the gold standard annotation prove beneficial for the results.

We see from Table 5, that the improvement in overall parse results is mainly in terms of dependency labeling, reflected in the LAS score. A closer error analysis shows that the performance of the two parsers employing gold and automatic animacy information is very similar with respect to dependency relations and we observe an improved analysis for subjects, (direct and indirect) objects and subject predicatives with only minor variations. This in itself is remarkable, since the covered set of animate instances is notably smaller in the automatically annotated data set. We furthermore find that the main difference between the gold standard and automatic Anim-experiments

[12] The SVM-classifiers generally show slightly lower results, however, only performance on the >1000 data set is significantly lower (p<.05).

[13] LAS and UAS report the percentage of tokens that are assigned the correct head *with* (labeled) or *without* (unlabeled) the correct dependency label.

[14] http://www.cis.upenn.edu/∼dbikel/software.html

[15] Recall that the Talbanken05 treebank contains animacy information for all nominal elements – pronouns, proper and common nouns. When the totality of this information is added the overall parse results are significantly improved (p<.0002) (Øvrelid and Nivre, 2007; Øvrelid, 2008).

does not reside in the analysis of syntactic arguments, but rather of non-arguments. One relation for which performance deteriorates with the added information in the gold Anim-experiment is the nominal postmodifier relation (ET) which is employed for relative clauses and nominal PP-attachment. With the automatically assigned feature, in contrast, we observe an improvement in the performance for the ET relation, compared to the gold standard experiment, from a F-score in the latter of 76.14 to 76.40 in the former. Since this is a quite common relation, with a frequency of 5% in the treebank as a whole, the improvement has a clear effect on the results.

The parser's analysis of postnominal modification is influenced by the differences in the added animacy annotation for the nominal head, as well as the internal dependent. If we examine the corrected errors in the automatic experiment, compared to the gold standard experiment, we find elements with differing annotation. Preferences with respect to the animacy of prepositional complements vary. In (14), the automatic annotation of the noun *djur* 'animal' as animate results in correct assignment of the ET relation to the preposition *hos* 'among', as well as correct nominal, as opposed to verbal, attachment. This preposition is one of the few with a preference for animate complements in the treebank. In contrast, the example in (15) illustrates an error where the automatic classification of *barn* 'children' as inanimate causes a correct analysis of the head preposition *om* 'about'.[16]

(14)	...*samhällsbildningar*	*hos*	*olika*	**djur**
	...societies	among	different	animals

'...social organizations among different animals'

(15)	*Föräldrar*	*har*	*vårdnaden*	*om*	*sina*	**barn**
	parents	have	custody-DEF	of	their	children

'Parents have the custody of their children'

A more thorough analysis of the different factors involved in PP-attachment is a complex task which is clearly beyond the scope of the present study. We may note, however, that the distinctions induced by the animacy classifier based purely on linguistic evidence proves useful for the analysis of both arguments and non-arguments.

6 Conclusion

This article has dealt with an empirical evaluation of animacy annotation in Swedish, where the main focus has been on the use of such annotation for computational purposes.

We have seen that human annotation for animacy shows little variation at the type-level for a binary animacy distinction. Following from this observation, we have shown how a type-level induction strategy based on morphosyntactic distributional features enables automatic animacy classification for noun lemmas which furthermore generalizes to different machine learning algorithms (MBL, SVM). We obtain results for animacy classification, ranging from 97.3% accuracy to 93.9% depending on the sparsity of the data. With an absolute frequency threshold of 10, we obtain an accuracy of 95.4%, which constitutes a 50% reduction of error rate. A detailed error analysis revealed some interesting results and we saw that more than half of the errors performed by the animacy classifier for the large class of inanimate nouns actually included elements which have been assigned an intermediate animacy status in theoretical work, such as animals and collective nouns.

The application of animacy annotation in the task of syntactic parsing provided a test bed for the applicability of the annotation, where we could contrast the manually assigned classes with the automatically acquired ones. The results showed that the automatically acquired information gives a slight, but significant improvement of overall parse results where the gold standard annotation does not, despite a considerably lower coverage. This is a suprising result which highlights important properties of the annotation. First of all, the automatic annotation is completely consistent at the type level. Second, the automatic animacy classifier captures important distributional properties of the nouns, exemplified by the case of nominal postmodifiers in PP-attachment. The automatic annotation thus captures a purely linguistic notion of animacy and abstracts over contextual influence in particular instances.

Animacy has been shown to be an important property in a range of languages, hence animacy classification of other languages constitutes an interesting line of work for the future, where empirical evaluations may point to similarities and differences in the linguistic expression of animacy.

[16] Recall that the classification is based purely on linguistic evidence and in this respect children largely pattern with the inanimate nouns. A child is probably more like a physical object in the sense that it is something one possesses and otherwise reacts *to*, rather than being an agent that acts upon its surroundings.

References

Judith Aissen. 2003. Differential Object Marking: Iconicity vs. economy. *Natural Language and Linguistic Theory*, 21(3):435–483.

Holly P. Branigan, Martin J. Pickering, and Mikihiro Tanaka. 2008. Contributions of animacy to grammatical function assignment and word order production. *Lingua*, 118(2):172–189.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2005. Predicting the dative alternation. In Gosse Bouma, Irene Kraemer, and Joost Zwarts, editors, *Cognitive foundations of interpretation*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164.

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Walter Daelemans, Jakub Zavrel, Ko Van der Sloot, and Antal Van den Bosch. 2004. TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. Technical report, ILK Technical Report Series 04-02.

Östen Dahl and Kari Fraurud. 1996. Animacy in grammar and discourse. In Thorstein Fretheim and Jeanette K. Gundel, editors, *Reference and referent accessibility*, pages 47–65. John Benjamins, Amsterdam.

Peter de Swart, Monique Lamers, and Sander Lestrade. 2008. Animacy, argument structure and argument encoding: Introduction to the special issue on animacy. *Lingua*, 118(2):131–140.

Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. 2005. Climbing the path to grammar: A maximum entropy model of subject/object learning. In *Proceedings of the 2nd Workshop on Psychocomputational Models of Human Language Acquisition*, pages 72–81.

Thomas G. Dietterich. 1998. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

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

Gregory Garretson, M. Catherine O'Connor, Barbora Skarabela, and Marjorie Hogan, 2004. *Optimal Typology of Determiner Phrases Coding Manual*. Boston University, version 3.2 edition. Downloaded from http://people.bu.edu/depot/coding_manual.html on 02/15/2006.

Johan Hall. 2003. A probabilistic part-of-speech tagger with suffix probabilities. Master's thesis, Växjö University, Sweden.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006a. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 2216–2219.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006b. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1392–1395.

Joakim Nivre, Jens Nilsson, Johan Hall, Gülşen Eryiğit, and Svetoslav Marinov. 2006c. Labeled pseudo-projective dependency parsing with Support Vector Machines. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Joakim Nivre. 2006. *Inductive Dependency Parsing*. Springer, Dordrecht.

Constantin Orăsan and Richard Evans. 2007. NP animacy resolution for anaphora resolution. *Journal of Artificial Intelligence Research*, 29:79–103.

Lilja Øvrelid and Joakim Nivre. 2007. When word order and part-of-speech tags are not enough – Swedish dependency parsing with rich linguistic features. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 447–451.

Lilja Øvrelid. 2008. Linguistic features in data-driven dependency parsing. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL 2008)*.

Anette Rosenbach. 2008. Animacy and grammatical variation - findings from English genitive variation. *Lingua*, 118(2):151–171.

Michael Silverstein. 1976. Hierarchy of features and ergativity. In Robert M.W. Dixon, editor, *Grammatical categories in Australian Languages*, pages 112–171. Australian Institute of Aboriginal Studies, Canberra.

Suzanne Stevenson and Eric Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 71–78.

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Studentlitteratur, Lund.

J. Weckerly and M. Kutas. 1999. An electrophysiological analysis of animacy effects in the processing of object relative sentences. *Psychophysiology*, 36:559–570.

Mutsumi Yamamoto. 1999. *Animacy and Reference: A cognitive approach to corpus linguistics*. John Benjamins, Amsterdam.

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 Donna Byron and Bonnie Webber, editors, *Proceedings of the ACL Workshop on Discourse Annotation*.