MULTILINGUAL ANIMACY CLASSIFICATION BY SPARSE LOGISTIC REGRESSION

Kirk Baker and Chris Brew

Abstract

This paper presents results from three experiments on automatic animacy classification in Japanese and English. We present experiments that focus on solutions to the problem of reliably classifying a large set of infrequent items using a small number of automatically extracted features. We labeled a set of Japanese nouns as ±animate on the basis of reliable, surface-obvious morphological features, producing an accurately but sparsely labeled data set. To classify these nouns, and to achieve good generalization to other nouns for which we do not have labels, we used feature vectors based on frequency counts of verb-argument relations that abstract away from item identity and into class-wide distributional tendencies of the feature set. Grouping items into suffix-based equivalence classes prior to classification increased data coverage and improved classification accuracy. For the items that occur at least once with our feature set, we obtained 95% classification accuracy. We used loanwords to transfer automatically acquired labels from English to classify items that are zero-frequency in the Japanese data set, giving increased precision on inanimate items and increased recall on animate items.

Kirk Baker & Chris Brew. Multilingual animacy classification by sparse logistic regression. *OSUWPL* Vol. 59, pp. 52-75.

1. Introduction

Distinguishing animate from inanimate noun phrases is important for a number of morphological and other linguistic processes. For example, most languages exhibit some type of syntactic or morphological alternations that reflect perceived distinctions in animacy (e.g., pronoun categories like s/he vs. it, or marking animate direct objects differently from inanimate ones). Animacy is associated with other potentially useful properties, including sentience, autonomy, and intentionality. The ability to make accurate inferences about the animacy of potential noun phrase referents is important for many natural language processing tasks such as pronoun resolution (Oräsan and Evans 2001), machine translation (Ilarraza et al. 2002), and language generation (Zaenen et al. 2004).

The core sense in which we are using the term "animacy" is straightforward. Animals (including humans) are animate; vegetables, minerals and manufactured objects are not. But it is far from straightforward to make detailed claims about the nature of the concepts that underly the linguistic distinctions that are typically labeled with the terms "animate" or "inanimate". There is no fully convincing way of establishing that two languages are operating with the same concept of animacy. Worse, there are no generally agreed criteria for deciding that a particular linguistic pattern is really a reflection of any concept of animacy. The relevant distinction could very well turn out to be agency, sentience or the ability to move in an apparently autonomous fashion. We have no intention to resolve either of these difficulties here. We will provide examples of animacy distinctions in action, and assume without detailed argument that the ontologies partly encoded in the linguistic distinctions of different languages are sufficiently similar for the enterprise of cross-linguistic information transfer to make sense.

Most of the literature on animacy distinctions and its relevance for human language makes reference to an animacy scale, which arranges objects along a sort of discretum from animate to inanimate. The number of categories is not fixed, but depends to some extent on the degree of resolution deemed appropriate for the description of the phenomenon at hand. Sometimes a three-way distinction between humans, other animates and inanimates is made (e.g., Zaenen et al. 2004), but often humans and animals are both treated as animate (e.g., van Nice and Dietrich 2003, Bresnan and Hay 2006) and contrasted with everything else. According to Bresnan and Hay (2006), Bresnan et al. (2005) used a four way animacy distinction that consisted of the categories human, organization, animal/intelligent machine and inanimate. Yamamoto (1999) provides an elaborate radial hierarchy view of animacy with 1st person pronouns at the center, 2nd person pronouns and other people connected to that and extending outward to supernatural being, animals, machines, plants, etc.

As we will see in the examples below, language users can and will make flexible use of their language's conventions about animacy. A simple example from English is that pet animals can be referred to using either the masculine and feminine pronouns "he" and "she" or the neutral pronoun "it". People who like pets and think of them as family members are more likely to choose the gender marked pronouns. If you do not like them, you will be more likely to use the neuter pronoun. In extremis the same thing can be done with babies, although the risk of opprobrium is correspondingly greater.

Patterns related to animacy show up in various ways in different languages. One relatively common way is for animate objects to be marked with some kind of case marker while inanimate objects are not. In Spanish, for example, patient noun phrases are marked with the preposition *a* if they are animate (1b), but not if they are inanimate (1a)(Yamamoto 1999:47, 13):

(1a) Ha comprado un nuevo libro.
have:3SG bought a new book
'He has bought a new book.'

(1b) Ha comprado *a* un nuevo caballo
have:3SG bought *to* a new horse
'He has bought a new horse.'

Similar facts are reported for Bantu languages like Swahili (Woolford 1999) and Native American languages like Blackfoot (Bliss 2005). Animacy has been claimed to play a role in languages with nominal classifier systems. For example, Japanese has an extensive system of numeral classifiers that selectively attach to subsets of the language's nouns. These are subsets are clearly related by properties such as animacy, shape, size and function (Iida 1999). A few examples are shown below. In the column labeled “Animacy” there are examples of four different classes of animate entities, each with its own numeral classifiers. Roughly, the -hiki class is for smallish animals, while the -tou class is for larger animals.

Animacy		Shape		Function	
-nin	*hito-ga san-nin*	*-hon*	*ohashi-ga san-bon*	*-dai*	*kuruma-ga san-dai*
	'three people'		'three chopsticks'		'three cars'
-hiki	*hebi-ga san-biki*	*-mai*	*shatsu-ga san-mai*	*-hatsu*	*juusei-ga san-batsu*
	'three snakes'		'three shirts'		'three gunshots'
-tou	*uma-ga san-tou*				
	'three horses'				
-wa	*tori-ga san-wa*				
	'three birds'				

Table 1. Some Japanese numeral classifiers.

Even in languages where animate and inanimate nouns have the same morphological properties, there are distributional differences in the occurrences of animate and inanimate nouns. For example, a corpus study of the syntactic distribution of animate and inanimate noun phrases in Swedish (Dahl and Fraurud 1996) found significant differences in the proportion of animate NPs that occurred as direct objects (13.0%) versus indirect objects (83.1%) and as transitive subjects (56.5%) versus intransitive subjects (26.0%). English evidences animacy preferences in double object word order (e.g., *give me money* versus *give money to me* and possessives (e.g., *-'s* used more with animates; *of* used more with inanimates) (Zaenen et al. 2004).

Taking animacy as a category that exists independently of linguistic categorizations of it necessitates a distinction between grammatical animacy and non-linguistic animacy.[1] Grammatical animacy is a type of lexical specification like person or number agreement that shows up in verb inflection and case marking. Like much conventionalized linguistic structure, the exact nature of the relationship between grammatical animacy and the correlates of non-linguistic animacy that supposedly underlie its use can be unclear. For example, Polish has been described as a language that employs grammatical animacy (Gawronska et al. 2002). Polish has four animacy classes that are distinguishable in terms of their accusative case marking and verb agreement:

- superanimate - grammatically masculine, human. Accusative form equals the genitive.
- animate - masculine or feminine living things. Singular accusative is the genitive; plural accusative is the same as the nominative.
- inanimate - masculine, feminine, or neuter non-living things. Accusative form equals the nominative.
- semi-animate - grammatically masculine, not living. Accusative form patterns like the animates.

The semi-animate nouns do not fall into any discernible semantic category, comprising things like names of dances, games or some actions; food, money, and negative mental states (Yamamoto 1999).

Similar phenomena exist for other languages noted for their sensitivity to animacy. For example, in Algonquian languages the distinction between animate and inanimate nouns overlaps in part with pronominal distinctions made in languages like English, but there are also many cases where correspondences are difficult to discern. For example, in Plains Cree, items like *opswākan* 'pipe', *mīhkwan* 'spoon', and *āpoy* 'paddle' are grammatically animate but do not correspond in a biological or grammatical sense to English nouns that are considered living (Joseph 1979). In Blackfoot as well, grammatical animacy cross-cuts sentience: nouns like *ato'ahsim* 'sock', *isttoan* `knife', *pokon* 'ball', *po'taa'tsis* and 'stove' are grammatically animate but not alive (Bliss 2005). Closely related Bantu languages encode animate object agreement differently from each other depending on additional factors like definiteness, agentivity, person, and number (Woolford 1999). In spite of this sometimes unintuitive disconnect between grammatical and non-linguistic animacy, a great deal of research has gone into exploring the ways in which animacy distinctions surface in language.

A variety of natural language phenomena are said to be sensitive to distinctions in animacy. For example, in English the choice of the genitive is influenced in part by the animacy of the possessor. Jäger and Rosenbach (2006) conducted an experimental study

[1] Some authors make this distinction in terms of 'syntactic animacy' versus 'semantic animacy' (e.g., Yamamoto 1999:50 and references therein).

in which subjects were asked to read a short text such as the example below and choose which of the two underlined possessive constructions was more natural:

> A helicopter waited on the nearby grass like a sleeping insect, its pilot standing outside with Marino. Whit, a perfect specimen of male fitness in a black suit, opened [the helicopter's doors/the doors of the helicopter] to help us board.

Subjects were more likely to choose the *'s* construction when the possessor was animate (e.g., *the boy's eyes*), than when the possessor was inanimate (e.g., *the fumes of the car*).

Animacy also plays a role in English in the syntax of verbs of giving (e.g., *give*, *take*, *bring*). Such verbs can be realized in two ways: the double object construction (2a), or the dative alternation (2b).

(2a) Who gave you that watch?
(2b) Who gave that watch to you?

Bresnan and Hay (2006) analyzed instances of the verb *give* that occurred in a corpus of New Zealand and American English talkers, and found that non-animate recipients were more than 11 times more likely to be realized in the prepositional dative than animate recipients were. (They also found that non-animates were more likely to appear in the double object construction in New Zealand versus American English.)

Dowty (1991) presents an analysis of a class of English verbs that are similar to verbs like *give* in that they permit a syntactic argument alternation that does not correspond to a difference in interpretation. Dowty (1991) refers to this as the *with/against* alternation, and an example is shown below (Dowty 1991:594, 62):

(3a) John hit the fence with the stick.
(3b) John hit the stick against the fence.

In contrast, a semantically similar class of verbs is claimed to not permit this alternation (Dowty1991:596, 65):

(4a) swat the boy with a stick
(4b) *swat the stick at/against the boy

This class, which includes verbs like *smack*, *wallop*, *swat*, *clobber* is said to be distinguished by the fact that its verbs restrict their objects to human or other animate beings, and entail a significant change of state in their direct object arguments.

Animacy has also been claimed to play a role in Japanese grammar as well. As illustrated in Table 1, numeral classifiers are sensitive to animacy distinctions. According to Iida (1999) animacy is the most important of four basic semantic features that play important roles in determining which classifier is selected for a given noun:

ANIMACY > FUNCTION > SHAPE > CONCRETENESS.

Thus, in assigning a classifier to an object that could be counted with either one, (e.g., a snake which is both long and animate), animacy takes precedence. Robot dogs can be counted with animate counters (*-hiki* or *-tou*); according to Iida (1999), when “one feels a high degree of animacy in a certain object, one can count it with animate classifiers; on the other hand, when one does not find animacy in what one is counting, there is no way to use animate classifiers”.

Japanese has a number of other lexical items which are correlated with the animacy of their arguments. Although nouns are typically not marked to indicate number, there is a plural marker *-tachi* which, when it is used, is largely restricted to animate nouns (e.g., *watashi-tachi* 'us', *hito-tachi* 'people'). Japanese has two distinct verbs meaning 'to exist', *iru* and *aru* that show tendencies for selecting animate and inanimate subjects, respectively:

(5a) dansa-ga *iru* = ANIMATE
dancer NOM *exist*
'There's a dancer.'
(5b) toosutaa-ga *aru* = INANIMATE
toaster NOM *exist*
'There's a toaster.'

We are interested in the problem of learning a statistical classifier that can reliably distinguish animate from inanimate noun phrase referents on the basis of their contextual distribution in a large text corpus. As the examples above illustrate, we cannot assume that there is a single sense of animacy that applies across all languages. However, we are comfortable with the assumption that at some cross-linguistic level, perceptual properties commonly associated with animacy such as sentience, agency, and intentionality overlap one another and with the presumed biological basis of an animate/inanimate distinction.

In the empirical work reported here, we correlate animacy information across two historically unrelated languages, Japanese and English. Our hypothesis is that combining lexical information from multiple languages allows for more reliable automatic semantic classification than is possible using cues from only one language. We make no claim that the notions of animacy that are involved are exactly parallel. If pressed, we would argue that two independently drawn languages may make similar choices about the gross distinction between clearly animate entities (for example: human beings and large wild animals) and clearly inanimate entities (for example: rocks and pieces of wood). We would not necessarily expect usable consensus on the trickier cases (statues, robots, thermostats or the personified form of The North Wind).

2. Previous Work on Automatic Animacy Classification

Oräsan and Evans (2001) describes a method for animacy classification of English nouns that applies an instance based learning method to WordNet sense annotated corpus data. In a two step process, they first categorize WordNet senses by animacy, and use that information to classify nouns whose sense is unknown. The assumption motivating their methodology is that a noun with lots of animate senses is more likely to refer to an animate entity in a discourse; conversely, a noun with a majority of inanimate senses is more likely to refer to an inanimate entity.

Oräsan and Evans (2001) defines animacy as the property of a NP whose referent, when singular, can be referred to pronominally with an element from the set {*he*, *him*, *his*, *himself*, *she*, *her*, *hers*, *herself*}. They explicitly reject classifying animals as animate, considering references to pets as *he* or *she* to be an "unusual usage of senses".

The first part of their method involved manually annotating 20026 NPs in a 52 file subset of SEMCOR (Palmer et al. 2005) for animacy. In order to determine the animacy of a given sense, they work bottom up to calculate the proportion of animate hyponyms of that sense. If all of the hyponyms of a given sense are animate, that sense is classified as animate. If not, they use a chi-square test to determine whether the proportion of animate to inanimate hyponyms is reliably skewed enough for that sense to be labeled animate.

A similar procedure is used to label verbs senses as ±animate. In this case, they use the proportion of animate or inanimate nouns that appear as subjects of a verb sense to decide the animacy label.

The number of animate and inanimate senses are used to make a vector representation of each noun. The features in this vector are:

- noun lemma
- number of animate noun senses
- number of inanimate noun senses
- number of animate verb senses
- number of inanimate verb senses
- ratio of animate to inanimate pronouns in the document

The last feature, the ratio of animate to inanimate pronouns, is calculated as the count of {*he*, *she*} in the document divided by the count of {*it*}.

The second part of their methodology involved using TiMBL (Daelemans et al. 2003) to classify nouns on the basis of these feature vectors. They settled on a k-nearest neighbors classifier (k=3) using gain ratio as a weighting feature. The most important feature was the number of animate noun senses followed by the number of inanimate noun senses.

Oräsan and Evans (2001) offers two evaluations of their method: 5 fold cross validation on SEMCOR, and training on SEMCOR and testing on a set of texts from Amnesty International. Their results are shown in Table 2. Overall accuracy on both corpora is around 98%, and precision and recall for both classes ranges from about 90% to 98%.

	Animate		Inanimate	
Accuracy (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)

97.5	88.9 n = 2512	91.0	98.7 n = 17514	98.4	SEMCOR
97.7	94.3 n = 537	92.2	98.4 n = 2585	98.8	Amnesty International

Table 2. Classification results from Oräsan and Evans (2001), Table 3

Oräsan and Evans (2001) note that the most problematic words for their system were named entities, as these tended not to occur in WordNet. Other problematic cases were animate nouns that did not have an animate sense in WordNet (e.g., “Bob”). Because named entities were “constantly misclassified”, Oräsan and Evans (2001) removed them from their dataset. Oräsan and Evans (2001) also report that their system learned that most unknown words were inanimate, and in most cases correctly classified them as such. Given that 87% of the NPs in SEMCOR were inanimate, and that 83% of the Amnesty International data were also inanimate, it is not clear how much is actually being generalized by their system. Removing only problematic items from consideration makes interpreting their results difficult and illustrates some of the problems associated with a system based mainly on static lexical lookup.

Ilarraza et al. (2002) presents a dictionary based method for assigning animacy labels and a reliability score to Basque nouns that relies on manually labeling a small set of items and using synonymy and hypernymy relations to extend those labels to a larger set. The insight motivating their proposal is that conventions for defining words in a dictionary can be exploited to iteratively classify nouns on the basis of selective annotation.

According to Ilarraza et al. (2002), dictionary definitions come in three forms that can be used to extrapolate semantic information about lexical entries. The classical definition relates the entry to a hypernym plus some differentiating information, i.e.,

> airplane: a **vehicle** (*hypernym*) that **can fly** (*differentia*)

Another method consists of specific relators that determine the semantic relationship between the entry and the definition:

> horsefly: **name** (*relator*) given to a kind of **insect** (*related term*)

The third method is to list synonyms of the entry:

> finish: **stop** (*synonym*), **terminate** (*synonym*),

Ilarraza et al. (2002) labeled the 100 most frequent hypernyms and relators that occurred in definitions in a Basque monolingual dictionary, and gave these a reliability rating of 1.0. Assuming that the animacy of a hypernym applies to all of its hyponyms, they search the dictionary and for each definition of an entry apply the animacy label of its relator or hypernym. That entry's reliability score is equal to the number of labeled keywords over the number of senses. For example, *armadura* 'armor' is labeled -animate with a reliability of 0.66 because it occurred with three definitions, two of which contained a labeled hypernym:

Noun	# def	# hype	Labeling process			Anim.	Rel.
armadura	3	2	**multzo**[-]1	**babesgarri**[-]1	**soineko**[]	[-]	0.66
armor			*collection*	*protector*	*garment*		

Table 3. Example animacy labeling process from Ilarraza et al. (2002).

This labeling process iterates such that if *armadura* 'armor' appears as a hypernym or relator of another noun in the dictionary, that entry can be labeled using the information assigned to *armadura* 'armor' in the previous iteration. After an iteration, synonyms are classified according to the labels just assigned. Ilarraza et al. (2002) states that automatic labeling asymptotes after about 8 iterations, with 75% of the nouns in the dictionary covered.

Ilarraza et al. (2002) offer two evaluations of their system, one in terms of items in the dictionary and another on corpus data. For the dictionary evaluation, they selected 1% of the nouns (123 items) for hand checking, and report 99.2% accuracy (75.1% recall). They do not report the breakdown of animate to inanimate items, do not state whether evaluation was in terms of senses of the checked items, nor explain the criteria for selecting the verification set. Ilarraza et al. (2002) report an overall recall of 47.6% of noun types, but do not report accuracy. Of the 3434 nouns labeled, 356 were classified as animate, and 3078 were classified as inanimate.

Øvrelid (2006) presents a machine learning approach to classifying nouns in Norwegian as ±animate that is based on decision trees trained on relative frequency measures of morphosyntactic features extracted from an automatically annotated corpus. Øvrelid (2006) uses a set of linguistically motivated features that potentially correlate with animacy. Nouns are represented as vectors containing the relative frequency of each of the following features:

- transitive subject/direct object: the prototypical transitive relation involves an animate subject and an inanimate direct object.
- demoted agent in passive: a correlation is assumed between animacy and agentivity
- reference by personal pronoun: Norwegian pronouns distinguish antecedents by animacy (e.g., *han/hun* 'he/she' vs. *den/det* 'it-MASC/NEUT').
- reference by reflexive pronoun: assumes that the agentive semantics of the reflexive might favor animate nouns
- genitive *-s*: assumes that possession is a property of animate entities

Øvrelid (2006) reports several experiments looking at the effect of frequency on classification accuracy. The first evaluation is based on 40 high frequency nouns, evenly split for animacy. She reports overall accuracy of 87.5% when all features are used, but does not report precision and recall for each class separately. Leave one out training and testing shows that SUBJECT and REFLEXIVE are the best individual features (85% and 82.5% accuracy). Her second experiment looks at the classification accuracy of nouns that occurred around 100, 50, and 10 times in the corpus (40 items in each bin). Classification accuracy drops to 70% using the classifier trained for the first evaluation. Although she does not report precision and recall, these values can be calculated from the confusion matrix for the nouns with frequency 100 in Øvrelid (2006:52, Table 5), for some indication of classifier performance.

	Animate		Inanimate	
Accuracy (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
70.0	83.3	50.0	64.3	90.0

Table 4. Precision and recall values calculated from Øvrelid (2006).

Øvrelid (2006) also investigated the effect of backing off to the most frequent features, and found that only using SUBJECT, OBJECT and GENITIVE performs slightly better for the nouns with frequency 100 and 50, but shows no improvement for the less frequent group. A backoff classifier, trained on nouns of similar frequency, showed some improvement when using all the features, but was unchanged for individual features.

3. Experimental Setup

3.1 Data Set

Japanese has a number of lexical features related to animacy that make it relatively easy to automatically label a subset of nouns without recourse to an external lexical resource. We used three of these to initially label a set of training data as animate or inanimate:

- the plural suffix *-tachi*, which typically attaches to animate nouns.
- Japanese has two verbs meaning `to exist'. *iru* is typically used with animate subjects, while *aru* typically occurs with inanimate ones.

We used CaboCha v0.53 (Kudo and Matsumoto 2002), a deterministic dependency analyzer for Japanese, to extract verb-argument pairs from the 229 MB (30M word) Japanese Business News Text corpus (Graff and Wu 1995). From its output, we extracted nouns marked by *-tachi* and subjects of *iru* and labeled them animate; subjects of *aru* were labeled inanimate. We defined subject as a nominative case-marked dependent of a verb, and object as an accusative case-marked dependent of a verb. We eliminated items that occurred as subjects of both verbs, and verified the remaining items against a native Japanese speaker's judgments.

We also identified and extracted English loanwords that appeared in the corpus and back-transliterated these into the original English source word (e.g., *herikoputaa* → helicopter). We used WordNet2.0 (Fellbaum 1998) to classify the English words as ±animate, and retained those nouns which had only animate or only inanimate senses. We assume that the animacy of the English transliteration is the same as that of its Japanese counterpart[2.]

[2] It is certainly possible to find cases where the WordNet animacy label of an English source word

We obtained animacy labels for 14389 words (4433 animate, 9956 inanimate); 63% were English loanwords. Table 5 contains a categorization of items within the two animacy classes in terms of a set of fairly coarse-grained WordNet distinctions.

Semantic Category	Percent	Examples
Animate		
Occupation	59%	*chiji* 'governor', *gakusei* 'student'
Proper Name	18%	*kiyouko* 'Kiyoko', *heminguwei* 'Hemingway'
Human Being	15%	*obaachan* 'grandma', *watashi* 'I'
Animal	8%	*saru* 'monkey', *iruka* 'dolphin'
Inanimate		
Concrete	40%	*jushi* 'resin', *opaaru* 'opal'
Activity	27%	*ragubii* 'rugby', *tsutae* 'conveying'
Abstract	20%	*gihou* 'technique', *konomi* 'taste'
Location	7%	*itarii* 'Italy', *sabaku* 'desert'
Organization	6%	*renmei* 'association', *hamasu* 'Hamas'

Table 5. Characteristics of the data set.

While these labeling methods are heuristic in nature, we have checked a sample of the results, and believe the labels to be of a quality comparable to what would be obtained by more costly methods. Of course, in focusing on the cases where simple indicators give reliable results, we sacrifice coverage relative to approaches that make heavier use of informed human judgment.

3.2 Feature Set

We considered three criteria in choosing a feature set: accuracy, coverage, and generality. One commonly used source of features is the pooled adjacent context approach (Redington et al. 1998), which uses counts of words that occur in a sliding window surrounding the target word. Variants of this approach are often used for similar lexical acquisition tasks such as word sense disambiguation, (e.g., Kaji and Morimoto 2005), and computing them requires little prior knowledge of the language at hand. However, contextually associated words are not always good cues to animacy: *soldier* and *tank* may often co-occur within the same word window, but *soldier* is animate while *tank* is inanimate. On the other hand, *soldier* and *tank* are not equally likely as subjects of a verb like *believe*. These facts suggest that some consideration of the structural relations between words will be useful in discriminating animate from inanimate items.

differs from the animacy of that loan in Japanese. For example, WordNet2.0 lists one sense for the noun *super*, which is animate (superintendent, super), whereas in Japanese *supaa* usually refers to a supermarket. In general, the transferred labels were reliable.

Many verbs are known to impose semantic restrictions on one or both arguments (e.g., the subject and object of *murder* are typically both people), indicating that verbs may work well for animacy classification. Relative to a set of specific syntactic or morphological features, verbs should provide reasonable coverage of the data set. Using verbs as features affords some measure of cross-linguistic consistency, as verb-argument relations are in principle language-independent (with language-specific instantiations). Therefore, we restricted our feature set to verbs governing as subject or object the nouns we wanted to classify. For each noun in our study we created three feature vectors based on the number of its subject occurrences, and three feature vectors based on the number of its object occurrences.

Subject (Object) Frequency: frequency with which a noun occurs as the subject (object) of a verb. Features are individual verbs, and values are counts. Of the three representations, we expect this one to do the best job of preserving distinctions between items. However, most verbs will occur with a small subset of nouns, making generalization relatively difficult.

Verb Animacy Ratio: for each verb in the training set, we calculated its subject animacy ratio as the number of animate subjects divided by the total number of subjects (likewise for the object animacy ratio) and substituted this new value for the original frequency of that verb when it occurred as a feature of a particular noun. These ratios were calculated over the training data, and applied to the features in the test set. Test features that did not occur in the training set were discarded.

The individual verbs are still used as features, but values are animacy ratios rather than counts. We are now paying attention to the general predictive power of the verb, rather than the frequency with which it occurs with a particular noun. We expect that with this representation, some distinctions between items will be lost because any time two items share a feature, they will have the same value for it. but expect generalization to improve,
because the feature values are now representative of class-wide tendencies across the training data, and are based on a larger quantity of data.

Average Verb Animacy Ratio: for each noun, we created a vector with a single feature whose value was the average animacy ratios of the verbs that occurred with that noun at least once. This representation essentially eliminates item-level distinctions, but should generalize across classes well even with sparse data. We expect that a single estimate of class-wide tendencies will be robust to some of the variation associated with a large number of infrequent features.

3.3 Classifier

Any of a number of machine learning techniques are suitable to the task of automatic animacy classification given the proposed feature set. We used Bayesian logistic regression (Genkin et al. 2004), a sparse logistic regression classifier that efficiently deals with a large number of features. This model has been shown to work well for other natural language classification tasks including lexical semantic verb classification (Li and

Brew 2008; Li, Baker and Brew 2008), text categorization (Genkin et al. 2004) and author identification (Madigan et al. 2005). Bayesian regression is based on a sparse logistic regression model that uses a prior distribution favoring feature weights of zero, simultaneously selecting features and providing shrinkage. Detailed description of the model is available in Baker (2008). We used a publicly available implementation of the model[3] and specified a Laplace prior with mode zero.

4. Experiments

We ran three experiments on animacy classification in English and Japanese. The first experiment establishes baseline classification accuracy using feature vectors based on frequency counts of verb-subject and verb-object relations. The second experiment examines the impact that grouping items into equivalence classes prior to classification has on data coverage and classification accuracy. The third experiment focuses on classifying zero-frequency items by training an English classifier on translations of the Japanese items and transferring those labels back onto the Japanese data set.

4.1 Experiment One

The purpose of this experiment is to classify Japanese nouns as ±animate, comparing the coverage and classification accuracy of feature vectors containing subject (object) counts, verb animacy ratios, and average verb animacy ratio. The results of the classification are shown in Table 6.

In terms of coverage, object counts accommodate slightly more of our data set (36%) than subject counts (33%). When combined, subject and object counts cover 40% of the data set (meaning that the remaining 60% were not parsed as subject or object of any of the verbs in the feature set[4]). The combined feature set tends to have the highest precision and recall within each feature type, and the best performing combination of feature set and feature type is average verb animacy ratio with combined subject and object.

Feat	Cvg (%)	Acc (%)	Inanimate		Animate	
			Prec (%)	Rec (%)	Prec (%)	Rec (%)
Subject (Object) Frequency						
Subject	33	83.6	84.7	95.3	78.3	49.8
Object	36	85.9	86.2	97.5	83.4	44.8
Subj+Obj	40	85.7	86.5	95.6	82.0	57.1
Verb Animacy Ratio						
Subject	33	83.1	85.9	92.6	71.9	55.6
Object	36	85.9	89.2	93.3	71.2	59.8
Subj+Obj	40	84.9	87.3	93.2	75.7	60.6
Average Verb Animacy Ratio						

[3] http://www.bayesianregression.org/

[4] We excluded *iru* and *aru* from the feature set because we used these two verbs to select the data set.

Subject	33	86.8	88.7	94.3	79.7	65.0
Object	36	88.1	89.1	96.6	82.6	57.9
Subj+Obj	40	88.0	89.2	95.4	83.5	66.7

Table 6. Classification results for Japanese.

The most frequent baseline for the covered portion of the data set is about 50%. All of the feature types outperform the baseline by 30-38%. Object counts perform as well as subject counts across the feature types. Overall, precision and recall for the animate nouns (p=79%, r=58%) tends to be considerably lower than for the inanimate nouns (p=88%, r=95%). Precision of the animate class is lowest when using verb animacy ratios as feature values (74% vs. 80%).

4.2 Experiment Two

The purpose of the second experiment is to group nouns into equivalence classes prior to classification and examine the corresponding effect on data coverage and classification accuracy. As mentioned in Experiment 1, the most comprehensive feature set (combined subject and object counts) only covers 40% of our data points. Therefore, we were interested in a way of forming noun classes that does not depend on feature counts.

We realized that many of the items in our data set are morphologically similar to compound nouns. Most compound nouns are subtypes of the head noun (e.g., *sports car* is a type of *car*), and compound nouns with a common head often share properties of the head, including its animacy. For example, in English, compounds such as *postman*, *fireman*, *salesman* are all types of *man*, and for the purposes of gross animacy classification further distinctions are not necessary. Many Japanese nouns are morphologically similar to the compounds in the English example above. In particular, it is common for words of Chinese origin to have a compound-like morphology, and Japanese orthography often makes this structure explicit. For example, a number of words end in the suffix *-jin* 'person' (orthographically the single character 人): *kajin* `poet', *kyojin* 'giant', *shuujin* 'prisoner', *tatsujin* 'expert', etc. Another class of words ends in the suffix *-hin* 'manufactured good' (orthographically the single character 品): *shinsouhin* 'bedding', *buhin* 'parts', *youhin* 'supplies', *shouhin* 'prize', etc. In both cases, the Japanese morphology and orthography provide a type of surface homogeneity not as readily available in the English equivalents.

We formed suffix classes of Japanese nouns by grouping all the items ending with the same kanji (i.e., the same character such as *-jin*, *-hin*, etc.). Although there are cases where the final character is not acting as the head (e.g., *satsujin* 'murder'), we were reasonably confident in the consistency afforded by this approach and did not try to eliminate such cases. Once the suffix classes were formed, we obtained the subject and object counts for the class. For example, given a suffix class of *-jin*, we incremented feature counts for this class any time *kajin*, *kyojin*, etc. appeared. We then applied this feature vector to each member of the class, so that *kajin*, *kyojin*, etc. have identical feature vectors. As with the average verb animacy ratio, this application of suffix classes eliminates many item-level distinctions. However, recall and precision for both classes should increase, given much denser feature vectors.

Figure 1 shows the effect of forming suffix classes on the distribution of item frequency in our data set. The dashed line shows the cumulative probability distribution of items based on their subject or object counts before applying suffix classes. As the dashed line in Figure 1 indicates, the data set is initially sparse, with about 77% of the items occurring fewer than 10 times.

The solid line in Figure 1 shows the cumulative probability distribution of item frequency after forming suffix classes. The effect of the suffix classes on the cumulative probability distribution manifests itself in the graph in short regions of steep slope, which correspond to groups of identical feature vectors occurring at a particular frequency. We are able to account for 11% our data set that does not occur with any of our features by virtue of inclusion in a suffix class. Moreover, about 75% of our data set now occurs 250 times or fewer, as opposed to fewer than 10 times, indicating that the mass of the cumulative probability distribution has shifted considerably.

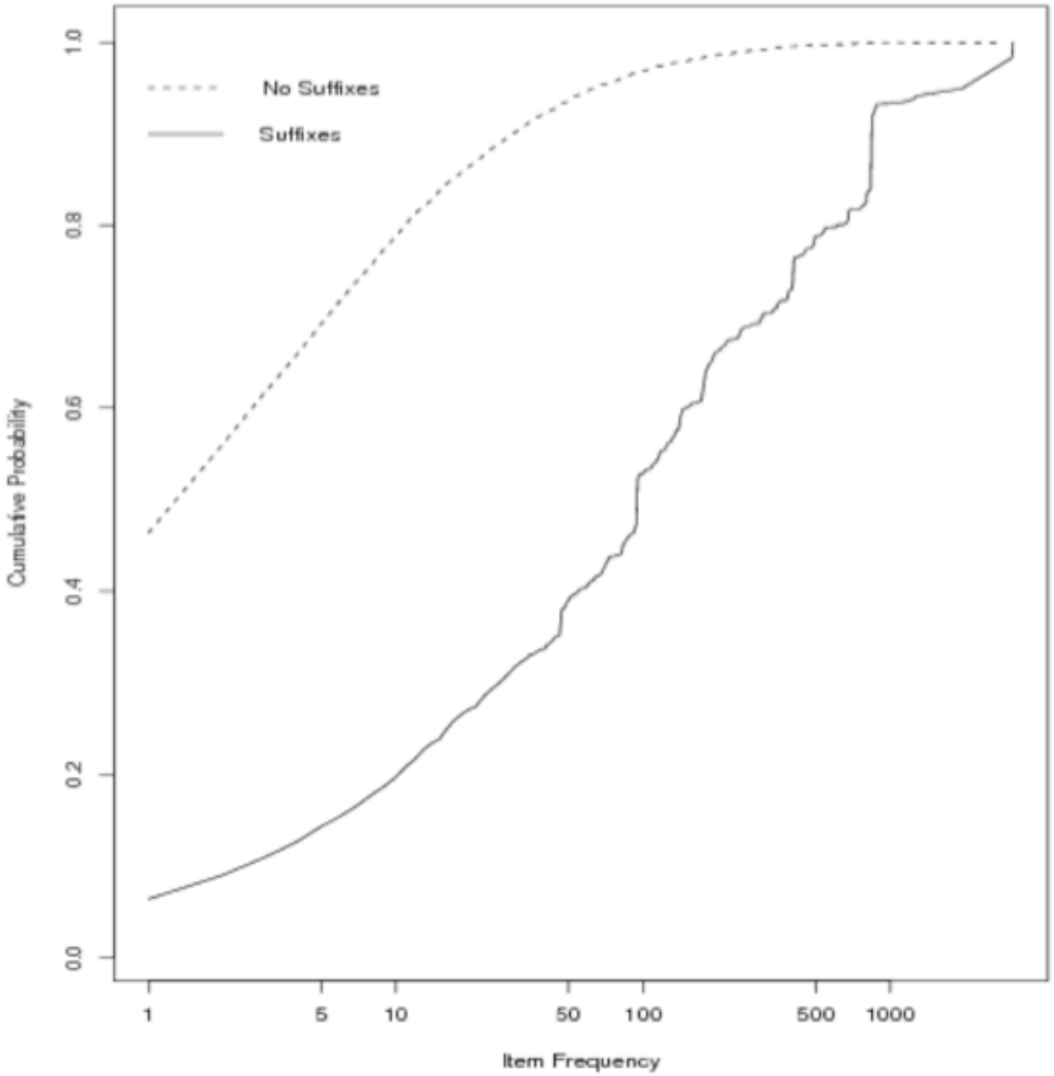

Figure 1. Cumulative probability distribution of item frequency with suffix classes

(dashed) and without (solid).

Table 7 shows the extent to which collapsing individual items into suffix classes reduced the size of the data set.

	Size	w/ suffix class	Reduction
Animate	4433	1892	57%
Inanimate	9956	8613	13%

Table 7. Effect of suffix classes on compressing the data set.

Applying suffix classes to the animate nouns resulted in a 57% reduction in the size of the class; for the inanimate items, we obtained a reduction of only 13%. The main reason for this disparity is the proportion of English origin items in the two classes: 74% of the inanimate nouns were English loanwords (hence, not members of a cohesive suffix class), but only 40% of the animate nouns were English loans.

Table 8 contains the classification results for the second experiment.

Feat	Cvg (%)	Acc (%)	Inanimate		Animate	
			Prec (%)	Rec (%)	Prec (%)	Rec (%)
Subject (Object) Frequency						
Subject	46	93.8	90.2	99.5	99.3	86.0
Object	49	94.3	91.2	99.5	99.3	87.7
Subj+Obj	51	93.6	90.4	99.5	99.3	85.6
Verb Animacy Ratio						
Subject	46	94.7	93.4	97.3	96.5	91.7
Object	49	95.3	94.4	97.5	96.6	92.6
Subj+Obj	51	94.6	93.5	97.2	96.1	91.0
Average Verb Animacy Ratio						
Subject	46	93.1	95.8	91.5	90.2	95.1
Object	49	93.8	95.4	93.4	91.8	94.2
Subj+Obj	51	93.4	95.4	93.0	90.8	94.0

Table 8. Classification results for Japanese using suffix classes.

The most frequent baseline for the covered portion of the data set is 58% (inanimate). Overall classification accuracy increases to above 95%, while coverage increases to 51% of the data set. Precision and recall of the animate class shows the largest improvement from Experiment One (up 16% and 34%, respectively) and is now on par with precision and recall for the inanimate class (93% and 96%, on average).

The effects of the ratios versus counts is visible in this less sparse data set. For the animate class, precision drops from 99% using counts to 96% using verb animacy ratios, and to 91% for the single-feature vector of average verb animacy ratio. Recall increases by a few points across each feature type from 87% (counts) to 94% (average verb animacy ratio). Precision and recall of the inanimate nouns shows the opposite effect: precision increases slightly from 91% (counts) to 96% (average verb animacy ratio) as recall decreases from 100% (counts) to 93% (average verb animacy ratio). We assume that the effect of using ratios is greater for the animate items, mainly because suffix classes resulted in a much greater reduction in the size of the animate noun class. As animate recall increases, inanimate precision increases because there are fewer incorrectly tagged animate items; as the number of incorrect animate predictions increases, recall of the inanimate class decreases.

4.3 Experiment Three

Even after forming suffix classes, we are able to cover only half of the data set. Most of the zero-frequency items are English loanwords. Since we have the English transliteration of each loanword, the purpose of Experiment 3 is to examine the feasibility of transferring animacy distinctions acquired in English onto Japanese data. We do not expect the English classifier to be as reliable as the Japanese one, because English is not particularly noted for robust sensitivity to animacy. However, we do expect performance to be better than chance.

For the English animacy classification, we extracted subject-verb pairs from the English Gigaword corpus (Graff 2003) using MiniPar (Lin 1995), a broad-coverage English dependency parser. Because data sparsity was less of an issue, we restricted our feature set to subject counts of transitive verb instances (i.e., verbs that occurred with a subject and object).

To create our training data, we translated the non-English words in the original data set into English using Babel Fish[5]. This resulted in 3302 training items (many items translated into the same word in English), two thirds of which were inanimate. There were 6917 test items (transliterations of the English loanwords); 5629 (81%) of these were inanimate and 1288 were animate. As with the Japanese suffix classes, we collapsed multi-word compounds into single categories before classification (e.g., *summer camp*, *day camp*, etc. → *camp*).

Table 9 contains the results of the English animacy classification. Overall, subject counts performed the best (88% correct), and precision and recall for the two animacy classes is similar to the results for Japanese using subject counts (Experiment 1, Table 6).

Accuracy	Inanimate		Animate	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Subject Frequency				
88.0	89.2	97.0	78.7	48.8

[5] http://babelfish.altavista.com

Verb Animacy Ratio				
79.7	92.4	81.9	47.0	70.3
Average Verb Animacy Ratio				
81.4	81.4	100	100	0

Table 9. English-only results on the loanwords.

The effect of substituting feature values with verb ratios is even more clearly visible on this less sparse data set (80% of the English items occurred more than 10 times, vs. 40% of the Japanese items with suffix classes applied).

The single feature vector containing a noun's average verb animacy ratio does not work for the English data. The most likely explanation for this fact is illustrated in Figure 2, which contains the distribution of verb animacy ratio for Japanese nouns versus the English test items.

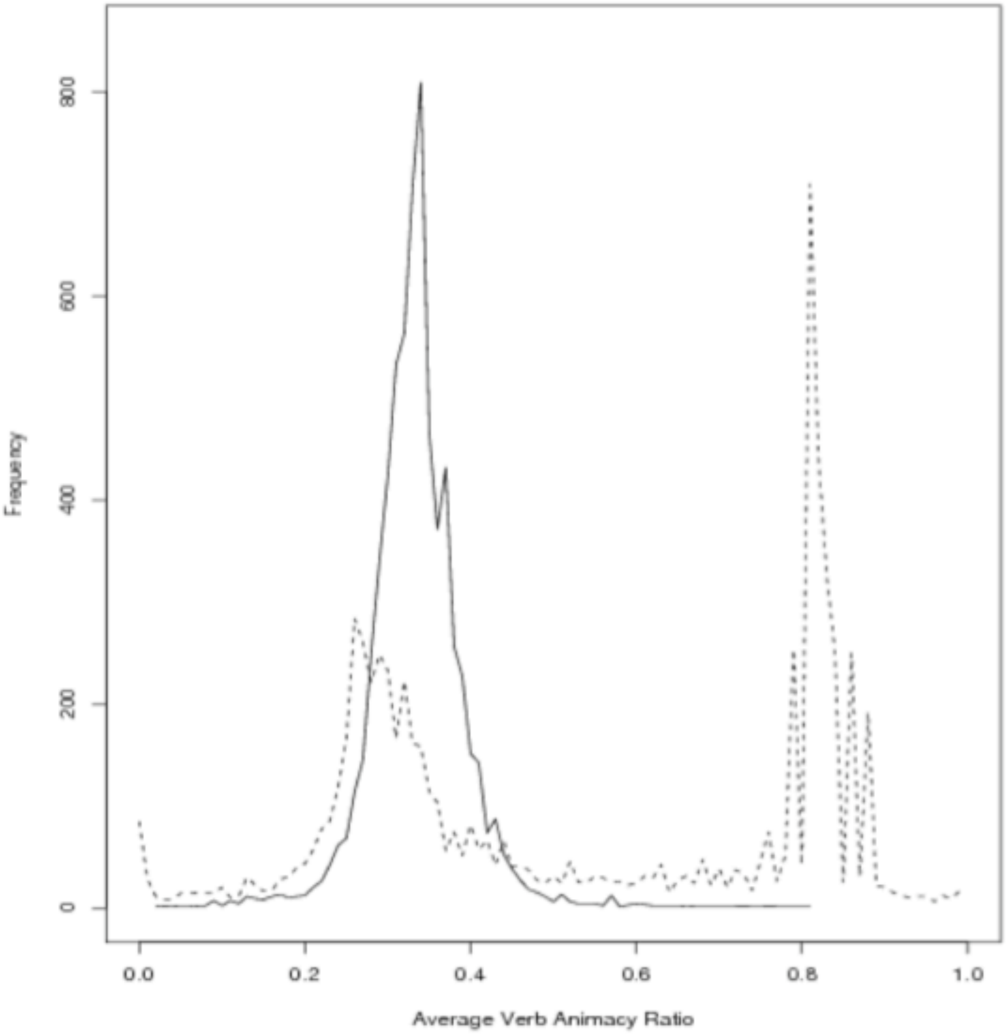

Figure 2. Distribution of Average Verb Animacy Ratio for Japanese nouns (dashed) and English test items (solid).

The English items are unimodally distributed around a mean of 0.36, meaning that most of their features are biased towards inanimates regardless of the item's true class. On the other hand, the Japanese data is bimodally distributed, indicating why this feature is a better predictor for the Japanese data.

Table 10 shows the classification results of the Japanese data with the English labels applied to the loanwords. The most frequent baseline is 70% (inanimate).

Feature	Coverage	Accuracy	Inanimate		Animate	
			Prec (%)	Rec (%)	Prec (%)	Rec (%)
Subject + Object Frequency						
Jp	97	87.6	84.9	99.8	99.1	60.2
Jp + En	97	86.6	91.1	89.3	77.1	80.5
Verb Animacy Ratio						
Jp	97	87.9	86.1	98.4	94.7	64.5
Jp + En	97	86.4	93.6	86.3	73.9	86.7
Average Verb Animacy Ratio						
Jp	97	86.8	87.1	95.0	85.8	68.5
Jp + En	97	85.8	94.4	84.5	71.9	88.7

Table 10. Japanese results with English transfer. Jp=Japanese baseline (inanimate), Jp+En=labels transferred from English.

We used subject and object occurrences for the Japanese feature set, as this performed the best for each feature type. Regardless of the Japanese feature type, the English labels were applied using the results of the subject counts.

With the English transfer, coverage increases to 97% of the data set (the remaining 3% are Japanese nouns that were not parsed as subjects or objects of verbs in the feature set). In every case, we get better inanimate precision and animate recall using transferred labels versus applying the default (inanimate) label to the same data set. Overall
accuracy is not different in each feature type, but within-class precision and recall are.

5. Discussion

Overall, classification accuracy was higher for Japanese than English using comparable feature sets. In particular, classifying animate items is more reliable for Japanese (precision ≈96, recall ≈91) than English (precision ≈79, recall ≈49). This disparity may result from noise arising from the language transfer or differences in how the two languages lexicalize animacy.

In general, the reliability of the transferred animacy labels seems to be reasonable: 99% of the English translations that appear in WordNet2.0 of the Japanese inanimate nouns (2293) have only inanimate senses, and 97% of the English translations that appear in WordNet2.0 of the Japanese animate nouns (1008) are listed as unambiguously animate. This fact suggests that the difficulty lies in English verbs' relative lack of sensitivity to the animacy of their subjects. Figure 3 compares the distribution of verb animacy ratios for English and Japanese. An animacy ratio of 0.0 means that verb occurred exclusively with inanimate subjects, and an animacy ratio of 1.0 means that verb appeared only with animate subjects.

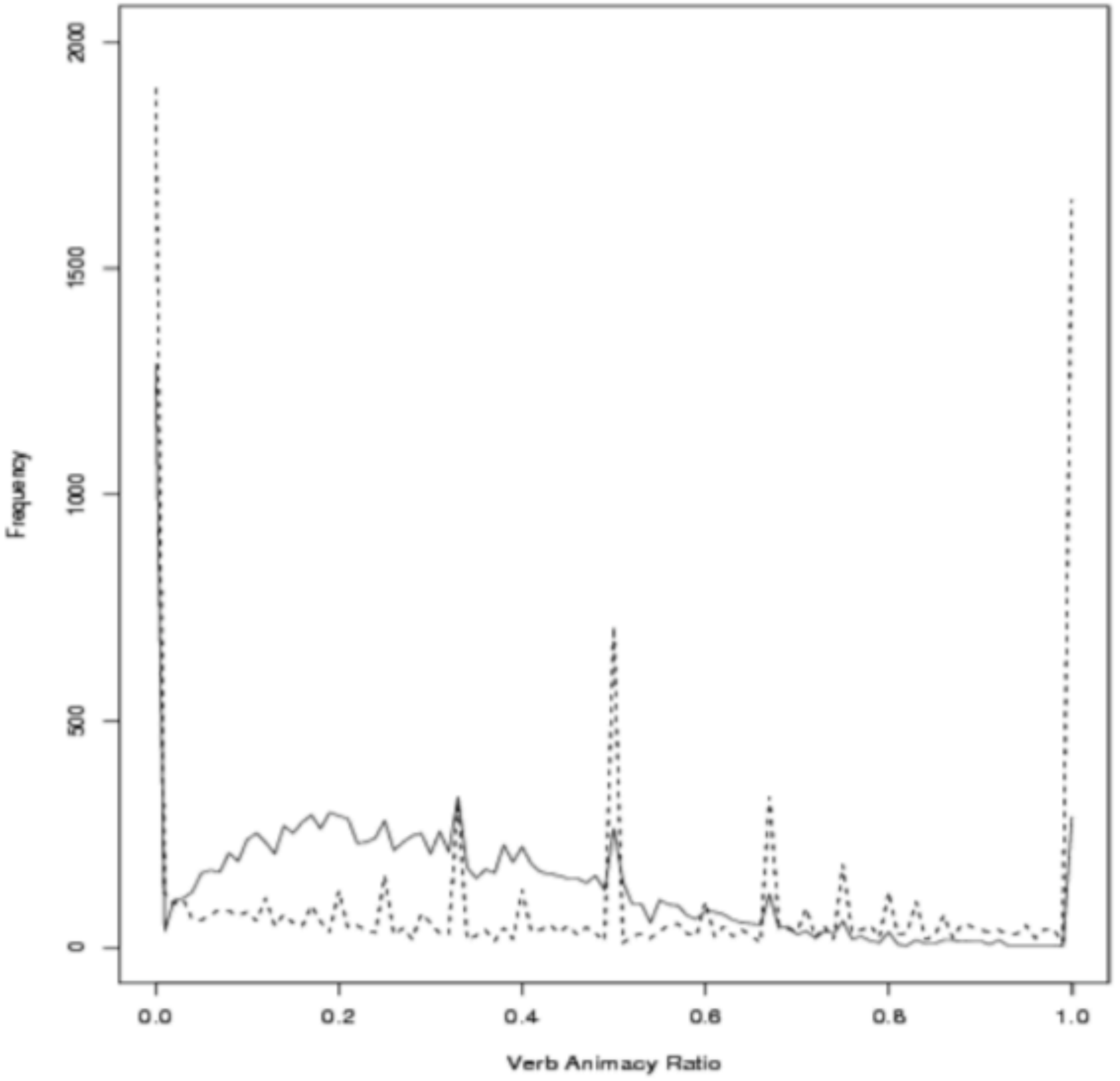

Figure 3. Distribution of Verb Animacy Ratios for Japanese (dashed) and English (solid) verbs.

Both languages exhibit peaks at the extremes of the scale, and for both languages the number of exclusively animate verbs is less than the number of exclusively inanimate verbs. In English, however, most verbs are biased towards inanimate subjects, with most of the frequency mass between 0.0-0.7. Japanese exhibits a third peak at 0.5, but most of the frequency mass is concentrated at the extrema. The distributions in Figure 3 indicate that the Japanese feature set better partitions the data into the two animacy classes than the English feature set does. This fact calls into question the general cross-language applicability of relying solely on verbs as features: the criterion of good coverage seems to hold, but the animacy distribution afforded by the English feature set appears too weak for reliable classification.

Like Oräsan and Evans (2001), we found that animate items are harder to classify than inanimate ones, and offer an explanation for this phenomenon on the basis of the distribution of items in our data set. Intuitively, animate things are capable of a wider range of different actions than inanimate things are. This difference may be reflected in language data as a disparity between the number and dispersion of verbs associated with the two animacy classes.

Examination of the data set shows that across the two languages, there are approximately three times as many animate features as animate items, whereas the number of inanimate features is roughly equal to the number of inanimate items. For both languages, animate subjects are associated with a larger number of verbs than inanimate subjects are. Conversely, each verb is associated with fewer animate subjects than inanimate subjects. Animate nouns may be harder to classify because each item occurs with a larger set of different features, making each animate feature vector relatively unlike any other.

6 Conclusion and Future Work

This paper presented the results of three experiments on automatic animacy classification in Japanese and English in which we focused on classifying infrequent items. Animacy classification for Japanese was more reliable than for English, largely because English verbs are less sensitive to the animacy of their arguments. Replacing feature counts with verb animacy ratios resulted in improved classification accuracy for the harder-to-classify animate items. The biggest gains in classification accuracy resulted from placing items into suffix-based equivalence classes prior to classification. We further demonstrated the feasibility of language transfer from English to Japanese using loanwords as conduits for lexical semantic annotation. By exploiting lexical surface cues to animacy in Japanese that are not available in English, we were able to create a training set for an English classifier and transfer the acquired labels back onto the loanwords in Japanese.

Future work will look at aspects of multilingual lexical acquisition touched on in this paper; in particular, it will focus on exploiting the robust animacy lexicalization in Japanese for making improved animacy distinctions in English. We will examine the feasibility of classifying the relatively small set of frequent English loanwords using Japanese corpus data, and extending those labels to a larger set of English words via a semi-supervised learning technique such as manifold regularization (e.g., Belkin et al. 2004). Using loanwords is appealing for this task because their transliteration can be automated (e.g., Knight and Graehl 1998) lessening the dependence on external lexical resources.

Acknowledgments

We are indebted to James Unger, Hiroko Morioka, Brian Joseph, Peter Culicover and Adriane Boyd for their help and suggestions on this paper. All errors are ours alone. Funded by NSF CAREER Grant 0347799 to Chris Brew.

References

Baker, Kirk. 2008. Multilingual distributional lexical acquisition. PhD Dissertation. The Ohio State University.

Belkin, Mikhail, Partha Niyogi and Vikas Sindhwani. 2004. Manifold regularization: a geometric framework for learning from examples. *University of Chicago CS Technical Report TR-2004-06*.

Bliss, Heather 2005. Topic, focus, and point of view in Blackfoot. In John Alderete (ed.), *Proceedings of the 24th West Coast Conference on Formal Linguistics*, 61-69. Cascadilla Proceedings Project. Somerville, MA.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina and Harald Baayen. 2005. Predicting the dative alternation. In *Proceedings of Royal Netherlands Academy of Science Workshop on Foundations of Interpretation.*

Bresnan, Joan and Jennifer Hay. 2006. Gradient grammar: an effect of animacy on the syntax of give in varieties of English. Electronic manuscript. www.stanford.edu/~bresnan/anim-spokensyntax-final.pdf.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch. 2003. TiMBL: Tilburg Memory-Based Learner version 5.0 Reference Guide. *ILK Technical Report* – ILK 03-10.

Dahl, Östen and Kari Fraurud. 1996. Animacy in grammar and discourse. In Thorstein Fretheim and Jeanette K. Gundel, eds. *Reference and Referent Accessibility*, pp 47-64. John Benjamins.

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67 (3): 547-619.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Gawronska, Barbara, Björn Erlendsson and Hanna Duczak. 2002. Extracting semantic classes and morphosyntactic features for English-Polish machine translation. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machines Translation*.

Genkin, Alexander, David D. Lewis and David Madigan. 2004. Large-scale Bayesian logistic regression for text categorization. *DIMACS Technical Report*.

Graff, David. 2003. English gigaword. *Linguistic Data Consortium*, Philadelphia. LDC2003T05.

Graff David and Zhibiao Wu, 1995. Japanese business news text. *Linguistic Data Consortium, Philadelphia*. LDC95T8.

de Ilarraza, Arantza Díaz, Aingeru Mayor and Kepa Sarasola. 2002. Semiautomatic Labelling of Semantic Features. In *Proceedings of the 19th International Conference on Computational Linguistics* (COLING 2002).

Iida Asako. 1999. A descriptive study of Japanese major classifiers. Electronic manuscript. http://www5b.biglobe.ne.jp/~aiida/ephd.html.

Jäger, Gerhard and Anette Rosenbach. 2006. The winner takes it all – almost: Cumulativity in grammatical variation. *Linguistics* 44(5): 937-972.

Joseph, Brian. 1979. On the animate-inanimate distinction in Cree. *Anthropological Linguistics* 21(7): 351-354.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2005. Unsupervised word sense disambiguation using bilingual comparable corpora. *IEICE Transactions on Information and Systems* E88 D (2): 289-301.

Knight Kevin and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics* 24: 599-612.

Kudo, Taku and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the CoNLL-2002*, 63-69.

Li, Jianguo, Kirk Baker and Chris Brew. 2008. A corpus study of Levin's verb classification. *American Association of Corpus Linguistics* (AACL-2008). Provo, Utah.

Li, Jianguo and Chris Brew. 2008. Which are the best features for automatic verb classification? *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Columbus, OH.

Lin, Dekang. 1995. A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of the 1995 International Joint Conference on Artificial Intelligence* (IJCAI).

Madigan, David, Alexander Genkin, David D. Lewis and Dmitriy Fradkin. 2005. Bayesian multinomial logistic regression for author identification. *DIMACS Technical Report*.

Oräsan, Constantin and Richard Evans. 2001. Learning to identify animate references. In *Proceedings of the Conference on Computational Natural Language Learning* (CoNLL-2001), 129-136.

Øvrelid, Lilja. 2006. Towards robust animacy classification using morphosyntactic distributional features. In *Proceedings of EACL 2006 Student Research Workshop*. Trento, Italy.

Palmer, Martha, Dan Gildea and Paul Kingsbury. 2005. The proposition bank: a corpus annotated with semantic roles. *Computational Linguistics* 31 (1): 71-106.

Redington, Martin, Nick Chater and Steve Finch. 1998. Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science* 22: 425-469.

van Nice, Kathy Y. and Rainer Dietrich. 2003. Task-sensitivity of animacy effects: evidence from German picture descriptions. *Linguistics* 5: 825-849.

Woolford, Ellen. 1999. Animacy hierarchy effects on object agreement. In Paul F. A. Kotey (ed.), *New Dimensions in African Linguistics and Languages*, 203-216. Africa World Press, Inc.

Yamamoto, Mutsumi. 1999. Animacy and Reference: *A Cognitive Approach to Corpus Linguistics*. John Benjamins Publishing, Cambridge, MA.

Zaenen, Annie, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor and Thomas Wasow. 2004. Animacy encoding in English: why and how. In *Proceedings of the ACL Workshop on Discourse Annotation*, 118-125.