Deep Lexical Acquisition of Type Properties in Low-resource Languages: A Case Study in Wambaya

Jeremy Nicholson,[†‡] Rachel Nordlinger[‡] and Timothy Baldwin[†‡]
† NICTA Victoria Research Laboratories
‡ The University of Melbourne, VIC 3010, Australia
{nj,racheln,tbaldwin}@unimelb.edu.au

Abstract

We present a case study on applying common methods for the prediction of lexical properties to a low-resource language, namely Wambaya. Leveraging a small corpus leads to a typical high-precision, low-recall system; using the Web as a corpus has no utility for this language, but a machine learning approach seems to utilise the available resources most effectively. This motivates a semi-supervised approach to lexicon extension.

1 Introduction

Deep lexical acquisition (DLA) is the process of (semi-)automatically creating or extending linguistically-rich lexical resources (Baldwin, 2005a; Baldwin, 2005b; Baldwin, 2007). Conventionally, DLA has been applied to high-resourced languages such as English, German or Japanese to broaden the coverage of a medium-coverage resource, or enrich existing linguistic annotation in resources. However, it also has tremendous potential in accelerating the documentation of low-density languages, a fact that is often discussed but very rarely delivered on in the literature. This paper attempts to deliver on this promise, and asks the question: how well do standard approaches to DLA perform over low-density languages? For example, one of the standard approaches to DLA is to extract n-gram counts for patterns involving a target lexeme from the web, and use these as the basis for predicting the lexical class membership of the lexeme. While there is little expectation that we will find significant amounts of text for low-density languages on the web, we nevertheless run the experiment to test the general applicability of this style of approach.

In this work, we take Wambaya as a real-world chronically low-density language, and examine the task of predicting the grammatical gender of nominal lexical items. Wambaya is a nearly extinct language (Gordon, 2005) from the Mirndi group of Australian languages. Like many languages from the Australian family, its complicated syntax and rich morphology makes natural language processing of Wambaya difficult. Unlike many of its neighbours, however, it has been well documented in a descriptive grammar (Nordlinger, 1998) and a Head-driven Phrase Structure Grammar (Bender, 2008). While resources for Wambaya are of little intrinsic value as it is doubtful that new text will be generated in the language, developing these resources is still instructive for parallel development in comparable languages (Warlpiri, for example, has a notable speech community (Gordon, 2005)). Additionally, it provides an invaluable test bed for DLA research, to test the potential of methods over similarly low-density languages, and truly test the bounds of DLA for the purposes of language preservation.

Lexicon extension for Wambaya is a task comparable to the state of many resources: the available lexicon is small, of only about 1500 entries. About half of these are nominals, which is the focus of this research. Furthermore, the sum total of available data in the language on which to base our methods is minimal: fewer than 5000 words across about 1100 sentences. We identify instances of the nominals in the small corpus, and examine standard machine learning approaches based on evidence in terms of lexically-disambiguating surface cues, which are intended as a proxy for features which could easily be

26th Pacific Asia Conference on Language,Information and Computation pages 481–490

designed with minimal assistance from a lexicographer familiar with the language.

While using surface cues to observe lexical properties has seen broad study in a number of languages, Wambaya represents a relatively extreme case in terms of difficulty: since NPs are often discontiguous, a given modifier that carries the grammatical marking of a token can be outside any reasonable context window.

(1) ***Garngunya*** *gin-aji* *yabu*
many.II 3SG.M.A-HAB.PST have
garirda-rdarra *garndawugini-ni.*
wife.II-GROUP one.I-ERG
"One [man] used to have many wives"

In (1), the modifier *garngunya* "many" of the class II noun *garirda* "wife", appears initial to the sentence, and agrees in gender and grammatical number with its displaced head. The behaviour is somewhat similar to referential pronouns in English, but can occur with any modifier. Having this discontiguity makes identifying surface cues problematic; in addition, the rich morphology means that even identifying token instances of a given type is non-trivial, as a lemma typically has hundreds of inflected forms.

Our approach is to take a standard inventory of DLA techniques and apply them naively to Wambaya, to gauge their effectiveness over a truly low-density language, with the added complexity of non-configurationality and complex morphology.

We will demonstrate that a number of strategies that have been shown to be competitive in some languages (primarily English) unsurprisingly perform poorly for Wambaya. Machine learning, on the other hand, is remarkably effective, with minimal feature engineering.

2 Background

2.1 Wambaya

Wambaya is a critically endangered Australian language (Nordlinger, 1998), spoken by only a handful of people in the Northern Territory, Australia. The language is radically non-configurational, with very free word order apart from a verb clitic cluster in second position. It is a split ergative language, with nominative–accusative pronouns and ergative–absolutive nominals otherwise. There are about nine nominal cases,[1] as well as four adnominal cases that further inflect for grammatical gender; there are furthermore three grammatical numbers: a singular, a dual, and a plural. In this work, we examine the four grammatical genders: semantically, class I and class II loosely correspond to masculine and feminine animates, class III to non-flesh food items and some round body parts, and class IV to the semantic residue. Gender morphology in Wambaya is mostly regular, but this is less true in other Australian languages, often because of vowel harmony, so we focus primarily on morphosyntax.

Nordlinger's grammar has been implented in a Head-driven Phrase Structure Grammar (HPSG; Bender (2008)) as part of an analysis of the LinGO Grammar Matrix (Bender et al., 2002; Bender and Flickinger, 2005; Drellishak and Bender, 2005; Bender et al., 2010). We use the lexical items from the HPSG lexicon to construct a set of nominal types marked for gender. There are 786 class assignments for 724 distinct nominals; their distribution is shown in Table 1.

I	II	III	IV
233	199	51	303

Table 1: Distribution of classes for Wambaya nominals.

Most of the multi-class items are animates (humans and animals) that belong to both class I and class II (masculine and feminine). These pairs have the same stem, but different forms in the absolutive, which is the unmarked case from which the lemma is derived. For example, *alag-* "child" can be realised as the class I absolutive nominal *alaji* "boy" or the class II absolutive nominal *alanga* "girl."

For each item in the lexicon, we use Bender's implementation of the grammar to generate the (absolutive) lemma from the stem, as well as all of the inflected forms that are licensed by the grammar. Nominals were observed to have between about 400 and about 2200 distinct inflected forms. We construct surface cues based on demonstratives in the language: Nordlinger indentifies four singular absolutive proximal demonstratives (one for each gender class), and 62 demonstratives overall (24 for each of class I and II, and 7 for III and IV), for proximal and

[1] There is some disagreement as to the exact number.

distal demonstratives in nominal classes. 28 of these do not occur in the corpus (described below). The demonstratives we examined appeared to usually act as deictic determiners qualifying a nominal, but they also occured as pronouns; we chose not to examine comitative and possessive demonstratives, or indefinites or interrogatives, which appeared to function more often as pronouns.

Along with the grammar is a treebank of sentences and phrases that occur in Nordlinger's descriptive grammar, combining the inline linguistic examples and eight provided transcribed texts. These amount to 1131 unique sentences (many of the sentences from the text were also used as linguistic examples): about a third of these were from the texts. We used these sentences — without the syntactico-semantic annotation from the treebank — as a raw corpus of Wambaya.

2.2 Lexical properties

The analysis of lexical information is often done on individual tokens, often under the banner of "lexical disambiguation". Some examples are context-sensitive spelling correction (Banko and Brill, 2001), selecting between target candidates for machine translation (Grefenstette, 1998), and determining the semantic gender of nouns in context (Bergsma et al., 2009). All of these were based on English data. Lapata and Keller (2005) examine a range of English tasks whereby frequencies of events can be used as evidence for the disambiguation. They assert that using Web page counts as a *de facto* corpus is a model that is generally as good as, or better than, established results in the field.

Lapata and Keller also examine a type-level task: that of the countability of English nouns (Baldwin and Bond, 2003). In this type of task, the token context is not available, and context must instead be generated to observe evidence. They construct a set of surface cues — *much* and *many* to disambiguate mass and count nouns respectively — and extract evidence from these. The performance is good, but not as high as that which Baldwin and Bond observe by using more sophisticated tools such as chunkers. A similar experiment was performed by Nicholson and Baldwin (2009), for a set of about 50 count classifiers in Malay; again, the Web was observed to be a strong performer for observing useful evidence.

As for grammatical gender, research has tended to focus on Indo-European languages. Hajič and Hladká (1997) examine grammatical gender in Czech as part of the part-of-speech tagging process. In Czech, morphological surface cues on a noun token give a strong indication of gender; more so in a stream of tokens where modifier inflection can also be taken into account. This method would probably also be effective for Wambaya, due to its mostly regular gender morphology. Morphological surface cues were also motivated for lexical semantics of derivational morphology by Light (1996). Finally, Cucerzan and Yarowsky (2003) explore a minimally-supervised approach for the prediction of grammatical gender of a mixture of tokens and types by extracting contextual cues from a seed set of nouns and bootstrapping to morphological cues. Token-level performance is high for the five languages they examine.

3 Methodology

Based on standard DLA methodology, we examine three prediction methods for Wambaya nominals:

- co-occurrence frequencies with demonstratives from a Wambaya corpus;
- co-occurrence frequencies with demonstratives from Web page counts estimated using the Yahoo! API[2]; and
- machine learning using context windows around token instances identified from the corpus.

As stated in Section 1, the selection of methods at this level is not intended to reflect any keen insights into Wambaya so much as a standard inventory of DLA methods, which we apply to the task.

Note that these features attempt to leverage token–level observations into type–level information; if we were examining token predictions in a tagging framework, then the feature engineering approaches for POS tagging as performed by Hajič and Hladká (1997) or morphological analysis in Chrupała et al. (2008) could provide further sophistication.

[2] http://developer.yahoo.com/search/

3.1 Corpus Frequency

Corpus frequency-based methods involve identifying lexical cues in the given language, and using observation of the relative frequency of each cue to classify instances. The frequencies are based on a monolingual corpus of the language, in our case, the small set of 1131 unique sentences of Wambaya.

We use cues observed from the Wambaya corpus as evidence for the gender of a lexical item. This is based on the intuition that a given nominal will only co-occur with demonstratives that agree in gender.

(2) ***Ngangaba*** ***yana*** *gi-n*
fire.IV.ABS this.IV.SG.ABS 3SG.S.PR-PROG
najbi.
burn
"There's a fire burning [here]."

We consider instances like (2) as evidence that the nominal *ngangaba* is of class IV, because *yana* is a class IV demonstrative.

Wambaya has a rich inflectional morphology, so that a given token instance of a nominal within a corpus can display one of hundreds of surface forms. The surface cues also display rich inflectional paradigms. When collating our corpus counts for a given lexical item, we attempt three different strategies of dealing with this phenomenon.

The first, ABS, assumes we have access to the absolutive form of the lexeme. This is the least-marked form, and also the lemma. For inflectional agreement, the corresponding surface cue must also be in the absolutive form; here we use the singular proximal absolutive demonstrative for each of the four gender classes. A sentence where both the absolutive nominal and an absolutive demonstrative occur is considered to be a positive count for the corresponding gender class. Although an NP can be discontiguous, sentences where the demonstrative is in direct apposition to the nominal can provide stronger evidence — as such, we also consider a cue strategy for instances in direct apposition (either pre-modifying, post-modifying, or either). A short example is shown for the absolutive nominal *alaji* across Examples (3)–(6) in Table 2.

Alternatively, with access to a morphological analyser, we could generate all of the possible inflected surface forms for a given nominal (INFL). On average, this is about 700 different forms. Here, we do not attempt to enforce morphological agreement: if any form of the nominal co-occurs with any cue, we consider that to be positive evidence. There are 62 fully-inflected demonstratives given by Nordlinger, and the aggregated count for a class is the sum of all of the sentences where one of the corresponding surface cues occurred. We again contrast direct apposition with sentence co-occurence. Table 3 shows the counts for *alaji* in the given examples.

If no morphological analysis tools are available, we could simply search for the stem (STEM); since morphology in Wambaya is primarily suffixing, we allow any number of other characters to optionally follow the stem. In this case, we consider both of the above cue strategies: the four absolutive nominals, where the stem is a proxy for the absolutive form, or all sixty-two, where the stem is a proxy for the entire set of inflected forms. This method fails for the given examples below, because none of the inflected forms of *alaji* begin with the stem *alag-*. About 30% of the stems in the lexicon are homologous with the corresponding absolutive form; many are proper prefixes thereof.

Classification proceeds by choosing the most frequent aggregated count; in most cases, this is the only non-zero count. Because of the small corpus and the sparsity of the cue set, we also explored a classification routine where any non-zero count is treated as a positive classification: predictably, a small boost in recall is traded off with a small drop in precision. Across the Wambaya corpus, these differences are not statistically significant at the 0.05 level, and are not reported in detail.

3.2 Web-as-Corpus Frequency

The methodology for using the Web as a corpus is very similar to the corpus frequency approach, except that page count estimates returned by a search engine are used in place of actual observed instances. The assumption that these values are strongly correlated was found to be accurate by Keller and Lapata (2003) for a range of classification tasks.

At first glance, using the Web to estimate corpus counts for a language close to extinction is patently absurd, as there is no speech community generat-

(3) *Gulug-ardi* *ng-u* ***ini*** ***alaji.***
sleep-CAUS(NF) 1SG.A-FUT this.I.SG.ABS boy.I.ABS
"I'm going to put this boy to bed."

(4) *Garnguji* *nyi-n* *yabu* ***alaji.***
many.I.ABS 2SG.A.PR-PROG have boy.I.ABS
"You have a lot of kids."

(5) ***Alangi-nka*** *yana* *jalyu.*
boy.I-DAT this.IV.SG.ABS bed.IV.ABS
"This is the boy's bed."

(6) *Jawaranya* *ng-u* *yidanyi* *ngaba* *ng-u* *yardi* *yaniya* *cool* *drink*
billycan.II.ABS 1SG.A-FUT get then 1SG.A-FUT put that.IV.SG.ABS cool drink.IV
ninaka *nanga* ***alangi-nka.***
this.I.SG.DAT 3SG.M.OBL boy.I-DAT
"I'm going to get the billycan and put that cool drink [in it] for the boy."

ABS	I (*ini*)	II (*nana*)	III (*mama*)	IV (*yana*)
Pre	1	0	0	0
Post	0	0	0	0
Pre/Post	1	0	0	0
No apposition	1	0	0	0

Table 2: Counts for examples (3)–(6) for the ABS paradigm of the class I nominal *alaji*

INFL	I	II	III	IV
Pre	1	0	0	0
Post	0	0	0	1
Pre/Post	1	0	0	1
No apposition	2	0	0	2

Table 3: Counts for examples (3)–(6) for the INFL paradigm of the class I nominal *alaji*

$t-4$	$t-3$	$t-2$	$t-1$	$t+1$	$t+2$	$t+3$	$t+4$
cool	*gulugardi*	*ngu*	*ini*	*yana*	*jalyu*		
	garnguji	*nyin*	*yabu*				
	drink	*ninaka*	*nanga*				

Table 4: Machine learning features based on the fully-inflected (INFL) forms of *alaji*, from examples (3)–(6)

$p1$	$p2$	$p3$	$p4$	$s1$	$s2$	$s3$	$s4$
a	*al*	*ala*	*alaj*	*i*	*ji*	*aji*	*laji*

Table 5: Machine learning features based on the prefixes and suffixes of the absolutive form of *alaji*

ing Web documents in that language. However, it may be the case that we actually observe documents which are linguistic descriptions of the target language, and not simply noise[3]. The ODIN project (Lewis and Xia, 2009) is an attempt to leverage such linguistic data into resources automatically. An additional reason for performing the experiment is that it is a standard DLA method which is used for higher-density languages, but there is no indication in the literature of how well to expect it to perform over low-density languages.

In our case, we experiment with the Yahoo! search engine API. Since the API rate-limits queries, we chose to only examine the ABS nominal set with the four proximal absolutive demonstratives, and the STEM set with the four demonstratives. We continued to contrast the surface cues in apposition, which were constructed as phrasal queries, with non-phrasal versions, i.e. that the demonstrative simply occurred in the same document as the nominal.

The frequencies that we observed from the Web were again sparse, but much less so than the corpus frequencies. Part of this was because of homology with wordforms in other languages; for example, the class I absolutive demonstrative is *ini*. Thresholding classification at zero frequency — that is, having a positive classification for any non-zero observation — becomes somewhat absurd over Web-scale data, particulary for the non-phrasal queries. Performance in these cases approaches that of the baseline classifier where every nominal is assigned to every class; the utility of this baseline is low.

3.3 Machine Learning

The third standard approach to DLA is machine learning, where a corpus provides not just frequency estimates of lexical patterns as for the corpus frequency approach, but the source of a potentially rich variety of features.

In applying the machine learning method to Wambaya, we relax the requirement for observing demonstratives. This is useful if a representative cue set is unknown or cannot be constructed. Instead, for each nominal in the data set, we identify corpus instances, and build feature vectors according to the tokens observed within a context window. We used a window size of up to four tokens, labelled for their distance from the target nominal; very little performance difference was observed when using different window sizes, possibly due to the fact that the average sentence length in the corpus was quite short. The feature values for all of the inflections of *alaji* for the given examples are shown in Table 4.

We then split the instances into training and test sets using 10-fold cross-validation. Our preferred machine learning model was the maximum entropy classifier.[4]; we do not expect substantial differences to result from using other types of machine learning models over this feature set. We also thresholded the classification so that if all classes were equally likely, no decision was made; otherwise, the class assigned with the greatest probability was chosen.

For contrast, we also built a model whose features were substrings of the nominal itself, rather than using contextual features. We considered prefixes of length 1 to 4 and suffixes of length 1 to 4, again varying this parameter was not observed to greatly affect performance. The feature vectors for the absolutive nominal *alaji* are shown in Table 5. This type of classification takes into account the regular morphological processes of Wambaya, and is consequently very effective, but would be less effective for many other Australian languages.

4 Results

For each methodology, we present the precision and recall, as well as the F-score. In fact, because of the low recall of most systems, the F-score is strongly correlated with recall, even though precision becomes the most interesting metric.

For the majority class baseline, that is, classifying every lexical item as class IV, the precision is 0.419 and the recall 0.385, for an F-score of 0.401. Most of the systems are well below this figure, due to low recall.

4.1 Corpus Frequency

The results of the corpus frequency assignment methods are shown in Tables 6 through 9, for Pre-modification frequencies, Post-modification fre-

[3]For example, of the top ten documents returned by Google for the query *ngabulu* "milk, breast", three are about Wambaya and another four are about Australian languages with a cognate.

[4]We used the OpenNLP implementation available at `http://www.sourceforge.net/projects/maxent/`.

quencies, the combination of those two, and frequencies where the demonstrative and nominal simply co-occur in a sentence.

The first notable fact is that recall is uniformly awful, where even the most generous system only classifies 49 instances correctly. On the other hand, precision is high, markedly higher than the baseline in almost all cases. Because there are so few instances being classified, it is difficult to draw significant comparisons between different systems; it seems though that there are fewer instances where the demonstrative post-modifies the noun, and that the methods that require apposition maintain higher precision and lower recall than the one that relaxes the requirement of contiguous NPs.

Context	Precision	Recall	$F_{\beta=1}$
Pre	0.944	0.022	0.043
Post	1.000	0.008	0.016
Pre/Post	0.950	0.024	0.047
No apposition	0.821	0.030	0.058

Table 6: Performance of corpus frequency assignment according to the four absolutive demonstratives over the set of absolutive nominals (ABS)

Context	Precision	Recall	$F_{\beta=1}$
Pre	0.824	0.018	0.035
Post	0.692	0.011	0.022
Pre/Post	0.783	0.023	0.045
No apposition	0.617	0.037	0.070

Table 7: Performance of corpus frequency assignment according to the four absolutive demonstratives over the set of nominal stems (STEM)

Context	Precision	Recall	$F_{\beta=1}$
Pre	0.794	0.034	0.065
Post	0.800	0.010	0.020
Pre/Post	0.784	0.037	0.071
No apposition	0.694	0.055	0.102

Table 8: Performance of corpus frequency assignment according to the full demonstrative set over the fully inflected nominal set (INFL)

4.2 Web-as-corpus Frequency

The results of the Web frequency assignment methods are shown in Tables 10 and 11, for the same observation methods as the corpus frequency approach, except that non-phrasal queries are now at the document level instead of the sentence level.

Context	Precision	Recall	$F_{\beta=1}$
Pre	0.718	0.036	0.069
Post	0.545	0.015	0.029
Pre/Post	0.673	0.042	0.079
No apposition	0.620	0.062	0.113

Table 9: Performance of corpus frequency assignment according to the full demonstrative set over the set of nominal stems (STEM)

Recall is generally not substantially higher than the corresponding approaches from the 1131 corpus sentences in Tables 6 and 7. As the precision is so much lower, significantly lower than the baseline in most cases, it appears that any classifications that this model makes correctly are completely accidental. If there is any useful evidence, it is swamped by extra-lingual or extra-linguistic material across the greater Web.

Context	Precision	Recall	$F_{\beta=1}$
Pre	0.304	0.031	0.056
Post	0.309	0.032	0.058
Pre/Post	0.315	0.037	0.066
Non-phrasal	0.238	0.086	0.121

Table 10: Performance of Web frequency assignment according to the four absolutive demonstratives over the set of absolutive nominals (ABS)

Context	Precision	Recall	$F_{\beta=1}$
Pre	0.216	0.031	0.054
Post	0.236	0.033	0.058
Pre/Post	0.244	0.041	0.070
Non-phrasal	0.177	0.074	0.104

Table 11: Performance of Web frequency assignment according to the four absolutive demonstratives over the set of nominal stems (STEM)

4.3 Machine Learning

The results of machine learning using the maximum entropy classifier are shown for the three nominal

sets in Table 12. The performance differences between the three sets were primarily caused by the number of corpus instances (and consequently feature windows) that could be observed for nominals in each set: for ABS, only 16.5% of the nominals were observed at least once in the corpus, compared to 26.7% and 31.7% for INFL and STEM respectively.

Feature set	Precision	Recall	$F_{\beta=1}$
ABS	0.511	0.087	0.149
INFL	0.634	0.214	0.320
STEM	0.713	0.281	0.403
MORPH	0.914	0.903	0.908

Table 12: Performance of the maximum entropy classifier over the various nominal sets

The precision of the model for all three datasets was markedly higher than that of the baseline. The low performance over the ABS dataset was primarily caused by sparsity of features: even though each instance was accurate, there were at best one or two windows with which to make a classification.

On the other hand, the fact that we saw higher performance across the STEM data set than the INFL dataset was surprising, particularly for precision ($\chi^2 = 3.87, p < 0.05$), somewhat less so for recall ($\chi^2 = 7.46, p < 0.01$). Examining the features, it was clear that there were erroneous corpus instances identified for the stem set, including verbs and other nouns that happened to share the first few characters. This makes its significantly higher performance all the more puzzling.

One observation from the feature sets was that some of the stems were wrong, or at least infelicitous in their interaction with the morphological generation component of the grammar to produce licensed wordforms in Wambaya. This would be artificially lowering the recall of the INFL set slightly, and possibly reducing the amount of discriminatory data for the model. This could also be affecting the STEM set, but it seems that many of the stems were proper prefixes of the lemma anyway.

The likely cause of the difference in precision between the STEM and INFL sets was that the machine learning model was picking up on spurious regularity in the corpus. Most of the sentences in the corpus were derived from inline linguistic citations, where it is often valuable to have a pair of sentences with minimal changes to highlight a particular property or construction. (For example, *Ngajbi gina ganggu yarruwarda* "He saw grandfather walking" and *Ngajbi gina gangguliji yarruwarda* "He saw his grandfather walking" to illustrate use of the reflexive-possessive suffix *-liji*.) If there was a morphological bias in the stems where corpus instances were observed for STEM and the inflected forms were not observed, that morphological bias could make classification easier because morphology is an accurate predictor of gender in Wambaya. To examine this, we attempted to construct the model using only the transcribed free text and not the linguistic inline citations, but this removed two-thirds of the data — consequently, the model struggled to classify any instances. One other possibility would be to train the model using features based on the linguistic citations and test on the features from the free texts.

Finally, we show results of using the pseudo-morphological features (prefixes and suffixes of the nominal of length 1 to 4) under MORPH in Table 12. When features were constructed from the lemma, both precision and recall were close to gold-standard, because gender is morphologically marked on the absolutive suffix. In some respects, this is a circular problem, because the gender must be known to correctly generate the absolutive form from the stem. If the morphological features are constructed from the stem instead of the lemma, accuracy drops to 68.6%. This approach is effective for Wambaya, but would be less so for many neighbouring languages.

5 Discussion

We presented several modes of classification for grammatical gender of nominals in Wambaya. Most of these had prohibitively low recall, showing that it is generally difficult to make such classifications, partly due to the complexity of the language, and partly due to the paucity of data available to provide evidence for one gender over another.

In general, it appears that for the few instances where evidence can be evinced from the small number of sentences in the corpus, that evidence leads to a correct classification. Presumably, if one had access to more Wambaya text, one could make more

correct classifications. This follows our intuition, and that of corpus-based computational linguistics over the past few decades.

However, the Web will not be the provider of that data. This may be because of the almost non-existent nature of the Wambaya speech community, but despite its gigantic size the value of the Web as a source of raw text for minority languages still remains to be demonstrated. Any Wambaya text that was returned by the search engine was swamped by other data, to the point where a Web frequency-based system was utterly hopeless — in contrast to other observations of such a system.

On the other hand, machine learning provided a promising approach in terms of having a high precision system that can actually make a non-trivial number of classifications, even from a small amount of data. The learner appears to be making best use of the data, without rigid constraints on co-occurrence with surface cues; this is possibly grounded in the distributional hypothesis. It is also possible that the model is overfitting to the regular structure of the linguistics-focussed corpus — this hypothesis is difficult to test, and, as little new text will be written for Wambaya, may remain unverified.

For the rich morphology in Wambaya, we contrasted identifying instances or cues based upon a simple set (the lemma, ABS, and the four proximal demonstratives), with a richly inflected set (INFL, and the full 62 demonstratives), and a resource-poor approach to morphology (STEM, where only the primarily-suffixing assumption is made). While performance between the systems was similar, it seems that having the full approach to morphology does indeed provide improvement in precision, at the cost of substantial development time. Simpler approaches, where assumptions about properties of the language can be quickly made and verified (using WALS Online[5] (Haspelmath et al., 2008), for example), seem like a reasonable trade-off.

All in all, the low-recall and moderate-to-high-precision results motivate a semi-automatic approach to lexicon extension: the model posits classifications, and the lexicographer examines these from high-confidence downward. As new entries are confirmed or corrected, the model can be re-run to suggest further classifications. This seems like a productive interaction for rapid lexicon extension.

While Wambaya probably presents the most extreme case of difficulty for the languages that have so far been analysed in deep lexical acquisition or lexical disambiguation, it also has its own idiosyncracies that make classification based on morphosyntax and contextual cues somewhat uninteresting. The very regular morphology may also be introducing biases in spite of its richness. As such, further analysis is required on other Australian languages — insofar as resources are available.

[5] http://wals.info

6 Conclusion

We have analysed a number of approaches to the prediction of grammatical gender of nominals in Wambaya, using Wambaya as a test case for a critically low-density language requiring documentation. While co-occurrence frequencies of gender-marking demonstratives give high precision in corpus frequency-based methods, recall is prohibitively low. Using the Web as a corpus does not allay this problem, as the Wambaya text available on the Web did not lead to useful frequency observations of the surface cues. Machine learning did appear to provide a more robust classification model, with some caveats for the nature of the data set; learning of morphological cues proved very effective, as these are distinctive in Wambaya. We envision these results as evidence for a semi-supervised approach to lexicon extension.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Timothy Baldwin and Francis Bond. 2003. Learning the countability of English nouns from corpus data. In *Proc. of the 41st Annual Meeting of the ACL*, pages 463–470, Sapporo, Japan.

Timothy Baldwin. 2005a. Bootstrapping deep lexical resources: Resources for courses. In *Proc. of the ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition*, pages 67–76, Ann Arbor, USA.

Timothy Baldwin. 2005b. General-purpose lexical acquisition: Procedures, questions and results. In *Proc. of the 6th Meeting of the Pacific Association for Computational Linguistics (PACLING 2005)*, pages 23–32, Tokyo, Japan. (Invited Paper).

Timothy Baldwin. 2007. Scalable deep linguistic processing: Mind the lexical gap. In *Proc. of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 3–12, Seoul, Korea.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proc. of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*, pages 26–33, Toulouse, France.

Emily M. Bender and Dan Flickinger. 2005. Rapid prototyping of scalable grammars: Towards modularity in extensions to a language-independent core. In *Proc. of the Second International Joint Conference on Natural Language Processing*, pages 203–208, Jeju Island, Korea.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proc. of the COLING 2002 Workshop on Grammar Engineering and Evaluation*, pages 8–14, Taipei, Taiwan.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, 8(1):23–72. 10.1007/s11168-010-9070-1.

Emily M. Bender. 2008. Evaluating a crosslinguistic grammar resource: A case study of Wambaya. In *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 977–985, Columbus, USA.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2009. Glen, Glenda or Glendale: Unsupervised and semi-supervised learning of English noun gender. In *Proc. of the Thirteenth Conference on Computational Natural Language Learning*, pages 120–128, Boulder, USA.

Grzegorz Chrupała, Georgiana Dina, and Josef can Genabith. 2008. Learning morphology with Morfette. In *Proc. of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Silviu Cucerzan and David Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proc. of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*, pages 40–47, Edmonton, Canada.

Scott Drellishak and Emily M. Bender. 2005. A coordination module for a crosslinguistic grammar resource. In *Proc. of the 12th International Conference on Head-Driven Phrase Structure Grammar*, pages 108–128, Stanford, USA.

Raymund G. Gordon, Jr, editor. 2005. *Ethnologue: Languages of the World, Fifteenth Edition*. SIL International.

Gregory Grefenstette. 1998. The World Wide Web as a resource for example-based machine translation tasks. In *Proc. of the ASLIB Conference on Translation and the Computer*, London, UK.

Jan Hajič and Barbora Hladká. 1997. Probabilistic and rule-based tagger of an inflective language - a comparison. In *Proc. of the Fifth Conference on Applied Natural Language Processing*, pages 111–118, Washington, USA.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors. 2008. The world atlas of linguistic structures online.

Frank Keller and Mirella Lapata. 2003. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29:459–484.

Mirella Lapata and Frank Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2:1–30.

William Lewis and Fei Xia. 2009. Parsing, projecting & prototypes: Repurposing linguistic data on the web. In *Proc. of the 12th Conference of the European Chapter of the ACL*, pages 41–44, Athens, Greece.

Marc Light. 1996. Morphological cues for lexical semantics. In *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 25–31, Santa Cruz, USA.

Jeremy Nicholson and Timothy Baldwin. 2009. Web and corpus methods for Malay count classifier prediction. In *Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 69–72, Boulder, USA.

Rachel Nordlinger. 1998. *A Grammar of Wambaya, Northern Territory (Australia)*. Pacific Linguistics, Canberra, Australia.