Automatic Discovery of Adposition Typology

Rishiraj Saha Roy
IIT Kharagpur
Kharagpur, India – 721302.
rishiraj@cse.iitkgp.ernet.in

Rahul Katare*
IIT Kharagpur
Kharagpur, India – 721302.
rah.ykg@gmail.com

Niloy Ganguly
IIT Kharagpur
Kharagpur, India – 721302.
niloy@cse.iitkgp.ernet.in

Monojit Choudhury
Microsoft Research India
Bangalore, India – 560001.
monojitc@microsoft.com

Abstract

Natural languages (NL) can be classified as prepositional or postpositional based on the order of the noun phrase and the adposition. Categorizing a language by its adposition typology helps in addressing several challenges in linguistics and natural language processing (NLP). Understanding the adposition typologies for less-studied languages by manual analysis of large text corpora can be quite expensive, yet automatic discovery of the same has received very little attention till date. This research presents a simple unsupervised technique to automatically predict the adposition typology for a language. Most of the function words of a language are adpositions, and we show that function words can be effectively separated from content words by leveraging differences in their distributional properties in a corpus. Using this principle, we show that languages can be classified as prepositional or postpositional based on the rank correlations derived from entropies of word co-occurrence distributions. Our claims are substantiated through experiments on 23 languages from ten diverse families, 19 of which are correctly classified by our technique.

1 Introduction

Adpositions form a subcategory of *function words* that combine with noun phrases to denote their semantic or grammatical relationships with verbs, and sometimes other noun phrases. NLs can be neatly divided into a few basic typologies based on the order of the noun phrase and its adposition. If the adposition is placed *before* the noun phrase, it is called a *preposition*. *Postpositions* and *inpositions*, on the other hand, are adpositions that are placed after and inside noun phrases respectively. If prepositions are predominantly used in the language, for example in English, Bulgarian and Russian, then the language is said to be *prepositional*. Similarly, Japanese, Hindi and Turkish are some examples of *postpositional languages*, which predominantly use postpositions. These two are the most commonly found adposition typologies across the globe. Out of 1185 languages analyzed on the World Atlas of Language Structures (WALS)[1] (Dryer and Haspelmath, 2011), there are 577 postpositional, 512 prepositional and only 8 inpositional languages. There are a few (30 and 58 respectively) languages which use no or both kinds of adpositions. The order of adpositions is strongly correlated with many other word order typologies. For instance, postpositional languages usually have Object-Verb ordering, whereas prepositional languages have Verb-Object ordering (Greenberg, 1963). Daumé and Campbell (2007) present a statistical model for automatically discovering such implications from a large typological database and discuss many other typological implications involving adpositions.

Motivation. Knowledge of the typological characteristics of languages is not only of interest to linguists, but also very useful in NLP for two main reasons. First, typological information, if appropriately exploited while designing computational methods, can lead to very promising results in tasks like coreference resolution and machine translation (Haghighi and Klein, 2007; Moore and Quirk, 2007). Second, as Bender and Langendoen (2010) have pointed out, in order to claim that a computational technique is truly language independent, one must show its usefulness for languages having diverse typological features. However, there is very little work on the automatic discovery of typological characteristics, primarily because it is assumed that such information is readily available. However, Hammarström et al. (2008)

* This work was done during the author's internship at Microsoft Research India.

[1] http://wals.info/

argue that documenting a language and its typological features is a time consuming process for the linguists and therefore, automatic methods for bootstrapping language description is a worthwhile effort towards language preservation. Lewis and Xia (2008) mine inter-linearized data from the Web and infer typological features for "low-density" languages, i.e. languages represented in scarce quantities on the Web. We argue that apart from documenting and understanding the typology of "low-density" languages, unsupervised discovery of adposition typology is also useful for analyzing undeciphered languages and scripts, such as the Indus valley script (Rao et al., 2009) and the Cypro-Minoan syllabary (Palaima, 1989), as well as newly emerging languages, such as the language of Web search queries (Guichard, 2002; Saha Roy et al., 2012) or the Nicaraguan sign language (Meir et al., 2010). While the former cases are interesting from historical and language change perspectives, the latter cases are useful for more practical reasons (for example, improvement in query understanding leading to better Web search, and development of interactive systems for deaf children).

Approach. In this work, we show that some simple word co-occurrence statistics, that can easily be computed from any medium-sized text corpus, can be used as reliable predictors of adposition typology of a language. These statistics have been arrived at based on two fundamental assumptions: (a) adpositions constitute a large fraction of function words; and (b) the strict ordering between the adposition and the noun phrase leads to differential co-occurrence characteristics on the left and right sides of the adposition. Therefore, if the function words of a language are automatically detected and the co-occurrence statistics on the left and right of those words are appropriately analyzed, then it should be possible to tell the prepositional languages apart from the postpositional ones. Specifically, we measure counts and entropies of left, right and total (either side) co-occurrence distributions for each word. We show that left co-occurrence statistics are better indicators of function words for prepositional languages, while right co-occurrence statistics perform better for postpositional languages. Interestingly, the performance of total co-occurrence statistics lie in between the two for both types of languages. Thus, the nature of the difference in performances of left (or right) and total co-occurrences is likely to be indicative of the adposition typology of the language. We formalize this intuition to devise our test for adposition typology. We demonstrate our technique on 23 languages from ten language families, of which 14 are prepositional and 9 are postpositional. Our technique is able to consistently predict the correct adposition typology for 19 of these languages. The remaining four languages are highly *inflectional* and *agglutinating* in nature, and hence not amenable to the present technique.

Organization. The rest of this paper is organized as follows. In Sec. 2, we present our method for function word detection using word co-occurrence statistics, along with results showing the effectiveness of such an approach. In Sec. 3, we propose our test for discovering the adposition typology of a language based on correlations inferred from different co-occurrence statistics. Sec. 4 discusses experiments conducted on diverse languages and inferences drawn from the observations. Finally, Sec. 5 summarizes our contribution and indicates possible directions for future work.

2 Function Word Detection

Our method for the prediction of the adposition typology of a language relies on the facts that most adpositions are function words, and the distributional properties of function words are very different from those of content words. We exploit this difference to first formulate a method for extracting the function words of a language from a corpus. We then proceed to use the same underlying principle to automatically discover the adposition typology for languages, where we do not assume that the true function word lists are available.

By function words, we refer to all the *closed-class* lexical items in a language, e.g., pronouns, determiners, prepositions, conjunctions, interjections and other particles (as opposed to open-class items, e.g., nouns, verbs, adjectives and adverbs). For the function word detection experiments, we shall look at four languages from different families: English, Italian, Hindi and Bangla. English is a *Germanic* language, Italian is a *Romanic* language, and Hindi and Bangla belong to the *Indo-Aryan* family. English and Italian are prepositional languages with *subject-verb-object* word order, while Hindi and Bangla are postpositional, relatively free word order with preference for *subject-object-verb*. Therefore, any func-

Language	Corpus source	S	N	V	Function word list source	F
English	Leipzig Corpora[a]	$1M$	$19.8M$	342157	Sequence Publishing[b]	229
Italian	*-do-*	$1M$	$20M$	434680	*-do-*	257
Hindi	*-do-*	$0.3M$	$5.5M$	127428	Manually constructed by linguists and augmented by extracting pronouns, determiners, prepositions, conjunctions and interjections from POS-tagged corpora available at LDC[c]	481
Bangla	Crawl of *Anandabazar Patrika*[d]	$0.05M$	$16.2M$	411878	*-do-*	510

[a] `http://corpora.informatik.uni-leipzig.de/download.html`
[b] `http://www.sequencepublishing.com/academic.html\#function-words`
[c] `http://www.ldc.upenn.edu` (Catalog Nos. LDC2010T24 and LDC2010T16 for Hindi and Bangla respectively)
[d] `http://www.anandabazar.com/`

Table 1: Details of NL corpora for function word detection experiments.

tion word characterization strategy that works across these languages is expected to work for a large variety of languages.

The details of the corpora used for these four languages are summarized in Table 1. M in the value columns denotes million. S, N, V and F denote the *numbers* of all sentences, all words, unique words (vocabulary size) and function words, respectively. We note that the Indian languages have almost twice as many function words as compared to the European ones. This is due to morphological richness and the existence of large numbers of modal and vector verbs.

Frequency is often used as an indicator for detecting function words, but the following factors affect its robustness. If the corpus size is not large, many function words will not occur a sufficient number of times. For example, even though `the` and `in` will be very frequent in most English corpora, `meanwhile` and `off` may not be so. As a result, if frequency is used as a function word detector with small datasets, we will have a problem of low recall. In our experiments, we measure corpus size, N, as the total number of words present. If our language corpus is restricted, or sampled only from specific domains, words specific to those domains will have high frequencies and will get detected as function words. For example, the word `government` will be much more frequent in political news corpora than `although`. The number of unique words in a corpus, or the vocabulary size, V, is a good indicator of its diversity. For restricted domain corpora, V grows much more slowly with N than in a general domain corpus.

We now introduce other properties of function words that may help in more robust detection. We observe the following interesting characteristics about the syntactic distributions of function and content words in NL, which can be summarized by the following two postulates.

Postulate I. Function words, in general, tend to co-occur with a larger number of distinct words than content words. What can occur to the immediate left or right of a content word is much more restricted than that in the case of function words. We hypothesize that even if a content word, e.g., *government*, might have high frequency owing to the nature of the domain, there will only be a relatively fewer number of words that can co-occur immediately after or before it. Therefore, the co-occurrence count may be a more robust indicator of function words.

Postulate II. The co-occurrence patterns of function words are less likely to show bias towards specific words than those for content words. For example, `and` will occur beside several other words like `school`, `elephant` and `pipe` with more or less equally distributed co-occurrence counts with all of these words. In contrast, the co-occurrence distribution of `school` will be skewed, with more bias towards `to, high` and `bus` than `over, through` and `coast`, with the list of words occurring beside `school` also being much smaller than that for `and`.

In order to test Postulate I, we measure the number of distinct words that occur to the immediate left, right and either side of each unique word in the sub-sampled corpora. We shall refer to these statistics as *left*, *right* and *total co-occurrence counts* (LCC, RCC and TCC) respectively. To test Postulate II, we compute the *entropy* of the co-occurrence distributions of the words occurring to the *left*, *right* and either

side (i.e., *total*) contexts of a word w:

$$\text{Entropy}(w) = -\sum_{t_i \in \text{context}(w)} p_{t_i|w} \log_2(p_{t_i|w}) \quad (1)$$

where, $\text{context}(w)$ is the set of all words co-occurring with w either in the left, the right or the total contexts, and $p(t_i|w)$ is the probability of observing word t_i in that specific context.

Context. In this paper, the left, right and total *contexts* of a word w respectively denote the immediately preceding (one) word, immediately succeeding (one) word and both the immediately preceding and the immediately succeeding words for w respectively, in sentences of the corpus. The definition of context (i.e., whether it includes the preceding or the succeeding one or two or three words) will change the absolute values of our results, but all the trends are expected to remain the same.

We shall refer to the co-occurrence entropies as *left*, *right* and *total Co-occurrence Entropies* (LCE, RCE and TCE respectively). Due to their pivotal role in syntactically connecting the different words or parts of a sentence to each other, we would expect LCC, RCC or TCC of function words to be higher than that of content words due to *Postulate I*; similarly, due to *Postulate II* we can expect the LCE, RCE or TCE to be higher for function words than for content words. If the LCE or LCC of a word w is high, it means that a large number of distinct words can *precede* w in the language (additionally, almost with equal probabilities for high LCE). Thus, predicting the *previous* word of w is difficult. Similarly, if RCE or RCC of w is high, it means that a large number of words can *follow* w in the language (additionally, almost with equal probabilities for high RCE). Thus, predicting the *next* word of w is difficult. A high TCE for a word implies that the word can be preceded and followed by a large number of words, making the prediction of either the next or the previous word (or both) for w difficult.

2.1 Experiments and Results

In our approach, the output is a ranked list of words sorted in descending order of the corresponding property. Here we adopt a popular metric, *Average Precision* (AP), used in Information Retrieval (IR) for the evaluation of ranked lists. More specifically, let $w_1, w_2, \ldots, w_n$ be a ranked list of words sorted according to some corpus statistic, say, frequency. Thus, if $i < j$, then frequency of w_i is greater than the frequency of w_j. *Precision at rank k*, denoted by P@k, is defined as

$$\text{P@}k = \frac{1}{k}\sum_{i=1}^{k} f(w_i) \quad (2)$$

where, $f(w_i)$ is one if w_i is a function word, and is zero otherwise. This function can be computed based on the gold standard lists of function words. Subsequently, *average precision at rank n*, denoted by AP@n, is defined as

$$\text{AP@}n = \frac{1}{n}\sum_{k=1}^{n} \text{P@}k \quad (3)$$

AP@n is a better metric than P@k because P@k is insensitive to the rank at which function words occur in the list. In our experiments, we compute AP@n averaged over $\mathcal{N}$ corpus sub-samples, which is given by $\frac{1}{\mathcal{N}}\sum_{r=1}^{\mathcal{N}}(\text{AP@}n)_r$ where $(\text{AP@}n)_r$ is the AP@n for the r^{th} sub-sample. We note that there are other metrics popularly used in IR, e.g. the Normalized Discounted Cumulative Gain (nDCG). However, these are more sensitive to the correctness of the top few items in the list and hence, are not suitable for us. Knowing that the number of function words in a popular NL is at least 200 (Table 1), we compute AP@200 with respect to the gold standard lists of function words for all our experiments.

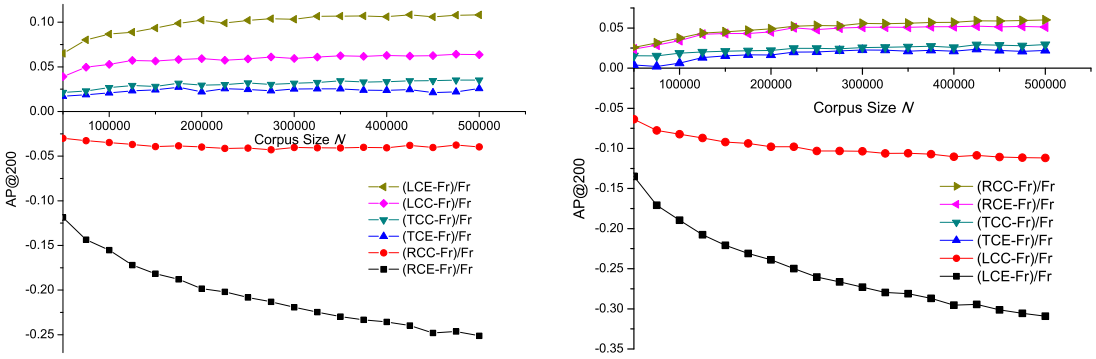
We now sort the list of all words in descending order of each of the seven indicators. We then compute AP@200 for these seven lists. To bring out the performance difference of each of the six co-occurrence features with respect to frequency, we plot (in Figs. 1 and 2) the following measure against N:

$$\text{Value plotted} = \frac{\text{Metric for indicator} - \text{Metric for Fr}}{\text{Metric for Fr}} \quad (4)$$

Language	Typology	Fr	LCC	LCE	TCC	TCE	RCC	RCE
English	Prepositional	0.663	**0.702**†	**0.729**†	**0.684**†	**0.679**†	0.637	0.527
Italian	Prepositional	0.611	**0.639**†	**0.645**†	**0.636**†	**0.620**	0.606	0.601
Hindi	Postpositional	0.682	0.614	0.510	**0.698**†	**0.694**†	**0.716**†	**0.713**†
Bangla	Postpositional	0.648	0.684†	0.691†	**0.730**†	**0.763**†	**0.741**†	**0.757**†

The four highest values in a row are marked in **boldface**. Statistically significant improvement over frequency is marked by †. The paired t-test was performed and the null hypothesis was rejected if p-value < 0.05.

Table 2: AP@200 for all indicators, averaged over 200 (N, V) pairs for each language.



(a) (b)

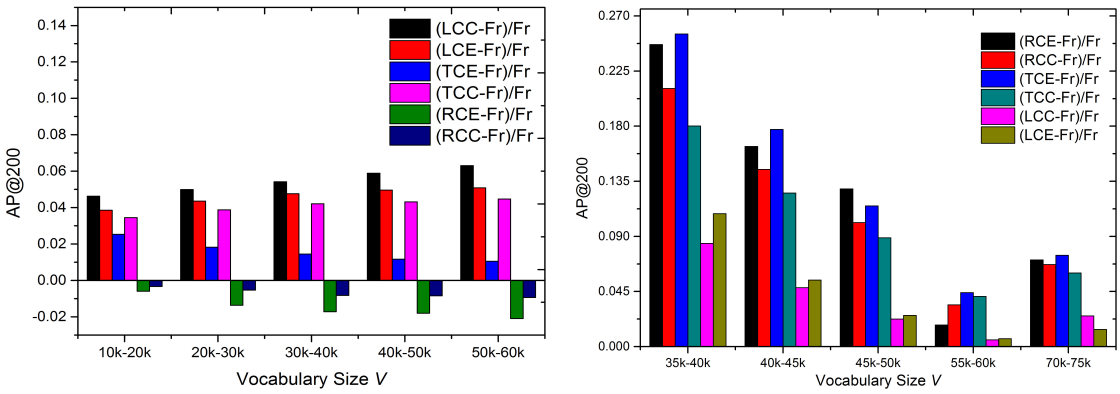
Figure 1: (Colour online) Performance of co-occurrence statistics for (a) English, and (b) Hindi, with respect to frequency for AP@200 with variation in *N*.

The x-axis can now be thought of as representing the performance of frequency. In Fig. 1, for a particular N, the data points were averaged over all (N, V) pairs (we had 20 (N, V) pairs for each N). For Fig. 2, V was binned into five zones, and for each zone, the AP was averaged over all corresponding (N, V) pairs. The observations (both N and V variation) for French and Italian were similar to that of English, while those for Hindi and Bangla were similar to each other. Table 2 reports AP values for all statistics for the four languages. From Table 2, we see that for all the languages, AP for some of the co-occurrence statistics are higher than AP obtained using frequency.

Regular improvements over frequency. From the plots and Table 2, it is evident that some of the co-occurrence statistics consistently beat frequency as indicators. In fact, as evident from Figs. 1 and 2, use of co-occurrence statistics results in systematic improvement over frequency with variations in N and V, and hence, are very robust indicators. Among the co-occurrence statistics, both entropies and counts are observed to have comparable performance.

3 Detection of Adposition Typology

From the results presented above, we observe that the best function word indicator depends upon language typology. Interestingly, while LCE and LCC are the best indicators of function words for the two prepositional languages of English and Italian, RCE and RCC perform better for Hindi and Bangla, the postpositional languages. This observation can be explained as follows. For a prepositional language, the function words, which are often the adpositions, precede the content word it is linked to. Therefore, the words following an adposition (or a function word) mark the beginnings of syntactic units such as noun phrases and are typically restricted to certain syntactic categories. However, the words that precede the adpositions have no or much weaker syntactic restrictions. Hence, the LCE and LCC are higher and consequently better and more robust indicators of function words for prepositional languages. For very similar reasons, the RCE and RCC are better indicators of function words for postpositional languages. Importantly, we observe that TCE and TCC seem to be reasonably good predictors of function words irrespective of the typology, with performances lying in between the poorest indicators (RCE and RCC



(a) (b)

Figure 2: (Colour online) Performance of co-occurrence statistics for (a) Italian, and (b) Bangla, with respect to frequency for AP@200 with variation in *V*.

for prepositional languages and LCE and LCC for postpositional languages) and the best indicators (LCE and LCC for prepositional languages and RCE and RCC for postpositional languages) for all the four languages. This makes them safe indicators to rely on when not much is known about the language syntax. In fact, the philosophy of this research is to be of assistance in these less-known cases. Thus, co-occurrence statistics have potential in predicting the adposition typology of a new language, which we leverage in this research.

We now describe our intuition and method behind our tests for automatically detecting the adposition typology of a language. In this context, we *do not know* the actual function words or adpositions of the language under consideration. Let us take the three lists of the top 200 words from a language corpus, sorted according to the statistics TCE, LCE and RCE. For a prepositional language, we can expect to see the highest number of function words towards the top of the list when sorted according to LCE, followed by the number of function words towards the top of the TCE list. The RCE list would be expected to be the poorest in this regard. Thus, we expect a higher overlap between the top 200 word lists for TCE and LCE, than for TCE and RCE. The reverse is expected to be true for postpositional languages. Similar arguments can be presented for LCC, RCC and TCC as well. We quantify this correlation between the lists using two different statistics – the *Pearson's correlation coefficient* (r) and *Spearman's Rank Correlation Coefficient* (ρ).

For computing Pearson's coefficients, we use the actual values of the distributional statistics, while for Spearman's rank coeffcients, we use the ranks of the words. Let $r(\mathrm{TL})$ and $\rho(\mathrm{TL})$ respectively denote the Pearson's and Spearman's Rank correlation coefficients of the lists sorted by TCE and LCE (or TCC and LCC), and similarly, let $r(\mathrm{TR})$ and $\rho(\mathrm{TR})$ denote the respective coefficients for the lists sorted by TCE and RCE (or TCC and RCC).

Postulate. For a prepositional language, the top-200 words by LCE will have a higher correlation with the top-200 words by TCE than the corresponding correlation of RCE with TCE. For a postpositional language, the top-200 words by RCE will have a higher correlation with the top-200 words by TCE. Formally, for *prepositional languages*, $r(\mathrm{TL}) > r(\mathrm{TR})$, and $\rho(\mathrm{TL}) > \rho(\mathrm{TR})$, while for *postpositional languages* $r(\mathrm{TL}) < r(\mathrm{TR})$ and $\rho(\mathrm{TL}) < \rho(\mathrm{TR})$.

4 Experimental Results and Observations

In this section, we first present our datasets, followed by detailed experiments on adposition typology detection and inferences drawn from the observations.

Language	Family	$\rho(\mathrm{TL})$	$\rho(\mathrm{TR})$	$\rho(\mathrm{Diff.})$	Predicted	True
Bulgarian	Slavic (Indo-European)	0.726	0.518	0.208	Pre-	Pre- (Scatton, 1984)
Danish	Germanic (Indo-European)	0.621	0.495	0.126	Pre-	Pre- (Allan et al., 1995)
Dutch	Germanic (Indo-European)	0.662	0.204	0.458	Pre-	Pre- (Shetter, 1958)
English	Germanic (Indo-European)	0.461	0.436	0.025	Pre-	Pre- (Selkirk, 1996)
German	Germanic (Indo-European)	0.563	0.517	0.046	Pre-	Pre- (Lederer, 1969)
Italian	Romance (Indo-European)	0.730	0.456	0.274	Pre-	Pre- (Sauer, 1891)
Macedonian	Slavic (Indo-European)	0.692	0.488	0.205	Pre-	Pre- (Friedman, 1993)
Norwegian	Germanic (Indo-European)	0.619	0.600	0.019	Pre-	Pre- (Olson, 1901)
Polish	Slavic (Indo-European)	0.798	0.554	0.243	Pre-	Pre- (Bielec, 1998)
Russian	Slavic (Indo-European)	0.743	0.652	0.091	Pre-	Pre- (Borras and Christian, 1959)
Slovenian	Slavic (Indo-European)	0.701	0.668	0.032	Pre-	Pre- (Priestly, 1993)
Swedish	Germanic (Indo-European)	0.663	0.525	0.138	Pre-	Pre- (Holmes and Hinchliffe, 1994)
Ukrainian	Slavic (Indo-European)	0.785	0.714	0.070	Pre-	Pre- (Stechishin, 1958)
Gujarati	Indic (Indo-European)	0.540	0.581	−0.041	Post-	Post- (Cardona, 1965)
Hindi	Indic (Indo-European)	0.529	0.731	−0.202	Post-	Post- (McGregor, 1977)
Japanese	Japanese (Japanese)	0.429	0.626	−0.197	Post-	Post- (Hinds, 1986)
Nepali	Indic (Indo-European)	0.495	0.719	−0.224	Post-	Post- (Bandhu, 1973)
Tamil	Southern Dravidian (Dravidian)	0.748	0.805	−0.057	Post-	Post- (Asher, 1982)
Turkish	Turkic (Altaic)	0.531	0.769	−0.238	Post-	Post- (Underhill, 1976)
Estonian	Finnic(Uralic)	0.790	0.733	0.057	Pre-	Post- (Tauli, 1983)
Finnish	Finnic (Uralic)	0.671	0.656	0.015	Pre-	Post- (Sulkala and Karjalainen, 1992)
Hungarian	Ugric (Uralic)	0.457	0.329	0.128	Pre-	Post- (Kenesei et al., 1998)
Lithuanian	Baltic (Indo-European)	0.715	0.724	−0.009	Post-	Pre- (Dambriunas et al., 1966)

Misclassified languages are marked in gray.

Table 3: Detecting adposition typology using Spearman's rank correlation coefficients on entropy lists.

4.1 Datasets

For all our typology detection experiments, we use datasets from the publicly available Leipzig Corpora[2]. We selected 23 languages from various families that are typologically diverse. A $(300,000)$-sentence corpora was used for all the languages so as to ensure similar-sized corpora for all the languages (many languages do not have a larger corpus). All languages examined have been listed in Table 3, along with their families and true adposition typologies (accompanied by appropriate references).

4.2 Experiments and Results

We extracted the top-200 words by TCE, LCE and RCE, and TCC, LCC and RCC from the $300k$-sentence corpora. We then computed $r(\mathrm{TL})$, $\rho(\mathrm{TL})$, $r(\mathrm{TR})$ and $\rho(\mathrm{TR})$, for both entropies and counts. As per our postulate, if $\rho(\mathrm{TL}) - \rho(\mathrm{TR})$ (= $\rho(\mathrm{Difference})$) is positive, the language is prepositional; if it is negative, the language is postpositional. The same can be expected for $r(\mathrm{Difference})$.

The performance of ρ as a predictor was found to be better than r. Results when the entropy lists are used are presented in Table 3. For only 4 out of 23 languages, the typology predictions are incorrect. We observe that three of these misclassified languages are from the Uralic family that are *synthetic* in nature characterized by extensive regular *agglutination* of modifiers to verbs, nouns, adjectives and numerals. The average number of characters in words of these languages were found to be in the relatively higher range of nine to eleven. Thus, function words, especially the adpositions, seldom occur as free words in these languages and hence our method cannot capture the distributional characteristics of the adpositions. It is worthwhile to note that the method can predict the correct typology for other languages that employ

[2] `http://corpora.informatik.uni-leipzig.de/download.html`

Corpus Size	Entropy lists (r)	Entropy lists (ρ)	Count lists (r)	Count lists (ρ)
$10k$	$17/23$	$\mathbf{21/23}$	$17/23$	$13/23$
$100k$	$17/23$	$\mathbf{19/23}$	$18/23$	$13/23$
$300k$	$16/23$	$\mathbf{19/23}$	$16/23$	$13/23$

The highest value in a row is marked in **boldface**.

Table 4: Correct predictions by strategy with varying factors.

agglutination to a lesser degree (Bulgarian, Dutch, German, Tamil and Turkish). Lithuanian, though not synthetic, is a highly inflectional language and therefore, instead of adpositions it makes extensive use of case-markers. With ρ(Difference) very close to zero, our prediction for Lithuanian is inconclusive.

A note on synthetic languages: For synthetic languages, the difference between the two rank correlation coefficients are close to zero, which provides us with a direct way to identify them. One could also employ unsupervised morphological analysis (Goldsmith, 2001) to automatically identify and segment affixes, which will provide deeper insight into the morpho-syntactic properties of the language. Nevertheless, affixes (like infixes in Arabic or case-marking suffixes in Bangla) are technically not considered as adpositions, and therefore, they do not really determine the adposition typology. Languages are divided into four classes according to their adposition usage: prepositional, postpositional, ambi-positional (use both types) and adposition-less (use none). Thus, as far as adposition typology is concerned, it suffices to identify whether a language is primarily adposition-less, which our technique is potentially capable of doing (we demonstrate it for four languages, but we believe more experimentation is needed to establish this claim). Note that a language may use case-marking affixes along with adpositions. In such cases our method is able to correctly determine the typology, as demonstrated for Bangla.

4.3 Experimental variations

We repeated the above experiments with lists of TCC, LCC and RCC instead of the co-occurrence entropies. The performance was found to be poorer than the entropy lists, with nine classification errors instead of the earlier four. Performance of these lists by co-occurrence counts was found to be poorer in other cases as well (Table 4). We systematically experimented with r instead of ρ. To test the performance of our method with even smaller corpora, we sub-sampled 3 and 30 corpora containing $100k$ and $10k$ sentences respectively from the $300k$ corpus. We computed the correlation between the original top 200 words obtained using TCE (or TCC) from the $300k$ corpus and the corresponding LCE and RCE (or LCC and RCC) lists obtained from the smaller corpora. For a given language, the mean of ρ(Difference) and r(Difference) were used to predict the typology (observed standard deviations were very low, of the order of 10^{-3}). The results of these experiments are summarized in Table 4. Out of 23 languages, 21 and 19 were correctly classified by ρ for corpora of $10k$ and $100k$ sentences. The corresponding number for r are 18 for both $10k$ and $100k$, and 17 for $300k$ corpora. Thus, the sensitivity of the method improves with slightly smaller corpora, provided that the TCE list, which is being used as a proxy for the gold standard function word list, is computed from a slightly larger corpus. Finally, we note that using Spearman's rank correlation coefficient with lists constructed by co-occurrence entropy consistently produces the best results.

5 Conclusions and Future Work

Knowing the adposition typology of a natural language can be useful in several NLP tasks, and can be especially useful in understanding new or undeciphered languages. In this research, we have taken one of the first steps towards automatic discovery of adposition typology. First, we have shown, through experiments on two prepositional and two postpositional languages, that function words can be effectively extracted from medium-sized corpora using word co-occurrence statistics, and such measures usually outperform simple frequency when used for the same task. Next, difference in behavior of various co-occurrence statistics for prepositional and postpositional languages has been exploited to devise a simple

strategy for predicting the adposition typology of a language. Simple differences of rank correlation coefficients among total, left and right word co-occurrence entropies have been shown to be potent signals towards automatic discovery of adposition and noun phrase typology in a language. Results show sufficient promise through an extensive evaluation over 23 languages.

We ventured into this study while solving a very practical and important problem: query understanding through analysis of the structure of Web search queries. While queries seem to have an emergent syntax, it is unclear whether they have function words, and if so what role they play in determining the query grammar. To this end, we conducted the current study. Thus, we envisage that this technique will be applicable for any such emergent linguistic system, such as pidgins, creoles, and computer mediated communications (CMCs) (Walther, 1996) like SMS and chats, where there is a large amount of text data available but the grammar is emerging or yet to be analyzed. Other examples are that of undeciphered languages, e.g., Indus valley language or script. In fact, our method can be applied to any system of symbols, be it linguistic or non-linguistic, such as musical note sequences.

As future work, it is important to improve our prediction accuracy further, while including more languages in the experimental setup. Combining clues from other sources to resolve uncertain cases and devising better ways of choosing corpus size and significance thresholds are some of the avenues in which effort may be channelized. Extending our approach to a morpheme-level analysis would also be beneficial in dealing with highly agglutinative and inflectional languages.

Acknowledgements

The first author was supported by Microsoft Corporation and Microsoft Research India under the Microsoft Research India PhD Fellowship Award. We would like to thank Amritayan Nayak, Walmart eCommerce (who was then a student of IIT Kharagpur working as an intern at Microsoft Research India) for contributing to some of the early experiments related to this study.

References

Robin Allan, Philip Holmes, and Tom Lundskær-Nielsen. 1995. *Danish: A Comprehensive Grammar*. Routledge, London.

R. E. Asher. 1982. *Tamil*, volume 7 of *Lingua Descriptive Studies*. North-Holland, Amsterdam.

Churamani Bandhu. 1973. Clause patterns in nepali. In Austin Hale, editor, *Clause, sentence, and discourse patterns in selected languages of Nepal 2*, volume 40.2 of *Summer Institute of Linguistics Publications in Linguistics and Related Fields*, pages 1–79. Summer Institute of Linguistics of the University of Oklahoma, Norman.

Emily M. Bender and D Terence Langendoen. 2010. Computational linguistics in support of linguistic theory. *Linguistic Issues in Language Technology*, 3(1).

Dana Bielec. 1998. *Polish: An Essential Grammar*. Routledge, London.

F. M. Borras and R. F. Christian. 1959. *Russian Syntax: Aspects of Modern Russian Syntax and Vocabulary*. Clarendon Press, Oxford.

George Cardona. 1965. *A Gujarati Reference Grammar*. The University of Pennsylvania Press, Philadelphia.

Leonardas Dambriunas, Antanas Klimas, and William R. Schmalstieg. 1966. *Introduction to Modern Lithuanian*. Franciscan Fathers Press, Brooklyn.

Hal Daumé and Lyle Campbell. 2007. A bayesian model for discovering typological implications. In *Annual Meeting of the Association for Computational Linguistics*, pages 65–72.

Matthew S. Dryer and Martin Haspelmath, editors. 2011. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich, 2011 edition.

Victor A. Friedman. 1993. Macedonian. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 249–305. Routledge, London / New York.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Comput. Linguist.*, 27(2):153–198, June.

Joseph H Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

Eric Guichard. 2002. *L'internet: Mesures des appropriations d'une technique intellectuelle*. These, Ecole des hautes études en sciences sociales, Oct.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Annual meeting-Association for Computational Linguistics*, pages 848–855.

Harald Hammarström, Christina Thornell, Malin Petzell, and Torbjörn Westerlund. 2008. Bootstrapping language description: The case of mpiemo (bantu a, central african republic). In *Proceedings of the Sixth international conference on Language Resources and Evaluation*, LREC '08.

John Hinds. 1986. *Japanese*, volume 4 of *Croom Helm Descriptive Grammars*. Croom Helm, Routledge, London.

Philip Holmes and Ian Hinchliffe. 1994. *Swedish: A Comprehensive Grammar*. Routledge, London.

Istvn Kenesei, Robert M. Vago, and Anna Fenyvesi. 1998. *Hungarian*. Descriptive Grammars. Routledge, London / New York.

Herbert Lederer. 1969. *Reference Grammar of the German Language*. Charles Scribner's Sons, New York. Based on Grammatik der Deutschen Sprache, by Doras Schulz and Heinz Griesbach.

W. Lewis and F. Xia. 2008. Automatically identifying computationally relevant typological features. In *Proc. of the Third International Joint Conference on Natural Language Processing (IJCNLP-2008)*.

R. S. McGregor. 1977. *Outline of Hindi Grammar*. Oxford University Press, Delhi. 2nd edition.

Irit Meir, Wendy Sandler, Carol Padden, and Mark Aronoff. 2010. Emerging sign languages. *Oxford handbook of deaf studies, language, and education*, 2:267–280.

R. C. Moore and C. Quirk. 2007. An iteratively-trained segmentation-free phrase translation model for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 112–119. Association for Computational Linguistics.

Julius E. Olson. 1901. *Norwegian Grammar and Reader*. Scott, Foresman and Co, Chicago.

Thomas G Palaima. 1989. Cypro-minoan scripts: Problems of historical context in problems in decipherment. *Bibliothèque des Cahiers de l'Institut de Linguistique de Louvain*, 49:121–187.

T. M. S. Priestly. 1993. Slovene. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 388–451. Routledge, London.

R.P.N. Rao, N. Yadav, M.N. Vahia, H. Joglekar, R. Adhikari, and I. Mahadevan. 2009. Entropic evidence for linguistic structure in the indus script. *Science*, 324(5931):1165–1165.

Rishiraj Saha Roy, Monojit Choudhury, and Kalika Bali. 2012. Are web search queries an evolving protolanguage? In *Proceedings of the 9th International Conference on the Evolution of Language*, Evolang 9, pages 304–311, Singapore. World Scientific Publishing Co.

Charles Marquard Sauer. 1891. *Italian Conversational Grammar*. Julius Gross, Heidelberg.

Ernest A. Scatton. 1984. *A Reference Grammar of Modern Bulgarian*. Slavica Publishers, Columbus, Ohio.

Elizabeth Selkirk. 1996. The Prosodic Structure of Function Words. In James L. Morgan and Katherine Demuth, editors, *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Routledge.

William Z. Shetter. 1958. *Introduction to Dutch*. Martinus Nijhoff, The Hague.

J. W. Stechishin. 1958. *Ukrainian Grammar*. Trident Press, Winnipeg.

Helena Sulkala and Merja Karjalainen. 1992. *Finnish*. Descriptive Grammar Series. Routledge, London.

Valter Tauli. 1983. *Standard Estonian Grammar. Volume 2: Syntax*, volume 14 of *Studia Uralica et Altaica Upsaliensia*. Almqvist and Wiksell, Uppsala.

Robert Underhill. 1976. *Turkish Grammar*. Massachusetts Institute of Technology (MIT) Press, Cambridge.

Joseph B. Walther. 1996. Computer-mediated communication. *Communication Research*, 23(1):3–43.