Syllable-Based Discovery of Word-Like Units: a model of unsupervised vocabulary acquisition

Abstract

This paper presents an unsupervised vocabulary acquisition model, along with its simulation results, that takes syllables as its basic units. The model is a bottom-up 'discovery' procedure in which strongly associated syllables are cumulatively joined, in contrast to the 'segmentation' strategy typical in the word acquisition work. Our model applies an information-theoretic association measure to N-grams with variable lengths, iteratively from unigram to N-gram until a threshold is reached, and as such, does not discriminate words and multi-word expressions. With the experiments on three languages with syllable-based orthography, Chinese, Japanese and Korean, it achieves up to over 80% precision, retrieving most frequent words and multi-word expressions. We also show the potential improvement that this addition brings to a language model.

1 Overview

In this paper we report a set of results from our experiments that simulate a syllable-based method for acquiring 'word-like units' from syllabified texts, and discuss its implications for humans' acquisition of unknown words as well as for engineering-oriented NLP. The method is entirely statistical and unsupervised, and thus can be considered to represent a domain-general manner of self-learning. It can also be of practical use for vocabulary extraction from unsegmented texts. The procedure can start from scratch, i.e. where no word is assumed to have been learnt, or from where some vocabulary has been acquired, and thus function as a word learning model at any learning stage, though our focus in this paper is on the former.

The proposed learning mechanism targets whatever sequences of syllables of *any length* that are strongly associated statistically, and thus does *not* discriminate between single words and multi-word expressions (hence the choice of the term 'word-like units'). The procedure returns a mixture of words and multi-word expressions, be it named entities or idiomatic expressions. The procedure thus learns linguistic units 'holistically', in presumably a similar way to what psycholinguists have observed in the acquisition of language in infants.

We use as the *association measure* (AM) an information-theoretic metric, *probability-weighted mutual dependency* (Thanopoulos et al., 2002), a variant of mutual information scores used in collocation detection. Starting from roughly syllabified texts, a procedure based on *variable length N-gram*(Kepler et al., 2012) is applied iteratively from unigram to N-gram, recording the association scores of the N-gram. We only keep those N-grams higher than a given threshold, and the iteration continues until no sequence is found above this threshold.

With three syllable-oriented languages, Chinese, Japanese and Korean, we show that approximately 80% of precision is achieved, as well as about the same rate of recall in the most frequent words and multi-word expressions.

2 Motivation and related work

It is a common observation that the human learners of a language, faced with continuous stream of sounds, do not distinguish between words and multi-word expressions (MWEs). The concept of word itself is far from uncontroversial too. Borderline cases are abundant, such as semantically monolithic 'hand bag' or 'none the less' (which could therefore be written without spaces), or semantically decomposable but phonologically monolithic 'aren't'. The wordhood is crucially dependent on meaning, of which the 'unity' is rather difficult to judge. Hence it appears more practical to be agnostic and defer the question of what really is a word, and search for 'word-like units' (WLUs), primarily phonological 'chunk'. The term dates back to (Zwicky, 1990) and the concept has been widely adopted by psycholinguists (Swingley, 2005; Endress and Mehler, 2009; Perruchet and Poulin-Charronat, 2012). Nevertheless such deliberate non-distinction is not so widespread amongst computationalists, who still generally work on space-segmented texts. In our view, however, a radical approach is required that once decomposes sentences to phonological level, where the contentious semantic issue simply is absent.

As for the computational (and statistical) work on the phonological level, unsupervised word acquisition tends to be tackled as the matter of *segmentation*, where the task is dividing a phoneme string into word units ('bracketing' model á la Perruchet et al. *ibid*). Generally, this line of work relies on probabilistic path selections, where some constraining top-down principle is also applied. Amongst such streams of work are (Goldsmith, 2006) and (Brent, 1999), based on the principle of Minimum Description Length, and more recently, (Goldwater et al., 2009), based on a Bayesian sampling method. However, as (Swingley, 2005) points out, whether this is the best approach as a vocabulary acquisition model for the very first stage of learning is suspect, as it assumes the learner to capture all the distinct phonemes presented to him/her in their totality, which appears rather implausible.

Our work is oriented, in contrast, in a bottom-up procedure that has been typically resorted to in research on MWEs. We don't segment but *join* successively the adjacent atomic units, in our case syllables, using association measures. AMs are intended to show the degree of association strengths between adjacent units. A wide variety of metrics have hitherto been proposed (see (Pecina, 2009) for the inventory), and there is nothing that should prevent us from applying them to sublexical levels. Some computational linguists have indeed started to apply them to character-level processing, for example in the context of the detection for Chinese of 'unknown words' on the basis of an existent lexicon (e.g. neologisms and named entities) (Ling-Xing et al., 2010; Liping et al., 2015). Our work can be taken as an attempt to generalise the application of AMs to the learning of all words from scratch, without a lexicon. Such a 'joining' approach as a general acquisition model has been pursued by psycholinguists (Swingley, 2005), but not much computational simulation has been attempted except on the basis of transitional probabilities (Perruchett and Vinter, 1998).

The reasons to opt for the level of syllable rather than that of phonemes are twofold, practical on one hand and plausibility-oriented on the other. The practical reasons are firstly that we have already languages thus segmented available, and that for these languages, where no word boundary is apparent on the surface even on the written form, word segmentation is a big issue in NLP. Moreover, starting from syllable level is obviously easier than from the phoneme level to reach word-level segmentation. This is not just expedient but can be cognitively motivated as well, as it is now broadly in agreement amongst the psycholinguists (Bertoncini and Mehler, 1981) that syllables, not phonemes, are the building blocks of words on the cognitive level.

Another major difference to the 'segmentation'-based work, in terms of the end product that the procedure produces, is that our procedure will not, and does not even attempt to, cover the whole string of a sentence. In contrast, we only produce the parts in a string that are more salient, leaving the rest untouched. It is close to the MWE work in spirit again. Our work can be taken to belong to the stream of work on multi-segment unit discovery, just as MWE extraction work, albeit on a different, lower-level granularity.

3 Datasets and their syllabification

We use social media as the data source for Chinese, Japanese and Korean specifically Weibo for Chinese, Twitter for the other two. Since our proposed units are syllables, the data first need to be 'syllabified', that is, segmented so that each segment corresponds one-to-one to a syllable.[1] This can be done most straightforwardly for Korean, as its orthography is entirely syllable-oriented. As a matter of fact almost nothing needs to be done, since every character represents a syllable.

The situation is a little more complicated for Japanese and Chinese because their standard orthography is (at least partially) ideographic and does not meet the condition of a one-to-one correspondence to syllables, but we can take advantage of the syllable-based orthography each language is equipped with. For Japanese, we can convert the ideographs (*kanji*s) into *kana*s, a syllable-based phonetic alphabet, while for Chinese, *pinyin* (with tones) is at hand, similarly a syllable-based orthography. We used ten million syllable tokens after cleaning in each language.

[1] We do not do syllabification in a linguistically strict sense. That is, we rely largely on orthographies and gloss over real phonetic deviations from the written surface form. We believe this is permissible given our global goal identifying words from syllabic strings.

4 Association metric and generalisation to N-grams

There are some well-known association measures, to quantify the cohesion between linguistic units. Amongst these one of the most used in the MWE context is an information-theoretic measure, *pointwise mutual information* (PMI), which is the co-ocurrence (joint) probability of two units divided by the 'expected' probability, the product of the two units' marginal probabilities ($log_2\frac{P(u_1,u_2)}{P(u_1)P(u_2)}$).

Since the denominator becomes smaller geometrically as the frequencies of the constituents become smaller, PMI very much favours sequences with infrequent constituents, with the joint probability, the numerator, being equal. An obvious extreme case is one in which each constituent occurs only once globally. Typical examples would be esoteric foreign names. In an English corpus for example, *Chiang Kai Shek* can be such an example on syllable level, assuming only one person mentions in it the former Taiwanese president once. We are precisely *not* targeting these type of sequences but the ones that 'stand out', i.e. are perceived by the humans to be *salient* as a unit. Frequency as a sequence is important factor for being salient. Thus the probability-weighted version of PMI or PMD (*probability-weighted mutual dependency*) will be used in this study (Thanopoulos et al., 2002), which is computed as follows:

$$PMD = log_2\frac{P(u_1,u_2)^2}{P(u_1)P(u_2)} + log_2 P(u_1,u_2)$$

.

where u_1 and u_2 are adjacent units of any level.

Another crucial requirement of this study, given that a WLU could stretch over a large number of syllables, is o extend the application of AMs beyond bigrams and generally to N-grams. Therefore we employ the *variable-length* N-grams (Kneser, 1996; Kepler et al., 2012), where we start with bigrams and extend them to N+1-grams by taking N-grams as the first item of a 'bigram' —or *generalised bigram* or GBGs as we henceforth call it— to compute the mutual information of N-grams in general for any N (Pitler et al., 2010; Liping et al., 2015). The schematic procedure is illustrated in the pseudocode below.

```
Data: $GBGsWithCount = \{((Units1, Unit2), Occ) \mid$ Units1 precede Unit2$\}$
where Units1 is an n-tuple of base units, Unit2 a single base unit and Occ the occurrences of (Units1,Unit2)
initialisation;
GBGsWithCount$\leftarrow$ BigramWithCount
GBGsWithAM$\leftarrow$ []
while $GBGs \neq \emptyset$ do
    for $GBG \in GBGs$ do
        calculate FrMD of GBG;
        GBGWithAM.append((GBG,AM));
    end
    GBGsWithAM$\leftarrow$ sort GBGsWithAM on AM;
    GBGsWithAM$\leftarrow$ apply filters on GBGWithAM;
end
```

Algorithm 1: Generalising bigrams to N-grams with filtering

An obvious practical problem is the complexity of such cumulative iterations with GBGs. The possible combinations in simple bigrams is already exponential in general case (the 2-permutation of vocabulary size or the type count of the corpus, whichever is greater). The vocabulary size of syllables is admittedly not as large as that of words, but it is rather the following fact that makes the whole procedure genuinely intractable: that after each iteration the resulting, bloated, N-grams are used in the next iteration. Furthermore, while word N-grams are heavily constrained to the extent that only a fraction of the possible combinations actually obtain (the very point of N-gram based modelling in the first place), syllable N-grams are much less constrained, which makes the complexity issue severer. We therefore need pruning between iterations to stay tractable, and this also accords with the spirit of cognitive modelling, as well as practical applications, where the amount of space (or 'memory') or time (either in humans or in a computer) is limited.

The end result of this procedure is a set of sequences of syllables with scores, which can therefore be ranked. They are candidates for WLUs, and the higher the score, the more plausible the learner judges as a unit.

We employ three types of pruning, two of which are applied between iterations. One is a simple frequency cutoff point, since, as stated earlier, we are not interested in infrequent items. The other is a threshold for the metric used, PMD, where its optimal point has to be experimentally established. Since all the sequences (GBGs) after an iteration are attached with metric values, those items that do not satisfy the conditions can be straightforwardly removed. The filtered items (including the ones filtered by the third method, to be discussed below) will then be left out in the following iteration to save time and not stored individually in the memory, though they will count as the 'tail' of the distribution assumed to be uniform, to be included in the global counts.

Notice that these prunings are in fact the source for the incompleteness we mentioned earlier. Without them the procedure would continue practically perpetually until all the sequences are exhausted, but our learner selectively search for good-looking sequences by giving up on likely hopeless ones. This is a practical necessity given the complexity of unrestricted search too. Such necessity can be said to reveal sources for superfluity in our scheme, unlike the 'bracketing' method. One of the most important redandancies concerns super/substring, which is discussed in a separate subsection below.

4.1 Super/sub-string reduction

One of the problems with our bottom-up method is how to deal with super/sub-string sets. There is so far nothing that stops a relatively long N-gram from 'moving on' to an N+1-gram that properly contains it, say ABCDE to ABCDEF, further to ABCDEFG and so on, even though it is likely that some or all of these heavily overlapped sequences are not separate WLUs but one.

We introduce therefore two manners of super/sub-string reduction. One is to stop the chain of 'upward' accumulation, and the other is a 'downward', post-hoc pruning. The first is proposed by (Liping et al., 2015) who, in a nutshell, propose to suppress the transition of an N-gram to its superstring N+1grams, if the drop in the score is much higher than the expected drop. The rationale for this is that if the drop is sharp enough, that position is likely to mark a definitive word boundary, beyond which there is not much hope to find more expressions. So there is a buffer b, and we continue the search only if $PMD_n < PMD_{n+1} + b$. This buffer[2] is necessary not to overprune and leave room for substring branching into different eventual superstring expressions.

While this chain halting is very effective in stemming a branch of search, there would still remain another type of unwanted sequences precisely due to our caution not to overprune: *transitional* stages to another WLU, particularly an MWE. When we keep going in a search of a further superstring WLU, it can come rather far away, leaving a prolonged intermediate sequence inbetween. Take an English case for example, *I-beg-your-par-don*, the procedure is not likely to stop at the boundary between *beg* and *your* and runs until the end of *don*, maintaining a relatively high score inbetween, e.g. between *your* and *par*. Thus the sequence *I-beg-your-par* will be kept, as an intermediate step to reach our desired MWE, but it just keeps the space that spawns unnecessary branches. This is essentially the same observation made by (Lu et al., 2004), and we do use their method of substring reduction. Whenever the procedure encounters overlapped character sequences that are similar in probabilities dynamically, we prune these probability-wise 'stable' parts considering them as likely transitional stages.

5 Evaluation

We evaluate the goodness of the produced WLU candidates in two ways: firstly against our 'gold standards', and secondly for their potential efficacy as an addition to a language model. Preparing a gold standard may not seem straightforward in our study due to two factors. First, our list will include MWEs, and it is widely considered elusive to prepare a definitive list for MWEs because of the absence of established explicit criteria. Second, since our procedure does not produce a complete result, if the standard is 'too complete', the recall will suffer from a massive shortfall.

We take a somewhat simplistic, pragmatic stance here. To deal with the MWE issue, we simply run the same procedure on the word-segmented text and add the results to our golden standard. Furthur, we bypass the incompletion issue by using only the most frequent words / MWEs that match the number of produced candidates. The number of produced candidates depends on the thresholds, and with our chosen set of thresholds, our proposed procedure typically yields, after 10 million tokens, 1,300 - 1,500 candidates depeding on the language. Therefore we take roughly the same amount of most frequent items each from two sources, words and of MWEs, make them serve as the standard. It's a disjunctive reference: if an item chosen is a word, we refer to the word list, if it is an MWE, we refer to the MWE list.

We also attempt to gauge the potential added value of newly found WLUs, and conduct for this purpose a prediction-based evaluation on the language models trained on the same data with and without the

[2] The buffer that is considered 'sharp enough' is the mean of the two immediately lower-order MIs, that is, given an N, those between $s_1, ..., s_{n-2}$ and s_{n-1} and between s_1 and $s_2, ..., s_{n-1}$.

	Chinese	Japanese	Korean
Precision	75.64	80.64	81.43
Recall	75.25	85.44	77.31

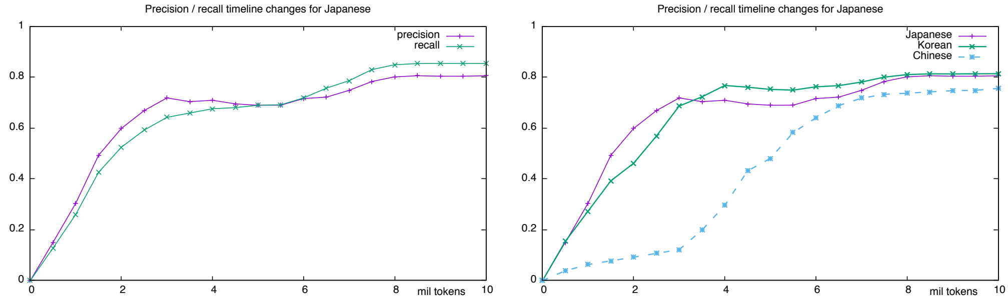

Figure 1: Precision and recall, with their timeline change

WLU additions (Wester, 2003). The baseline is a purely syllable-based trigram model, against which our augmented model, with WLU candidates in its vocabulary, is compared. As a further comparison we also test an 'upper bound' model, which has been trained with the customary word trigrams with a full lexicon augmented with our MWEs. Note that these models will be hugely different from each other in terms of vocabulary, so the conventional perplexity figures would not be comparable. This is why we use the prediction rate, on the level of syllables at that, to include 'partial' results in the equation.

5.1 Results

Results are shown in Figures 1, 2 and 3, the first two corresponding to the comparison with our gold standard and to the prediction rates, as discussed above. The third figure shows samples of top-ranked WLUs, which are discussed later in the analysis subsection.

The table in Figure 5.1 shows the precision and recall of the models trained with full data, that is approx. 10 million syllables.

The graph below on the left hand side shows an accumulation effect —we henceforth call it timeline change, though it's not strictly to do with time— with both precision and recall for Japanese. As can be seen the precision tends to hit the peak earlier where growth stops, whereas recall seems slower but more constant.

We only show the results on precision for the other two languages, since they follow the same trend as for the contrast between precision and recall, on the right hand graph. Nevertheless, there is a good deal of cross-language variations.

Figure 3 shows the character prediction rates of our main language model in question built with a vocabulary of syllables plus the obtained WLU candidates ('syl-WLU'), pitched against the baseline model wth a syllable-only vocabulary ('syl-only') and the upper-bound model (the word model). The bracketed numbers are the differences from the baseline model and the upper-bound model.

5.2 Analysis and discussion

The model performs reasonably well successfully picking up most of the frequent words and MWEs, although with a possible exception of Chinese, which lags behind, if it goes well above the chance level. It also is a reasonably quick learner, particularly given the light computational cost it requires, reaching the plateau after around 3-4 million tokens, that is, given the average syllable length at about 30-40, comes down to about ten thousand 'teweets'. It nevertheless is not a perfect learner, seeming to reach its limit around 80%, although the recall shows sign of a possible further growth.

The language model results also point to a significant 'added value' in terms of processing of language, if there stil is a distance from the word+WLU model. Again there is a cross-linguistic variation, but it is

Chinese			Japanese			Korean		
syl-only	**syl-WLU**	word	syl-only	**syl-WLU**	word	syl-only	**syl-WLU**	word
21.41	23.82	33.11	23.34	26.81	34.40	22.14	25.96	30.71
(+1.42,-9.29)			(+3.47,-7.59)			(+3.82,-4.75)		

Figure 2: Prediction-rate changes for different granularities

Chinese	*shui4 bu4 zhao2, bu4 zhi1 dao4, shi2 jun1 zhi1 lu4, chun1 wan3, hen3 bang4 de5 *de5 ren2, *wo3 de5, bu4 zhao2, *tang3 zai4 chuang2 shang4, zhe4 ji3 tian1, ao4 ba1 ma3, hai2 zi5, liang3 ge4, *zhu4 da4 xin1 kuai4 *ai4 wo3 de5 ren2 he2 wo3 ai4 de5 ren2, shen2 me5,keng1 die1, *qiu1 ling3 dao4 feng4, yi2 ren2, *rang4 wo3, ming2 tian1, jia1 you2 jia1 you2
Japanese	ッテ, マス, デス, ッタ, シタ, カラ, シテ, キョオ, ナイ, * ンダ, コオ, ワ, チュウ, ユウ, ヨロシクオネガイイタシマス, * ジョオ, サン, カイ, トリアエズ, ダイジョオブ, タイ, * カワイ, ショオ, セン, ケド, アリ, カクサンキボオ, * オモ
Korean	니다, 습니, 으로, 월호, 에서, 오늘, 사람, * 합니, 입니, 생각, 함께, * 하는, * 지않, 세요, 는데, * 유병언, 하고, 국민, 네요, 마지막, 너무, 우리, 는것, 처럼, 가,* 괜찮, 있는, 부터, 프로그램, 화이팅

??

Figure 3: Top-ranked WLU candidates

apparent that the additions of WLU candidates, despite included errors and possible added confusability, brings improvements over the syllable-based model. Apart from the use as an additional component, the technique may still not in a mature state, given the presense of superior supervised methods, as well as additional necessity of syllabification which may be deemed gratuitous. However, it is important to seek a universal method that could be applied to under-resourced languages too, in a manner that can connect the sound processing.

Another important observation is that the performance varies across the languages. In our three target languages, somewhat surprisingly, Chinese, which has an isolating morphology, lags behind inflectional Japanese and Korean. This could be due to the fact that the word lengths are shorter in the former two than the latter two, making the distribution dense. Furthermore, the learning curves are also different from language to language, though they form roughly an asymptotic pattern. Particularly interesting is the onset of growth can be very different, which warrants a further qualitative investigation.

In Figure 3, the top thirty candidates are shown that have been produced by our procedure for the feel of what kind of sequences you can get. The ones with '*' are the 'errors', i.e. ones not found in the golden standard. As the speakers of these language would see, the model picks up 'phonological chunks' rather than syntactic words, bundling together frequently co-occurring syllables. Particularly prominent are common greetings, which are very often used in our particular medium in a fixed form.

The types of errors frequently observed include, for Japanese and Korean, the case of not capturing inflectional endings. Amongst such examples are, for Japanese, カワイ *kawai* and オモ *omo* are part of adjective *kawaii* and verb *omou* respectively, or for Korean, 아니 *ani* and 지않 *j*ianh both lack their ending. These however are predictable results, since they are exactly the phonologically stable part of the words. *kawaii* inflects to *kawai-ku*, *kawai-i* etc, and *omou* to *omo-wa*, *omo-i* etc,. whereas the Korean examples are followed by a variety of endings, *da*, *kka*, *go* etc. They can in fact be considered not so much errors as a segmentation issue, because a modified analysis may lead to the corresponding segmentations. Nevertheless, children simply make no such mistakes, and an additional mechanism would be in order as a cognitive model.

There are a large number of two-syllable candidates (25%-30%), in contrast with a very few instances of one-syllable extractions (under 4%), creating 'undersegmentation' issue for the former and 'oversegmentation' for the latter. The proportions of syllable counts certainly vary from language to language, but both figures are consistently larger (disyllabic) and smaller (monosyllabic) in the candidate lists than their golden standard counterparts in all the languages. For the one-syllable entries, the largest difference is observed for Chinese, where about a fifth of the word golden standard are one-syllable words, whereas only 2.5% of the found candidates are monosyllabic. On the other hand the largest difference for two-syllable items is found for Korean, where about 28% of the items in the candidates fall into this category against 9% in the golden standard.

Single syllables are essentially unigrams, and their PMI, if generalised, becomes nil. Our metric, PMD, though it normally owes much of its value to PMI, then boils down to just the log probability $log_2 P(\sigma)$. Unless extremely frequent, the log probability will not rise above two-syllable PMD, though it can happen (as it does with the Japanese particle *wa* and Korean particle *ga*). On the other hand two-syllable items benefit both from the PMI and log-probability components, if they are coherent enough (relatively more frequent than other combinations) and are themselves frequent. It may further benefit from the 'upward' superstring chain pruning mechanism. *wo de* ('of me') in Chinese, for example, is a somewhat inevitable inclusion of our procedure in this sense.

6 Conclusion and future tasks

We have shown a set of simulations for the bottom-up and cumulative method of vocabulary discovery based on syllables, which have revealed the potential of the method as well as its possible weaknesses. Syllable-based bottom-up strategy does go a long way towards the acquisition of vocabulary, but does not seem like a complete picture. This imperfection is not unexpected, and hence, warrants further studies.

There are two directions that this research can take. One is broaden the range of languages, since our choice is admittedly expedient, and picking the languages with a stable syllable structure, as we do, may give too favourable light to the strategy in question. It is therefore essential to extend the simulation to other languages, particularly ones in which syllabification itself is a more difficult task because boundary ambiguity is very common, such as English and French.

The other direction is to combine it with a top-down strategy for word acquisition, which the humans undoubtedly employ. This might just take the form of the 'segmentation' strategy, but other forms which are complementary to the present approach would also be possible, such as (Chen et al., 2011), simpler and computationally lightweight method just as our method is.

References

J Bertoncini and J Mehler. 1981. Syllables as units in infant speech perception. *Infant Behaviour and Development*, 4:247–60.

Michael R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105.

Songjian Chen, Yabo Xu, and Huiyou Chang. 2011. A simple and effective unsupervised word segmentation approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence.*

A D Endress and J Mehler. 2009. The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language.*

John Goldsmith. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering.*

Sharon Goldwater, Thomas Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112:21–54.

Fabio Kepler, Sergio Mergen, and Cleo Billa. 2012. Simple variable length N-grams for probabilistic automata learning. In *JMLR Workshop and Conference Proceedings 21*. Journal of Machine Learning Research.

Reinhard Kneser. 1996. Statistical language modeling using a variable context length. In *Proceedings of the 4th International Conference on Spoken Language Processing.*

Ling-Xing, Tang Shlomo Geva, and Andrew Trotman. 2010. A boundary-oriented chinese segmentation method using ngram mutual information.

Du Liping, Li Xiaoge, Yu Gen, Liu Chunli, and Liu Rui. 2015. New word detection based on an improved pmi algorithm for enhancing segmentation system. *Acta Acientiarum Naturalium Universitatis Pekinensis.*

X Lu, L Zhang, and J Hu. 2004. Statistical substring reduction in linear time. In *Proceedings of IJC-NLP.*

Pavel Pecina. 2009. *Lexical Association Measures: Collocation Extraction.* Institute of Formal and Applied Linguistics, Charles University of Prague.

Pierre Perruchet and Benedicte Poulin-Charronat. 2012. Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, (66):80–818.

P Perruchett and A Vinter. 1998. PARSER: A model for word segmentation. *Journal of Memory and Language*, 39(2):246–263.

Emily Pitler, Shane Bergsma, Dekang Lin, and Kenneth Church. 2010. Using web-scale N-grams to improve base NP parsing performance. In *Proceedings of the 23rd International Conference on Computational Linguistics.*

Daniel Swingley. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50:86–132.

Aristomenis Thanopoulos, Nikos Fakotakis, and George Kokkinakis. 2002. Comparative evaluation of collocation extraction metrics. In *Proceedings of the 3rd Language Resources Evaluation Conference*, pages 620–625.

Malin Wester. 2003. User evaluation of a word prediction system. Master's thesis, Uppsala University.

Arnold Zwicky. 1990. Syntactic words, morphological words, simple and compsite. In *Morphology Yearbook*.