Building a Lexicon of Formulaic Language for Language Learners

Julian Brooke[*†] **Adam Hammond**[‡] **David Jacob**[†]
Vivian Tsang[†] **Graeme Hirst**[*] **Fraser Shein**[*†]

[*]Department of Computer Science
University of Toronto
jbrooke@cs.toronto.edu
gh@cs.toronto.edu

[†]Quillsoft Ltd.
djacob@quillsoft.ca
vtsang@quillsoft.ca
fshein@quillsoft.ca

[‡]School of English and Theatre Studies
University of Guelph
adam.hammond@uguelph.ca

Abstract

Though the multiword lexicon has long been of interest in computational linguistics, most relevant work is targeted at only a small portion of it. Our work is motivated by the needs of learners for more comprehensive resources reflecting formulaic language that goes beyond what is likely to be codified in a dictionary. Working from an initial sequential segmentation approach, we present two enhancements: the use of a new measure to promote the identification of lexicalized sequences, and an expansion to include sequences with gaps. We evaluate using a novel method that allows us to calculate an estimate of recall without a reference lexicon, showing that good performance in the second enhancement depends crucially on the first, and that our lexicon conforms much more with human judgment of formulaic language than alternatives.

1 Introduction

A significant portion of a speaker's lexical knowledge consists not of atomic lexical entries, i.e. words, but rather sequences built from their combination; in fact, the working multiword lexicon of the average native speaker is almost certainly much larger than the single-word lexicon (Church, 2011). Language learners, due to lack of exposure to the new language and interference from their native language, often fail to use these larger sequences proficiently, a fact which has been demonstrated via corpus analysis using high frequency *n*-grams (Chen and Baker, 2010; Granger and Bestgen, 2014). Although high frequency *n*-grams, known in corpus linguistics as lexical bundles, are useful for certain kinds of analysis, they are inappropriate for a fully-featured multiword learning system, which would ideally involve an electronic lexicon corresponding roughly to the internal lexicon of native speakers. In this work, we adopt the creation of such a lexicon as our goal.

Though much work has been done and many resources created which focus on specific aspects of the multiword vocabulary, most notably in fields such as multiword expressions (MWEs) (Baldwin and Kim, 2010) and keyphrase/term extraction (Newman et al., 2012), our pedagogical perspective leads us towards a somewhat broader theoretical foundation, the *formulaic sequence* theory of Wray (2002; 2008). We are interested in any multiword sequence that could plausibly be lexicalized, not simply those that are noncompositional (idiomatic) or that are otherwise useful for information retrieval applications. With our goal of helping advanced learners produce more fluent language, we are more interested in sequences that underpin the structure of sentences and not just terms that reflect its topic. As much as possible, we do not want to limit the syntactic composition, size, or frequency of our lexical items, and we want methods that allow us to build distinct, high-coverage lexicons for varying genres.

Working on top of an existing pipeline for unsupervised multiword unit segmentation (Brooke et al., 2014), the current work presents two key improvements on that initial model that allow us to build high-coverage lexicons of formulaic language. With respect to improving the quality of the sequences, we present a new measure for distinguishing true (lexicalized) affinity from background syntactic effects, the *lexical predictability ratio*, and integrate it into the model to improve the quality of the out-

put lexicon. The second major advance expands the coverage of the lexicon beyond directly contiguous sequences, allowing for sequences with gaps. Note that these are not independent, since the class imbalance between possible and actual gap phrases means that the second depends on the first.

Our main evaluation is novel for this space: rather than comparing with (necessarily) incomplete reference lexicons, we view our task as a *n*-gram (or gapped *n*-gram) filtering task, sampling *n*-grams to annotate from our full (frequency-filtered) set, which allows us to calculate a reliable precision, recall, and F-score. We also test the relevance of our lexicon to contextual recognition of multiword expressions, using a recently released dataset. In both cases, our method outperforms a variety of alternatives, including the original segmentation approach that was our starting point; like that original approach, our lexicon creation method is highly scalable and deterministic, and has only one key parameter (minimum frequency in the corpus).

2 Related Work

There is a long-standing area of research in computational linguistics focusing on lexical association measures, often, though not exclusively, for the creation of multiword lexicons (Church and Hanks, 1990; Schone and Jurafsky, 2001; Evert, 2004; Pecina, 2010): for two-word sequences there are, in fact, far too many to list in this context, though most of the research has centered upon popular options such as the *t*-test, log-likelihood, and pointwise mutual information (PMI). When these methods are used to build a lexicon, particular syntactic patterns and thresholds for the metrics are typically chosen. Critics note that many of the statistical metrics do not generalize at all beyond two words, but PMI (Church and Hanks, 1990), the log ratio of the joint probability to the product of the marginal probabilities, is a prominent exception. Other measures specifically designed to address sequences of larger than two words include the *c*-value (Frantzi et al., 2000), a metric designed for term extraction which weights term frequency by the log length of the *n*-gram while penalizing *n*-grams that appear in frequent larger ones, and mutual expectation (Dias et al., 1999), which produces a normalized statistic that reflects how much a candidate phrase resists the omission of any particular word.

Overlapping with this area is the research on multiword expressions (Baldwin and Kim, 2010), which is generally (though not exclusively) understood to refer to idiomatic, non-compositional multiword units; even so restricted, there is a huge variety of distinct types, and research in the area has tended to be rather focused, looking at, for instance, just verb/noun combinations (Fazly et al., 2009). The recent work of Schneider et al. (2014a) is a rare example of a comprehensive MWE identification model which distinguishes a full range of MWE sequences, including those involving gaps, using a supervised sequence tagging model; like other models in this space, Schneider et al. make use of existing manual lexical resources and they note that an (unsupervised) automatic lexical resource could be useful addition to the model. Otherwise, gaps in MWEs have generally addressed by using full syntactic representations (Seretan, 2011).

Beyond association metrics, other unsupervised approaches to the multiword problem include that of Newman et al. (2012), who used a generative Dirichlet Process model which jointly creates a linear segmentation of the corpus and a multiword vocabulary. Gimpel and Smith (2011) focus specifically on deriving word sequences with gaps using a generative model, with the intent of improving machine translation. The drawback to these generative methods, relative to association metrics, is scalability and a certain degree of randomness, since these methods generally involve Gibbs sampling with many iterations through the corpus to reach an acceptable model. The approach presented here is based on that of Brooke et al. (2014), which was developed explicitly to work well for larger corpora, in the order of a billion words or more; we will leave further discussion of that work for Section 4.

3 Theory and Rationale

Though the approach to identification of phrases presented in this paper should not be viewed as entirely distinct from work on multiword expressions, collocations, lexical bundles, or phraseology, we nonetheless will make use of a somewhat less familiar term to refer to our objects of interest: *for-*

mulaic sequences. A formulaic sequence is defined by Wray (2002; 2008) as "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar." In other words, a formulaic sequence shows signs of being lexicalized. Other than the psycholinguistic fact of being a lexical item for native speakers of a language, there is no other single necessary condition for some collection of words to be a formulaic sequence, but there are many indicators: Wray (2008) lists 11 diagnostic criteria, including exact repetition, a lack of semantic transparency, genre associations, pragmatic effects, nonstandard syntax, and phonological properties; she reports that native speaker intuition is usually sufficient to make a reliable judgment of whether or not a sequence is formulaic.

Wray's conception of formulaic language is explicitly not that of mere exception to the combinatorial creativity of syntax and semantics; she argues that most language can be viewed to some degree as formulaic, and that the use of formulaic sequences is the default mode for most genres, both written and oral. Moreover, her view is that the processing of language in general should be viewed not so much as a bottom-up construction of larger phrases from individual lexical units, but rather as a top-down process where larger chunks are split apart and analyzed as discrete parts only when there is clear evidence for flexibility, a strategy that has a direct analogy in the decomposition approach used here. Another important aspect of the theory is a focus on the linear sequence rather than some other kind of syntactic abstraction (e.g. a dependency relationship) as being primary to the internal representation of multiword phenomena, a perspective which allows for much cleaner analysis of longer and more varied expressions: when cases of sequence-internal flexibility occur, they are handled by the inclusion of a slot or gap which is also part of the sequence. Note that, since humans are fairly skilled at interpreting noisy input of various kinds, the notion of sequence as the default glue of the internal multiword lexicon does not rule out the possibility of greater creativity (e.g. reversing word order), but this should be understood as the speaker abandoning one of the benefits of formulaic sequences (easy processing) for other communicative purposes (e.g. humor).

Second language acquisition is one of the major areas of application for work on formulaic sequences (Ellis et al., 2008). Wray (2008) posits that the difficulty many adult second language learners have reaching fluency reflects, at least in part, an inattention to the role of formulaic sequences, coupled with an expectation that a language should allow for free combination of words governed only by the basic rules of syntax. Modern communicative approaches to teaching tend to encourage learners to express themselves freely so long as they are able to make themselves understood, i.e. to satisfy the short-term communicative goal. However, if full fluency and social integration into the culture of native speakers is a long-term goal, as it is for many immigrant learners for instance, these learners also need to correctly process and eventually produce a wide range of formulaic sequences. Creating high-coverage vocabularies based on real, modern language usage is a first step in helping learners with these challenging but ubiquitous units of language.

4 Method

4.1 Preliminaries

Although there are several key additions that bring the resulting vocabulary much closer to being a comprehensive collection of formulaic sequences, the overarching structure of our method is adapted from Brooke et al. (2014): first, basic statistics are collected from the corpus, and, based on these an initial segmentation of the corpus is carried out. Once a preliminary lexicon is built from these segments, the lexicon is refined based on both the initial statistics as well as the initial segmentation. Brooke et al. applied this refinement process in the corpus to create a final segmentation, but, since the lexicon is our main interest, we will not address that step here. We will first present the use of lexical predictability in the context of the basic (no-gap) model, and then introduce the changes required to accommodate gaps.

First, a few details that would distract from the main discussion of the method below. Following Brooke et al, we set our frequency threshold to be at least one instance in 10 million tokens; all of the work here (including alternatives to our method) are

based on that restriction. Our corpus is a filtered version of the tier 1 blogs in the ICWSM 2009 Spinn3r dataset (Burton et al., 2009), including about 2.4 million blogs or about 890 million tokens of text; for this and other work, we have made a significant effort to exclude texts with spurious repetition (e.g. spam, multiple postings). The part of speech (PoS) information is provided by the TreeTagger (Schmid, 1995), which relative to our needs is quite fast and available for many languages. We collected our statistics using the lemmatized, lower-case forms of words, and accordingly dropped the inflectional information from the PoS tag. We will not discuss the specifics of the algorithms and representations used to collect the statistics except to say that a great deal of attention was paid to keeping the process both fast and memory efficient.

4.2 N-gram decomposition using the lexical predictability ratio

The central mechanism in the n-gram decomposition approach is a measure for choosing among a set of possible segmentations of a text span. Brooke et al. (2014) select a segmentation based on maximizing the conditional probability of each word when the conditioning context is limited to words within the same segment. Our measure also uses conditional probabilities, but we need to distinguish between two types: let $p(w_i|w_{j,k})$ refer to the conditional probability of some particular word/tag pair given a surrounding sequence of word/tag pairs $w_{j,k}$, and $p(w_i|t_{j,k})$ refer to the probability of a particular word/tag pair given only the PoS tags ($j \leq i < k$). For some w_i within some segment whose endpoints are m and n, the *lexical predictability ratio* (LPR) of w_i within span (m, n) is:

$$LPR(w_i, w_{m,n}) = \max_{m \leq j < k \leq n} \frac{p(w_i|w_{j,k})}{p(w_i|t_{j,k})}$$

The LPR for the entire span is defined as:

$$LPR(w_{m,n}) = \prod_{i=m}^{n-1} LPR(w_i, w_{m,n})$$

We use the word *lexical* to refer to this measure because it represents an increase in probability that is apparently due to a lexical rather than syntactic affinity. Other than eliminating syntactic noise, one obvious advantage of using a ratio here is that it naturally emphasizes open-class lexical items, which will tend to have low probability independent of any lexical context, and minimizes the influence of closed-class words; the opposite is true for a measure based on difference, where the influence of a relatively small change to a word with a relatively high initial probability might dwarf a huge relative increase in a low probability word. Given our lexical interests, it is important that our measure be especially sensitive to words in the general vocabulary.

Other than this key change and those relevant to the inclusion of gaps to be discussed in the next section, we preserve intact the initial segmentation algorithm of Brooke et al. (2014). Briefly, the key steps of this process are as follows: First, for each sentence in the corpus, we identify maximal length n-grams, i.e. n-grams above our frequency threshold where any $n+1$-grams that contain them are below the threshold. Where these maximal n-grams overlap, one or more segment boundaries must be inserted in order to create a proper segmentation, with all segments corresponding to an n-gram in our statistics; in this case, the best segmentation is chosen based on maximizing the lexical prediction ratio of the relevant segments, and the segments are counted and taken as the initial vocabulary of the vocabulary decomposition step.

Vocabulary decomposition proceeds by considering each sequence in the initial vocabulary, starting with the longest, and deciding whether or not to break it into two smaller pieces: the counts are added to the smaller pieces which are considered later on in the process. The original algorithm treated the two substrings equally, but here we do not: in practice, in most decompositions there is one, rarer piece that contains the core lexical information, which we call the nucleus (u), while the other is the satellite (s) and is most often a function word or other relatively common word or phrase; the vocabulary decomposition process should be viewed as a process of shaving off satellites until we are left with a lexical nucleus (possibly a single word) that resists further splitting. For each sequence length n, we proceed from the n-gram with the lowest count to the highest. An entry w is broken when either its count $c(w)$ in the lexicon is below the frequency threshold, or when inequality (1) is false for at least one break

index b, $0 < b < n$; here $y(t)$ refers to the number of word types for a given tag, and $c(*)$ refers to the count of all tokens in the corpus:

$$\frac{LPR(w_{0,n})}{LPR(w_{0,b})LPR(w_{b,n})} > \frac{\frac{c(u)}{c(t_u)/y(t_u)}}{\frac{c(w)}{c(t_w)/y(t_w)}} \log \frac{c(*)}{c(s)} \quad (1)$$

The left-hand side of the inequality represents the amount of lexical predictability that is lost (over all words in w) when a break is inserted at b. The higher this number, the more the sequence resists decomposition. The first term on the right side represents a ratio of the counts of the lexical nucleus to the full entry: the higher the count of the lexical nucleus relative to the count of the full entry, the more likely we are to break. However, we do not compare these counts directly: mirroring what we have done with the conditional probability in the calculation of LPR, we consider these marginal probabilities relative to the expected marginal probability (count) of a term with that tag sequence, which is simply the total count for the tag divided by the number of types. All these counts are derived from the statistics of the initial vocabulary. In the second term on the right, more common satellites decrease the chance of a break, which counters the property, mentioned earlier, that the LPR can be rather low even for entirely predictable satellites if the marginal probability is already high. Since $c(*)$ is necessarily larger than $c(s)$, this term also serves an initial threshold that must be overcome by increased LPR and/or a higher than expected count. Finally, when there are multiple breakpoints which render the inequality false, or when a break is forced due to low counts, the break which is actually carried out is that which maximizes the difference between the right-hand side and the left-hand side. When all entries have been examined in this fashion, the entries which have been preserved are the final vocabulary.

4.3 Including gaps in the decomposition model

Although it is essentially impossible to describe all formulaic sequences using a single syntactic representation, the slots or gaps within English formulaic sequences are relatively well behaved: in the MWE corpus of Schneider et al. (2014b), for instance, a manual analysis revealed that essentially every gap consisted of a noun phrase (e.g. *point * out*) or a noun modifier (*have * complaints about*). Although it is possible for a gap to have complex content, this is not typical, and anyway it is not necessary to cover all possible cases to do successful lexical induction; for English, we define our gaps as a sequence of 1 to 4 words whose tags satisfy this regular expression:

PP|[(PDT)(DT)JJ*[NN|NP]*(POS|PP$)JJ*NN*]

For us, a gap n-gram is just a regular n-gram with an additional index indicating the location of the gap: in essence, we can collect gap n-gram statistics by first searching for a tag sequence that matches our gap regex, and then counting n-grams around the sequence as if it were not there. This is efficient and defensible, since in many cases knowing the syntactic content of the gap would be redundant, since it entirely predictable from the surrounding context. We do not consider the possibility of multiple gaps: though such patterns exist, they are quite rare (Schneider et al., 2014a).

When we have collected the same statistics for our gap n-grams as we had previously collected for our regular n-grams, we can carry out an initial segmentation. When we are able to match a gap n-gram with a gap size of m, for the purposes of proposing initial segmentation alternatives we treat it as if it were a regular $n+m$-gram spanning the full extent of the gap n-gram. When a segment which corresponds to one or more possible gap n-grams is considered, we have to solve a new segmentation problem: inserting two special gap breaks which define the outer gap n-gram, plus (possibly) additional breaks within the gap if needed. For the purposes of calculating LPR, we treat the two outer pieces as a single span, and the contents of the gap as an entirely independent segment. Under those restrictions, we choose breaks to optimize LPR across the entire segment, and, eventually, the entire local context.

After segmentation, the resulting initial lexicon has a mixture of contiguous and gap n-grams. During the lexicon decomposition process, the two kinds are not differentiated with respect to the order in which they are examined. The main difference is that when decomposing regular contiguous n-grams we now have a new option: we can split to create a gap n-gram. For gap n-grams, we do not allow additional gaps; only a single break is possible, though a

break at the gap creates two regular n-grams, while one in any other location preserves a gap n-gram. There are only minor changes to the inequality that decides whether a break should occur; if we are considering decomposing by adding a gap, then the denominator of the left-hand term of (1) is now $LPR(w_{b_1,b_2})LPR(w_{0,b_1+b_2,n})$, where $w_{0,b_1+b_2,n}$ is understood to be the string consisting of the concatenation of w_{0,b_1} and $w_{b_2,n}$; when calculating LPR, any conditional probabilities involving spans that cross the gap must use the appropriate gap statistics.

5 Evaluation

Evaluation of large-scale automatically-generated lexicons is notoriously problematic: comparing to a reference lexicon is usually not valid because the reference lexicon, if one exists, is not complete (if it were, why build an automatic lexicon at all?) and therefore it is impossible to accurately estimate precision. The output of a particular approach (i.e. the lexicon) can be judged directly, but this only measures precision, not recall, and it is a short-sighted approach with regards to evaluating future improvements. In this work, we take advantage of the fact that we are assuming an initial n-gram frequency threshold, which greatly reduces the space of all possible n-grams (both contiguous and gap) that we are actually considering as possible formulaic sequences. Although there are still many more bad n-grams than good, the imbalance is not so great as to make annotation impossible: we can sample from the set of possible n-grams, judge them as being good or bad formulaic sequences, and then compare with the output of lexicon creation processes to calculate precision, recall, and F-score.

Our annotation project involved 3 judges, a number chosen so we could use consensus for the creation of a gold standard. The judges, all college-educated native English speakers, were introduced to the basic theory of formulaic sequences and their diagnostics (Wray, 2008), and then instructed that their main task was to identify canonical formulaic sequences, where *canonical* was understood to mean a sequence that contains all the words that would most commonly be used as part of the formula, and no words whose presence seems incidental or the result of rule-driven processes: if an n-gram was larger, smaller, or otherwise distinct from a canonical sequence but the formulaic sequence was nonetheless identifiable, we offered another option (the n-gram "recalls" a formulaic sequence), which we don't use directly in our evaluation, but which we used to help the judges focus in on canonical formulaic sequences. To help them make their annotation, the judges were presented with 5 sample sentences from our corpus.

We annotated 1000 contiguous n-grams and 1000 gap n-grams in this fashion, with the n-grams randomly selected from sets of roughly 1.5 million n-grams in both cases. For contiguous n-grams, 16.9% of the n-grams were judged to be canonical formulaic sequences, but from gap n-grams this number was much lower, only 2.9%. Kappa is problematic with such a serious class imbalance (Di Eugenio and Glass, 2004), so instead we calculated an average F-score across the 3 annotations[1], which was found to be 0.62 for contiguous n-grams and 0.42 for gap n-grams, numbers which reflect a certain amount of subjectivity in the judgment task, but also considerable agreement. These F-scores also provide an estimate of a practical upper bound for our evaluation. To create a gold standard annotation, we used the majority judgment. We also had a single judge produce separate sets for development purposes.

Our second evaluation uses an existing resource, a section of the English Web TreeBank (Bies et al., 2012) that has been annotated for a full range of MWEs (Schneider et al., 2014b). As mentioned earlier, formulaic sequences are a broader category than MWEs (as traditionally understood), and indeed a manual analysis of a portion of the corpus revealed many formulaic sequences in this set which are not annotated. Nevertheless, since all MWEs are formulaic expressions, we can make use of the annotation as a secondary evaluation: for positive examples, we extracted all MWEs in the corpus (except for MWE-internal MWEs, which we ignored) above the frequency threshold (which was the vast majority of them, since the genres of the ICWSM and the Web TreeBank are similar), and as negative examples we extracted all n-grams (both contiguous and

[1] That is, treating one set of judgments as a gold standard and each of the others as an attempt to reproduce it. For all calculations of F-score in this paper, a "positive" classification is a judgment that the sequence is indeed formulaic.

Table 1: Comparison of various automatically generated lexicons with two annotated test sets. P = Precision, R = Recall, F = F-score, ME = mutual expectation, pred decomp = prediction decomposition method of Brooke et al. (2014). Bold is best in column.

	FS test set						**MWE test set**					
	Regular			**Gap**			**Regular**			**Gap**		
Method	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Count	0.21	0.24	0.23	0.00	0.00	0.00	0.18	0.77	0.29	0.05	0.47	0.09
c-value	0.22	0.23	0.23	0.05	0.07	0.06	0.11	0.08	0.10	0.01	0.01	0.01
PMI	0.23	0.22	0.22	0.12	0.14	0.13	0.13	0.11	0.12	0.03	0.01	0.02
ME	0.23	0.23	0.23	0.00	0.00	0.00	0.18	0.59	0.28	0.05	0.27	0.09
Pred decomp	0.35	0.50	0.42	0.09	0.14	0.11	0.26	**0.83**	0.40	0.09	**0.59**	0.16
Simple LPR	0.40	**0.54**	0.46	0.10	**0.48**	0.17	0.38	0.74	0.50	0.17	**0.59**	0.27
LPR decomp	**0.51**	0.45	**0.48**	**0.22**	0.31	**0.26**	**0.47**	0.72	**0.57**	**0.33**	0.50	**0.40**

gap) above the frequency threshold where at least one word in the *n*-gram is contained within a MWE, and at least one word is not. This tests to see whether our lexicon would be potentially useful for this task while at the same staying agnostic about the status of other potential formulaic expressions beyond the scope of the MWEs. For regular *n*-grams, this process yields 1273 positive examples and 7272 negative examples: for gap *n*-grams, there are 263 positive examples, and 6764 negative examples, for both types the class imbalance corresponds roughly to the class imbalances in our formulaic sequence annotation. Note that, relative to our main evaluation, this test set is populated with common expressions; for comparison, only 5.2% of postitively identified formulaic sequences from our test set are in WordNet, whereas 31.5% of the MWEs from the Web Tree-Bank test set are. As with our main evaluation, we use precision, recall, and F-score.

We compare our model first with lexicons built using established measures which can be applied to general sequences beyond 2 words: pointwise mutual information (Church and Hanks, 1990), mutual expectation (Dias et al., 1999), *c*-value (Frantzi et al., 2000), and raw frequency: all can be calculated for both regular and gap *n*-grams using the statistics extracted for our LPR-based method, and then a threshold selected which builds a lexicon of the size we would expect to be ideal given the ratio of good to bad sequences found in our annotation (i.e. the best 16.9% of regular *n*-grams, the best 2.9% of gap *n*-grams). We also build a vocabulary using the original Brooke et al. (2014) prediction-based *n*-gram decomposition method (pred comp), using the same statistics; though it did not originally handle gaps, we updated it to allow gaps, in the same way as our approach. Finally, we consider a simplified version of the LPR approach which does not carry out an initial segmentation: Starting with all *n*-grams, we use inequality (1) to make a decision whether to keep the *n*-gram in the lexicon. In this version of (1), $c()$ is now the original count from the full corpus statistics, not the initial lexicon, except that we subtract from their counts the occurrences of u and s that are also occurrences of w.

6 Results and Analysis

The results for the various automatically generated lexicons for both test sets are in Table 1. First, we note that none of the simple measure-based lexicons offer competitive results, and the results for gap *n*-grams are consistently poor. There is also no clear standout, though ME seems to have the edge on average, a result which is consistent with previous work. Relative to these simpler methods, the original *n*-gram decomposition approach does fairly well in the regular test sets; its results for gap *n*-grams, however, are not impressive. The simpler LPR method is almost indistinguishable from our full method with respect to regular *n*-grams, but its performance with regards to gap *n*-grams indicates a benefit from using the full decomposition pro-

cess, though it is not as large an effect as the use of LPR. Our LPR *n*-gram decomposition is consistently the best for both test sets and *n*-gram types and the F-scores in our test set indicate that, relative to the original *n*-gram decomposition technique, we have made real progress towards the practical upper bound suggested by the between-human F-score.

Our final formulaic sequence lexicon has 227,188 entries; 184,246 are contiguous, and 42,942 have gaps. For comparison, our single-word vocabulary with the same frequency cutoff is 72,117, supporting the long-standing claim that the multiword lexicon of a language is significantly larger than the single-word lexicon. For contiguous *n*-grams, 2-word entries compose 36.5%, 3-word entries 33.3%, 4-word entries 20.2%, and 5-word entries 7.7%; for non-contiguous entries, 3-word entries are the most common (44.0%), followed by 4-word entries (27.1%), 2-word entries (17.2%), and 5-word entries (9.9%). With respect to variety, although three 2-word part-of-speech combinations (NN NN, NP NP, and JJ NN) make up close to 21% of the contiguous lexicon, beyond those three there is significant variety, with no single PoS combination accounting for more than 2%, and the top 20 part-of-speech combinations covering only 37.6%. The situation for gaps is even more extreme: only verb/noun combinations (4.9%) stand out as being particularly common. Though a certain amount of this variety might be due to error, in general we believe it reflects the huge variety of potential syntactic realizations of formulaic sequences; essentially any words that regularly appear in sequence could be formulaic.

Looking at just the first 50 (randomly ordered) entries in our lexicon for each type we indeed see much variety, clearly formulaic contiguous phrases like *just the two of us*, *into the depths of*, *would not have been possible without*, *interestingly enough* and gap sequences like *watch * in action*, *about * or so*, *millions of * worldwide*, *implementation of * program*, *gave * a heart attack*, *scold * for not*, *beyond * capabilities* and *back to where * started*. There are some systematic errors, however: probably the biggest single problem is pronouns, which are often highly predictable in a particular context despite being theoretically flexible, e.g. *find myself wanting to*. Another clear problem is lexical predictability that is due to word classes (e.g. *in Long Beach*); information about these classes should be integrated into our background syntactic predictability. When there is enough variability in usage that smaller pieces of a larger phrase get segmented, LPR will often hold these incomplete pieces together, e.g. *your way through*. Looking at the gap lexicon, there are some syntactic patterns (*a * or a*), some semantic patterns (*parents of * kids*), and other cases where it is not clear why a gap was necessary since we would expect little or no variation (*as * weapon against*): often these last cases were close to the frequency threshold and there was just enough variation that the canonical sequence (in this case, *use * as a weapon against*) fell below the threshold. Future work should look at having a more flexible threshold.

7 Conclusion

We have presented here a very general approach to automatic acquisition of multiword lexicons, to our knowledge the broadest to date. By focusing on (apparently) lexical effects using the lexical predictability ratio, while at the same expanding the scope of the output to include gap phrases, we can make a genuine claim that our lexicon reflects a significant portion of the formulaic vocabulary of the language, especially given the size of our corpus that this method can accommodate and the choice to avoid filtering of particular syntactic types, which was justified by the diversity we found in our output lexicon. Our interest here is in educational applications, where having an explicit representation (rather than the implicit lexical information contained in, for instance, language models) can be used to help a learner expand their multiword vocabulary; this is particularly true for formulaic language which is fairly compositional, and therefore may not be obviously formulaic to a learner nor likely to appear in a standard dictionary. There is still work to be done in addressing the errors we see in our lexicon, but our results nonetheless represent significant progress towards the human upper bound suggested by our annotation project, and the evaluation method and resources introduced here should spur future work.[2]

[2]The test set, the automatically-generated lexicon, and the lexicon-creation software are available at http://www.cs.toronto.edu/~jbrooke

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada and the Mitacs Elevate program.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank. LDC2012T13, Linguistic Data Consortium, Philadelphia.

Julian Brooke, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2014. Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proceedings of The 25th International Conference on Computational Linguistics (COLING 2014)*.

Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM '09)*, San Jose, CA.

Yu-Hua Chen and Paul Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2):30–49.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Kenneth Church. 2011. How many multiword expressions do people know? In *Proceedings of the Workshop on Multiword Expressions*, pages 137–144.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101, March.

Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In *Proceedings of Conférence Traitement Automatique des Langues Naturelles (TALN) 1999*.

Nick C. Ellis, Rita Simpson-Vlach, and Carson Maynard. 2008. Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42:375–396.

Stefan Evert. 2004. *The statistics of word cooccurrences–word pairs and collocations*. Ph.D. thesis, University of Stuttgart.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3:115–130.

Kevin Gimpel and Noah A. Smith. 2011. Generative models of monolingual and bilingual gappy patterns. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Sylviane Granger and Yves Bestgen. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52:229–252.

David Newman, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin. 2012. Bayesian text segmentation for index term identification and keyphrase extraction. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, pages 47–50.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014a. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206, April.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 455–461.

Patrick Schone and Dan Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP '01)*.

Violeta Seretan. 2011. *Syntax-Based Collocation Extraction*. Springer.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge.

Alison Wray. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford University Press, Oxford.