# Resubmission of TACL #1137, Unsupervised Acquisition of Comprehensive Multiword Lexicons using Competition in an $n$-gram Lattice.

May 30, 2017

## Contents

## 1 Author(s) cover letter responding to the original reviews

Starts on next page.

Submission # 1137
Action Editor: Noah Smith

We would like to thank all of our reviewers for their detailed comments. We have tried to address the important points as best we can, but were limited to 3 extra pages (beyond the 10 pages of the original submission). As all the reviews had numerous suggestions for improvement, with only one obvious area of overlap, we have had to pick and choose what we could add. In cases where we did not change the paper, we provide a justification below.

We should mention one major change that wasn't prompted by individual reviewer comments: there were clear misinterpretations by our original Japanese annotators which led to overannotation of FS, leading us to do another round of annotation since the original submission. The Japanese results have therefore changed, and are in fact more consistent with the other languages.

REVIEWER A

> In the introduction, the proposed approach is claimed to be "effective, expandable, and tractable", however, it is not clear how each of these claims is evaluated in the paper (effective, expandable and tractable). The approach is also claimed to have the candidates compete for "the best (most parsimonious) explanation for statistical irregularities due to lexical affinity", but concepts like parsimony and lexical affinity need to be defined (what are they, how candidates display them, and how they are measured, ...).

Our intent is that effectiveness is demonstrated by our main evaluation. Tractability is demonstrated implicitly by the fact that we apply our method to corpora much larger than many of our competitors can handle. We have added more details about this to the paper, though we do not want to belabour the point, since it is not the primary purpose. Our claim of expandability relates to our discussion of future work, however we have added a bit more discussion about exactly how expansions would be added without major changes to the model.

> The paper should explicitly state what the hypotheses are in relation to the utility functions, particularly in relation to the approach by Brooke et al. (2015), which is simpler than the proposed approach. Moreover, the state-of-the-art has several simpler formulations for the identification of general formulaic sequences of flexible sizes (e.g. Silva et al. 1999 for contiguous & non-contiguous cases and Villavicencio et al. 2007 for contiguous cases), and the text should justify why this particular approach is appealing, and advantageous in relation to the others. The addition of an error analysis would help highligh these points, helping to clarify the advantages of the particular approach especially in relation to Brooke et al.

We have added a new paragraph contrasting our method to other algorithms, and a qualitative discussion which doubles as an error analysis.

Note that we looked at Villavicencio et al. 2007, but it is very limited in scope (they only deal with trigrams) and doesn't provide any workable comparison in the context of unrestricted FS extraction beyond ranking with simple association measures, which, as we mention, have already been shown to be inferior to the more sophisticated models we are comparing against here.

> The paper should also discuss the motivation for each of the node

> interaction functions for rewarding/penalizing competitive candidates in the lattice, including how they are expected to contribute to the overall performance of the approach and if the authors considered alternative functions or why these in particular, which linguistic intuition they wanted to capture, why define E0 in terms of the minLPR and E1 of a cost parameter, why they are introduced in the explainedness as exponents (why not using a linear dependence, for example. A suggestion for helping the presentation is for the authors to rewrite E0 and E1 into a single formula that combines the three measures in a utility function, as it would be easier to explain (how E0 and E1 compete, under which circumstances each of them wins), particularly with the addition of an example. In addition, when presentating these functions, one problem is that although d0 and d1 are used to define E0 and E1 in section 3.2, they are only defined in later sections (sections 3.3.2 and 3.3.3). As a result the discussion appears too abstract as the functions have not been completely defined yet. There should be some explanation for their motivation/intuition, as it's not clear what they are trying to capture, how E0 and E1 would vary for FS and non-FS, etc.  My suggestion is to first introduce and motivate the ideas behind them, and later define them together. For d1, there is also a difference in the number of arguments of "oc" in the text (oc(x,y)) and in the definition of d1 (oc(oi)). d0 is defined as a uniform combination of clear and cover, but it would be helpful to have a discussion of the effects of different weights for each of them for the different languages.

We have substantially reorganized the presentation of the model, and are now providing an explicit objective function, at the same time getting rid of the confusing forward pointers and many of the extra symbols.

We hope that the general purpose of each interaction in terms of the way that it inhibits other nodes is clear. We have added a bit more detail about the choice of the clearing function to both that section as well as the discussion.

> In terms of the methodology, a discussion is needed about the motivation for adopting Brooke et al.'s lexical predictability ratio (LPR), and for the modifications proposed, such as the use of minLPR instead of the product of the LPRs and why the focus is on the weakest link (was there any comparison between the use of minLPR and the product of LPRs, is it done for efficiency reasons, etc).  Moreover, the construction of the lattice needs to be explained, including how the nodes relate to one another. For instance, in figure 2, the solid lines indicate subsumption, but "keep everything under wraps" is linked with "be keep * under" via a subsumption relation, even though the latter contains elements that are not in the former (shouldn't it be an overlap link?).

We have added a bit more to the discussions of LPR, to be more explicit why LPR is a good metric for our lattice. We've also fixed the incorrect relations you mention, which should hopefully resolve the confusion.

> In terms of the evaluation, the distinction between contiguous and gapped test sets, which are later merged into a single test set per language when presenting the results, is not helpful. Instead, maintaining the distinction would help clarify the advantages of the proposed approach in each language, particularly in relation to the error analysis.

As we note in the paper, adding the distinction between continguous and gapped simply isn't defensible given the lack of positive gapped examples across some of the datasets (since both precision and recall rely on primarily on true positives). We would very much like to present this distinction, but the scores depend primarily on a handful of examples for some of datasets and the results are erratic, they would confuse rather than elucidate.

> Additionally, for the contiguous n-grams in these languages a comparison with LocalMaxs could be done.

Actually that's not possible. LocalMaxs (like DP-seg) does not scale, due to the fact that it needs to calculate counts for every n-gram in the corpus; it is quicker than DP-Seg, but the memory usage is immense for large corpora. We have added a mention of this fact in the paper.

> Moreover, the authors need to add the statistical significance of the results in tables 2 and 3 as the approaches often differ only in the second decimal place.

We have added some statistical significance information.

> In Table 2, it would be helpful to have also the results for only the cover function (Lattice-cl-ovr), to determine how much the other functions contribute.

Thank you for the suggestion. This has been added to the revised version.

> A figure with the top/bottom candidates proposed by the approach be helpful.

There are no explicit top/bottom candidates in our model, as it's not a ranking model and it would be incorrect to present it as such. N-grams could be ranked by the explainedness of their node, of course, but that doesn't properly reflect their overall influence in the model. We do have a new qualitative analysis with random samples from our lexicon.

> Add the list in an appendix for English and explicitly explain the lists for Croatian and Japanese. How many expressions and how many cases they cover.

We have added some samples for each language and some general statistics for the lexicons of each language. As should be now clear, the lists of expressions generated by the model are far larger than can be included in a paper. We will be separately releasing all the lexicons.

> The paper also needs to add a list of existing MWE lexica or lexical resources with MWEs available to the community (e.g. those listed in the MWE community website, or available in WordNet or other resources).

We have to demurr on this. Our paper is not about MWEs specifically, we do not use MWE resources to build our model nor to evaluate it, so a focus on available MWE resources seems a significant detour, especially given the space constraints.  It would also be very difficult to do the topic justice, especially given our cross-linguistic interest.

> Stylistically the text would also benefit from a revision that splits sentences that are too long (there are several throughout the text). For instance, the last sentence in the first paragraph of the introduction is 9 lines long: 'We present an effective, expand- able, and above all tractable new approach to com- prehensive multiword lexicon acquisition that aims to find a middle ground between standard MWE ac- quisition approaches based on association measures (Ramisch, 2014), and more sophisticated statistical models (Newman et al., 2012) which fail to scale to the large corpora which are the main sources of the distributional information in modern NLP

> systems.').

Thank you for your suggestion. We have split up some of the longer sentences, hopefully improving the readability.

> Moreover, some passages are copied from Brooke et al. (2015), but lack the
> citation, such as:
> *"Other mea- sures 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." (page 2) and
> *"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" (page 3)
> There are also other instances, and these need to be rephrased.

We have rephrased the relevant parts of the literature review, though the basic content is unchanged.

> For the annotations, the text mentions 3 annotators, but doesn't specify
> whether they were different for each of the 3 languages.

Yes, they were different for each of the 3 languages. Since they were native (first language) speakers of each of the relevant languages, we believe this implies that they were different across languages.

> When comparing the languages, the paper also mentions that "free word order
> actually results in more of a tendency towards contiguous FS, not less.",
> but it is unclear where this is shown.

We've tried to explain this a bit better.

> * “accessible by analogy (e.g., glass limb or government ambiguity)” by
> analogy with what? Specify and add the original expression

We have made the analogy explicit.

> * “define the explainedness of a node in terms of two functions” Which two
> functions?

Hopefully this is clearer in the new formulation.

> * In section 3.2 can the authors explain how two valid but similar FSs would
> be treated (e.g. keep * under wraps/surveillance)?

The only kind of interaction between these nodes involves their shared n-gram subsequence (keep * under), as described in the example in the paper. We have expanded the example a bit.

> * "We also use the concept of hard covering to ad- dress the issue of
> pronouns, based on the observation that pronouns often have high LPR
> values". This sentence is too vague. Explain using an example.

We've added an example.

> * How was the 2/3 threshold for hard covering determined? Why also have soft
> covering?

Using performance in the (English) development set, we have made that clearer. Our intention is that the "keep * under wraps/surveillance" example shows why soft covering is also important, we have added a bit more explanation.

> * "has been turned on, the covering, blocking, or overlap- ping effects of
> these other nodes" --> blocking shoud be clearing

Yes, thanks.

> * For the efficiency restrictions explain if other values have been tested
> for them. How did they affect the overall performance? A table detailing the
> behavior with different values would be helpful.

We've added a bit more detail, though perhaps not as much as you're asking for (due to space limitations).

> * "which was roughly the same size (in terms of token count) as the English
> corpus" specify which English corpus

The ICWSM. We've added that.

> * In Table 1 the values are less than 2000 (for ICWSM) and 1000 for the
> other corpora. Add also the initial numbers for each type of FS/non-FS.

As we mention in the text, to improve the reliability of our evaluation we drop instances where just one annotator tagged as FS. This obviously results in test sets that are less than the original sample size. This should now be clearer.

REVIEWER B

> Assuming that the solution is indeed scalable and empirically viable -- I am
> wondering if the same idea could be used in the context of other tasks
> (morphological composition in morphological lattices, preferred expressions
> in lattices of proposed machine translations, etc). It would be nice if the
> author could comment on the applicability of the technical solution in other
> domains (if there is any).

This is interesting, and so we've added a discussion of the morphology idea to the paper (The short answer is yes).

> In the case of syntax, much research in theoretical linguistics and in
> language technology has proposed the idea that idiomatic expressions (of
> which MWEs are a special case) are the rule, rather than the exception, in
> NL grammar. Among the theoretical proposal one can find versions of
> Construction Grammar (Goldberg 1996) where language knowledge is composed of
> complete constructions, rather than primitive units (words) and in
> technological proposals one can find the idea of Data-Oriented Parsing

> (Remko Scha, Rens Bod), wherein each syntactic subtree (with potential slots
> for substitution) is an FS-like element that can potentially be used and
> reused for analysis and generation. And so: while the relation of the
> proposed FS to existing theories of strongly lexicalized fixed expressions
> on the one end is clear, it is unclear what is the relation larger idiomatic
> syntactic constructions with lexicalized or unlexicalized slots on the other
> end -- and this relation should be made clear, for the work to be properly
> situated.

Note that we are not ourselves proposing FS, we are borrowing the theory behind FS from the work of Wray (2002;2008), who includes an extremely thorough discussion of its relationship to other linguistic theories (Chapter 7 of the 2008 book in particular), including Construction Grammar. We have added a bit more detail to our introduction of FS to make it clear that, yes, most "constructions" are FS, and we have added constructions to Figure 1. However, ultimately this is a technical paper, and we are limited in the space we can dedicate to theoretical concerns.

> The case of morphology is far less clear, however it is equally important,
> and it has to do with the basic question of what is the unit that constructs
> the FS (or the lattices themselves in empirical terms). Are these words?
> morphemes? lemmas? POS? inflections? In the case of words, what is the
> relation of inflected sequences forms / inflected idioms (kicked his bucket
> vs kicked her bucket, for instance) -- are they the same? related?
> overalapping? subsumed?. As it stands, the discussion of morphology comes
> out only later as an afterthought -- where the construction of the lattice
> in Japanese (in the evaluation section) involves some morphological
> segmentation whereas the lattice construction for the other languages
> doesn't. It is thus unclear if the algorithm finds sequences that are of the
> same type, or are comparable for that matter, across those languages. The
> selected FS may be incomparable in terms of effective length (morpheme
> sequences tend to be longer), the status of function words vs. inflection in
> FS, and so on. The authors should put forth clearly what is the status of
> morphology inside their general theory (and in terms of their regular
> expressions), and then later revisit the empirical ramifications of their
> decision -- for instance what happens when moving between typologically
> different languages.

We have added a bit more to both the introduction of FS as well as a paragraph in the discussion that addresses this point. Ultimately, in all three languages we are tokenizing and lemmatizing using standards for each language, and building our FS from the result. However, this is not a theoretical choice, but a practical one; it is undeniable that FSes often involve morphological components that we are overlooking. With Japanese, due to the lack of morpheme boundaries, morphological segmentation is a more standard form of tokenization, and so we have taken advantage of this to bring Japanese somewhat closer to our theoretical ideal. Given that many of the case markers of Japanese correspond to prepositions in European languages (or the possesive marker in English), the actual difference in generated FS is fairly minor, except for the extra overhead of word segmentation in Japanese that is provided by word boundaries in English (in Wray's theory, most words would be FS by default).

> The author discuss at length the method, the construction of the lattice,
> and possible relation between arcs (covering clearing and overlapping) and
> their related scores. The objective function however and the respective
> optimisation algorithm are discussed informally in passing. I would much
> prefer a more precise formalisation of the objective function and greedy
> algorithm, possibly with a running example of several states of progress of
> the algorithm over a sample lattice. This would also help to clarify what

> "locality properties" are at play here, which justify a greedy solution that is locally optimised. As a more general note, I think it should be made clear if the target of the lattice construction is a single lattice for the entire text, optimized at once, or multiple lattices, reflecting different aspects of the vocabulary, optimized separately.  As a minor note, there are two many forward references in the text, IMO, which hinder understandability. Example "where d1 and d2 are ... as detailed in the next section". I would prefer the details closer to where they are relevant.

Thank you for your suggestions. We have reorganized the expanation of the method to remove forward references, and have provided an explict objective function and algorithm. We decided, though, that a full example of the optimization algorithm involved too much overhead with only modest explanatory benefits.

By "text", we assume you mean corpus. It is hopefully clear that the algorithm works at the level of all n-grams for the corpus, and not texts within the corpus (this is what makes the model practical relative to token-level models like DP-seg); the nodes of the lattice are n-gram types, not tokens. As it happens, most nodes are connected to most other nodes by some pathway in the lattice, but there are nodes (mostly individual bigrams) which do not, and thus are not part of the main lattice.

> It is unclear to me why comparing the collected lexicon using the proposed method to a sample set of ngrams that respond to a certain frequency threashold is a sound comparison-- isnt the frequency threashold just a simpler and less sophisticated baseline for MWE extraction? It appears that you use this method not as a baseline, but as an actual method to extract gold standard for comparison with your method. I understand that others have used such evaluation methods before -- but I would need much more convincing in order to accept the claim of efficacy. If I misunderstood, and you do not use it as a gold for comparison, or you use a different gold, please clarify.

We are not entirely sure we are intepreting your concern correctly. It is certainly not the case that we are using high-frequency n-grams directly as a gold standard (i.e. we assume that high frequency n-grams are FS simply by virtue of their high frequency): this should be clear given the discussion of the annotation process, where we taken the n-grams and annotated them as being FS or not. To make this absolutely clear, we have included frequency as a baseline association score. It is poor, but obviously would be unbeatable if our gold standard was built based simply on frequency.

A more subtle issue is that using frequency-filtered n-grams as the sampling pool for our FS gold standard means that there are FS in the corpus that cannot possibilty be considered in the evaluation because they do not appear sufficiently. This is true, and unavoidable in this evaluation setup. As discussed in Brooke et al. 2015, sampling from frequent n-grams in the corpus brings with it a number of benefits: in particular, the test set reflects the variety of FS that appear throughout the entire corpus (many of which would not appear in any existing lexicon or in a small sample of texts), annotators can look at variety of examples to make an judgement, and (most importantly) it is possible to calculate a proper F-score. The trade-off, however, is that there needs to be some initial filtering of the sample so that annotation process is not "finding a needle in the haystack" as it would be if you considered all n-grams in the corpus. As we discuss in 3.1, we investigated coverage in an existing English MWE annotated corpus using n-grams from the ICWSM after applying frequency cutoffs and found that type coverage is quite good (token coverage is even better, of course); that is, we are not losing very many FS with our filtering, at least not ones that occur across corpora of the similar genre (both are social media corpora). With a large corpus and an appropriately low frequency cutoff, we believe that annotations of frequency filtered n-gram samples is a reasonable way to solve what is without a doubt a very challenging evaluation problem. We have added a slightly longer justification of the evaluation method to the paper.

REVIEWER C:

> Where I had trouble was with the details of the model. First of all, much ofit seemed to be ad hoc; even accepting the notion of LPR from prior work, there are many heuristic (or at least not obviously natural) choices made to operationalize the general lattice idea mathematically. Second, even if the model is reasonable, it is not entirely clear how the different parts fit together into an optimization problem. There is mention of a parameter C; and there is a notion of nodes being activated or not (I don’t think notation is introduced for these binary variables, though it would be appropriate). Are these the only things that have to be tuned? Is the calculation of explainedness deterministic based on these? There is a greedy search to optimize explainedness of the nodes in the lattice, subject to various heuristics. Presenting a formal algorithm for the optimization may clarify things considerably.

We have reorganized the method section and added an explicit objective function, which should help with some of the problems you identified. In terms of the central optimization problem, the only variables are exactly the on/off status of each node, and, if we exclude the parameters used in the creation of the lattice (e.g. the n-gram frequency threshold) and those involved in the efficiency restrictions, the only (hyper)parameter in the optimization is C; we have a included psuedocode for optimization algorithm. With regards to the ad hoc parts of the model, we have added a discussion separating the core aspects of the model from some of the more ad hoc choices made in the particular instantions of component functions (such as clearing), making it clear that we believe these can potentially be improved on (both in terms of performance as well as mathematical elegance), and that our primary purpose here is to show that overall architecture of the model is a promising one.

> I think it would also be worth discussing how the lattice model relates back to the cognitive/psychological literature. Would you predict (or are you aware of any evidence) that node interactions such as covering, clearing, and overlapping have an analogue in human language processing? Do any of the design decisions intentionally sacrifice cognitive plausibility for engineering reasons?

In general, we did not consider cognitive plausibility when creating the model, and we are certainly making no claims as such. For starters, our model makes use of more data than a human being could possibly read in a lifetime, and it is clear humans identify FS with far less evidence than our model requires. With space, we could discuss our model in the context of the more cognitive aspects of the theory of formulaic sequences as proposed by Wray, but given that this wasn't a focus in the development of the model, we think it is really too much of a digression to justify.

> - After the equation for LPR, it would be good to give the intuition: choosing a context span such that w_i is more likely given its context words than its context POS tags.

This was removed from an earlier draft due to space, and we've added back in.

> - Inter-annotator agreement is surprisingly high. Could you elaborate on Wray’s criteria for FS that were applied by annotators?

Your surprise about the Kappa being high is based on a misunderstanding: the kappa number reported in the table is after filtering to remove borderline cases (i.e. it is the kappa for the final test set, not the original sampling). To avoid confusion, we've added the original kappa score which is quite a bit lower. Your intution is correct that there is significant subjectivity to FS identification (more so than MWE annotation), and as such we would never rely on a single annotator for building a FS test set. We have not added more information about Wray's criteria; though it provide a starting point for annotation, more relevant to achieving reasonable agreement are the specific annotation guidelines which we created. Presenting them in the paper would be too much of a digression, but we promise to make them public.

> - I would like to see qualitative analysis of the sequences that were identified as formulaic (or not) by annotators/the method. E.g.: How many of them are syntactic constituents? Are there a wide variety of grammatical kinds of expressions? Are high-frequency as well as low-frequency sequences detected? What is the distribution in terms of length? What kinds of gapped sequences are there? It would be nice to see some examples.

In the discussion, we have added a thorough analysis of specific examples and some statistics, including most of the information you ask for (implicitly if not explicitly).

> - Related to the previous point: While noting the distinction between FS and MWEs, is it perhaps worth measuring recall against MWE datasets? Most MWEs apart from proper names should count as FS, right?

Do you mean as a way of evaluating different systems built on the same corpus, or just to get a sense of coverage more generally? One central problem with using recall for evaluation is that in order to make it fair, there must be some control of the size of the lexicon. Neither our system nor our best performing competitors are ranking models, none of them offer an easy or fair way to simply take the best N words and look at recall. And even if they did, there's other kinds of biases that would come into play: for instance, PMI and similar metrics do well with rare terms that correspond well to the MWE lexicon of WordNet, for instance, which (at the type level) is dominated by certain kinds of terminology (biomedical terms, particularly) which are not really used outside a very specific community. This relates to a more fundamental problem with the idea of using recall in external lexicons as any kind of useful measure: we could not expect any particular corpus (even a very large one) to do justice to the full range of FS/MWE for a language, since most of which appear regularly only within specific genres. For example, we could not expect to build a good lexicon of English verb-particle constructions from a biomedical corpus, because many are too informal to be appropriate for that genre (e.g. "mess up"). With the qualitative analysis we have added, the genre-specificity of the FS we extract is hopefully more clear.

Though it perhaps does not directly address your comment (since it doesn't involve an external MWE lexicon), we were nonetheless inspired by your comments to include an analysis of the recall of MWE vs. non-MWE FS in the context of our existing ICWSM test set, and a small investigation of the recall (also in the ICWSM) of extremely common MWE from another MWE-annotated social media corpus.

> Typographical issues:

<snip>

We have addressed these typos, thanks.

## 2 Revised submission

Starts on next page.

# Unsupervised Acquisition of Comprehensive Multiword Lexicons using Competition in an $n$-gram Lattice

**First Author**
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

**Second Author**
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

## Abstract

We present a new model for acquiring comprehensive multiword lexicons from large corpora based on competition among $n$-gram candidates. In contrast to the standard approach of simple ranking by association measure, in our model $n$-grams are arranged in a lattice structure based on subsumption and overlap relationships, with nodes inhibiting other nodes in their vicinity when they are selected as a lexical item. We show how the configuration of such a lattice can be optimized tractably, and demonstrate using annotations of sampled $n$-grams that our method consistently outperforms alternatives by at least 0.05 F-score across several corpora and languages.

## 1 Introduction

Despite over 25 years of research in computational linguistics aimed at acquiring multiword lexicons using corpora statistics, and growing evidence that speakers process language primarily in terms of memorized sequences (Wray, 2008), the individual word nonetheless stubbornly remains the *de facto* standard processing unit for most research in modern NLP. The potential of multiword knowledge to improve both the automatic processing of language as well as offer new understanding of human acquisition and usage of language is the primary motivator of this work. Here, we present an effective, expandable, and above all tractable new approach to comprehensive multiword lexicon acquisition. Our aim is to find a middle ground between standard MWE acquisition approaches based on association measures (Ramisch, 2014) and more sophisticated statistical models (Newman et al., 2012) that do not scale to the large corpora, the main source of the distributional information in modern NLP systems.

A central challenge in building comprehensive multiword lexicons is paring down the huge space of possibilities without imposing restrictions which disregard a major portion of the multiword vocabulary of a language: allowing for diversity creates significant redundancy among statistically promising candidates. The lattice model proposed here addresses this primarily by having the candidates—contiguous and non-contiguous $n$-gram types—compete with each other based on subsumption and overlap relations to be selected as the best (i.e., most parsimonious) explanation for statistical irregularities due to lexical affinity. We test this approach across four large corpora in three languages, including two relatively free-word-order languages (Croatian and Japanese), and find that this approach consistency outperforms alternatives, offering scalability and many avenues for future enhancement.

## 2 Background and Related Work

In this paper we will refer to the targets of our lexicon creation efforts as **formulaic sequences**, following the terminology of Wray (2002; 2008), wherein a formulaic sequence (FS) is defined 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." That is, a FS shows signs of being part of a mental lexicon.[1] As noted by Wray (2008), formulaic sequence theory is compatible with other highly multiword, lexicalized approaches to language structure, in particular Pattern Grammar (Hunston and Francis, 2000) and Construction Grammar (Goldberg, 1995); an important distinction, though, is that these sorts of theories often posit entirely abstract grammatical constructions/patterns/frames which do not fit well into the FS framework. Nevertheless, since many such constructions *are* composed of sequences of specific words, the FS inventory of a language includes many flexible constructions (e.g., *ask $*$ for*) along with entirely fixed combinations (e.g., *rhetorical question*) not typically of interest to grammarians. Note that the FS framework allows for individual morphemes be part of a formulaic sequence, but for practical reasons we focus primarily on lemmatized words as the unit out of which FS are built.

In computational linguistics, the most common term used to describe multiword lexical units is *multiword expression* ("MWE": Sag et al. (2002), Baldwin and Kim (2010)), but here we wish to make a principled distinction between at least somewhat non-compositional, strongly lexicalized MWEs and FS, a near superset which includes many MWEs but also compositional linguistic formulas. This distinction is not a new one; it exists, for example, in the original paper of Sag et al. (2002) in the distinction between lexicalized and institutionalized phrases, and also to some extent in the MWE annotation of Schneider et al. (2014b), who distinguish between weak (collocational)[2] and strong (non-compositional) MWEs. It is our contention, however, that separate, precise terminology is useful for research targeted at either class: we need not strain the concept of MWE to include items which do not require special semantics, nor are we inclined to disregard the larger formulaticity of language simply because it is not the dominant focus of MWE research. Many MWE researchers might defensibly balk at including in their MWE lexicons and corpus annotations (English) FS such as *there is something going on*, *it is more important than ever to ...*, *... do not know what it is like to ...*, *there is no shortage of ...*, *the rise and fall of ...*, *now is not the time to ...*, etc. as well as tens of thousands of other such phrases which, along with less compositional MWEs like *be worth ...'s weight in gold*, fall under the FS umbrella. Another reason to introduce a different terminology is that there are classes of phrases which are typically considered MWEs that do not fit well into an FS framework, for instance novel compound nouns whose semantics are accessible by analogy (e.g., *glass limb*, analogous to *wooden leg*). Also, we exclude from the definition of both FS and MWE those named entities which refer to people or places which are little-known and/or whose surface form appears derived (e.g., *Mrs. Barbara W. Smith* or *Smith Garden Supplies Ltd*). Figure 1 shows the conception of the relationship between FS, (multiword) constructions, MWE, and (multiword) named entities that we assume for this paper.

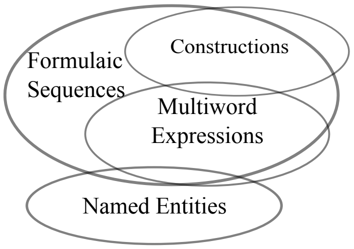

Figure 1: Multiword Terminology

Regardless of the terminology used to describe them, the starting point for multiword lexicon creation has typically been lexical association measures (Church and Hanks, 1990; Dunning, 1993; Schone and Jurafsky, 2001; Evert, 2004; Pecina, 2010; Araujo et al., 2011; Kulkarni and Finlayson, 2011; Ramisch, 2014). When these methods are used to build a lexicon, particular binary syntactic patterns are typically chosen. Only some of these measures generalize tractably beyond two words, for example PMI (Church and Hanks, 1990), i.e., the log ratio of the joint probability to the product of the marginal probabilities of the individual words. Another measure which addresses sequences of longer than two words is the $c$-value (Frantzi et al., 2000) which

[1] Though by this definition individuals or small groups may have their own FS, here we are only interested in FS that are shared by a recognizable language community.

[2] Here we avoid the term *collocation* entirely due to confusion with respect to its interpretation. Though some define it similarly to our definition of FS, it can be applied to any words that show a statistical tendency to appear in the vicinity of one another for any reason: for instance, the pair of words *doctor/nurse* might be considered a collocation (Ramisch, 2014).

weights term frequency by the log length of the $n$-gram while penalizing $n$-grams that appear in frequent larger ones. Mutual expectation (Dias et al., 1999) involves deriving a normalized statistic that reflects the extent to which a phrase resists the omission of any constituent word. Similarly, the lexical predictability ratio (LPR) of Brooke et al. (2015) is an association measure intended for any possible syntactic pattern which is calculated by discounting syntactic predictability from the overall conditional probability for each word given the other words in the phrase. Though most association measures involve only usage statistics of the phrase and its subparts, the DRUID measure (Riedl and Biemann, 2015) is an exception which uses distributional semantics around the phrase to identify how easily an $n$-gram could be replaced by a single word.

Typically multiword lexicons are created by ranking $n$-grams according to an association measure and applying a threshold. The algorithm of da Silva and Lopes (1999) is somewhat more sophisticated, in that it identifies the local maxima of association measures across subsuming $n$-grams within a sentence to identify MWEs of unrestricted length and syntactic composition; its effectiveness beyond noun phrases, however, seems relatively limited (Ramisch et al., 2012). Brooke et al. (2014; 2015) developed a heuristic method intended for general FS extraction in larger corpora, first using conditional probabilities to do an initial (single pass) coarse-grained segmentation of the corpus, followed by a pass through the resulting vocabulary, breaking larger units into smaller ones based on a tradeoff between marginal and conditional statistics. The work of Newman et al. (2012) is an example of an unsupervised approach which does not use association measures: it extends the Bayesian word segmentation approach of Goldwater et al. (2009) to multiword tokenization, applying a generative Dirichlet Process model which jointly constructs a segmentation of the corpus and a corresponding multiword vocabulary.

Other research in MWEs has tended to be rather focused on particular syntactic patterns such as verb-noun combinations (Fazly et al., 2009). The system of Schneider et al. (2014a) distinguishes a full range of MWE sequences in the English Web Treebank, including gapped expressions, using a supervised sequence tagging model. Though in theory automatic lexical resources could be a useful addition to the Schneider et al. model, which uses only manual lexical resources, attempts to do so have achieved mixed success (Riedl and Biemann, 2016).

The motivation for building lexicons of FS naturally overlaps with those for MWE: models of distributional semantics, in particular, can benefit from sensitivity to multiword units (Cohen and Widdows, 2009), as can parsing (Constant and Nivre, 2016) and topic models (Lau et al., 2013). One major motivation for looking beyond MWEs is the ability to carry out broader linguistic analyses. Within corpus linguistics, multiword sequences have been studied in the form of *lexical bundles* (Biber et al., 2004), which are simply $n$-grams that occur above a certain frequency threshold. Like FS, lexical bundles generally involve larger phrasal chunks that would be missed by traditional MWE extraction, and so research in this area has tended to focus on how particular formulaic phrases (e.g., *if you look at*) are indicative of particular genres (e.g., university lectures). Lexical bundles have been applied, in particular, to learner language: for example, Chen and Baker (2010) show that non-native student writers use a severely restricted range of lexical bundle types, and tend to overuse those types, while Granger and Bestgen (2014) investigate the role of proficiency, demonstrating that intermediate learners underuse lower-frequency bigrams and overuse high-frequency bigrams relative to advanced learners. Sakaguchi et al. (2016) demonstrate that improving fluency (closely linked to the use of linguistic formulas) is more important than improving strict grammaticality with respect to native speaker judgments of non-native productions; Brooke et al. (2015) explicitly argue for FS lexicons as a way to identify, track, and improve learner proficiency.

## 3 Method

Our approach to FS identification involves optimization of the total explanatory power of a lattice, where each node corresponds to an $n$-gram type. The explanatory power of the whole lattice is defined simply as a product of the **explainedness** of the individual nodes. Each node can be considered either "on" (*is an FS*) or "off" (*is not an FS*). The basis of the calculation of explainedness is the syntax-sensitive

LPR association measure of Brooke et al. (2015), but it is calculated differently depending on the on/off status of the node as well as the status of the nodes in its vicinity. Nodes are linked based on $n$-gram subsumption and corpus overlap relationships (see Figure 2), with "on" nodes typically explaining other nodes. Given these relationships, we iterate over the nodes and greedily optimize the on/off choice relative to explainedness in the local neighborhood of each node, until convergence.

### 3.1 Collecting statistics

The first step in the process is to derive a set of $n$-grams and related statistics from a large, unlabeled corpus of text. Since our primary association measure is an adaption of LPR, our approach in this section mostly follows Brooke et al. (2015) up until the last stage. An initial requirement of any such method is an $n$-gram frequency threshold, which we set to 1 instance per 10 million words, following Brooke et al. (2015).[3]

We include gapped or non-contiguous $n$-grams in our analysis, in acknowledgment of the fact that many languages have MWEs where the components can be "separated", including verb particle constructions in English (Dehé, 2002), and noun-verb idioms in Japanese (Hashimoto and Kawahara, 2008). Having said this, there are generally strong syntactic and length restrictions on what can constitute a gap (Wasow, 2002), which we capture in the form of a language-specific POS-based regular expression (see Section 4 for details). This greatly lowers the number of potentially gapped $n$-gram types, increasing precision and efficiency for negligible loss of recall. We also exclude punctuation and lemmatize the corpus, and enforce an $n$-gram count threshold. As long as the count threshold is substantially above 1, efficient extraction of all $n$-grams can be done iteratively: in iteration $i$, $i$-grams are filtered by the frequency threshold, and then pairs of instances of these $i$-grams with $(i-1)$ words of overlap are found, which derives a set of $(i+1)$-grams which necessarily includes all those over the frequency threshold.

Once a set of relevant $n$-grams is identified and counted, other statistics required to calculate the **Lexical Predictability Ratio** ("LPR") for each word in the $n$-gram are collected. LPR is a measure of how predictable a word is in a lexical context, as compared to how predictable it is given only syntactic context (over the same span of words). Formally, the LPR for word $w_i$ in the context of a word sequence $w_1, ..., w_i, ..., w_n$ with POS tag sequence $t_1, ..., t_n$ is given by:

$$\text{LPR}(w_i, w_{1,n}) = \max_{1 \le j < k \le n} \frac{p(w_i|w_{j,k})}{p(w_i|t_{j,k})}$$

where $w_{j,k}$ denotes the word sequence $w_j, ..., w_{i-1}, w_{i+1}, ..., w_k$ excluding $w_i$ (similarly for $t_{j,k}$). Note that the lower bound of LPR is 1, since the ratio for a word with no context is trivially 1. We use the same equation for gapped $n$-grams, with the caveat that quantities involving sequences which include the location where the gap occurs are derived from special gapped $n$-gram statistics. Note that the identification of the best ratio across all possible choices of context, not just the largest, is important for longer FS, where the entire POS context alone might uniquely identify the phrase, resulting in the minimum LPR of 1 even for entirely formulaic sequences—an undesirable result.

In the segmentation approach of Brooke et al. (2015), LPR for an entire span is calculated as a product of the individual LPRs, but here we will use the minimum LPR across the words in the sequence:

$$\text{minLPR}(w_{1,n}) = \min_{1 \le i \le n} \text{LPR}(w_i, w_{1,n})$$

Here, minLPR for a particular $n$-gram does not reflect the *overall* degree to which it holds together, but rather focuses on the word which is its weakest link. For example, in the case of *be keep $*$ under wraps* (Figure 2), a general statistical metric might assign it a high score due to the strong association between *keep* and *under* or *under* and *wraps*, but minLPR is focused on the weaker relationship between *be* and the rest of the phrase. This makes it particularly suited to use in a lattice model of competing $n$-grams, where the choice of *be keep $*$ under wraps* versus *keep $*$ under wraps* should be based

[3] Based on manual analysis using the MWE corpus of Schneider et al. (2014b), this achieves very good (over 90%) type-level MWE coverage using the frequency filtered $n$-gram statistics from the ICWSM blog corpus (see Section 4) after filtering out proper names.

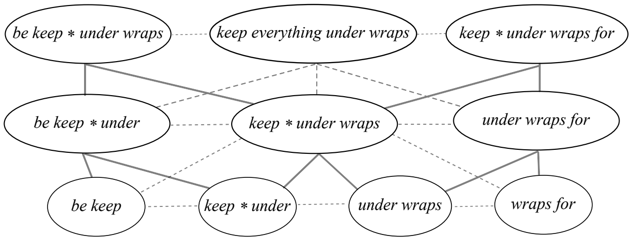

Figure 2: A portion of an $n$-gram lattice. Solid lines indicate subsumption, dotted lines overlaps

exactly on the extent to which *be* is an essential part of the phrase; the other affinities are, in effect, irrelevant, because they occur in the smaller $n$-gram as well.

### 3.2 Node interactions

The $n$-gram nodes in the lattice are directionally connected to nodes consisting of $(n + 1)$-grams which subsume them and $(n - 1)$-grams which they subsume. For example, as detailed in Figure 2, the (gapped) $n$-gram *keep* $*$ *under wraps* would be connected "upwards" to the node *keep everything under wraps* and connected "downwards" to *under wraps*. These directional relationships allow for two basic interactions between nodes in the lattice when a node is turned on: **covering**, which inhibits nodes below (subsumed by) a turned-on node (e.g., if *keep* $*$ *under wraps* is on, the model will tend not to choose *under wraps* as an FS); and **clearing**, which inhibits nodes above a turned-on node (e.g., if *keep* $*$ *under wraps* is on, the model would avoid selecting *keep everything under wraps* as an FS). A third, undirected mechanism is **overlapping**, where nodes inhibit each other due to overlaps in the corpus (e.g., having both *keep* $*$ *under wraps* and *be keep* $*$ *under* as FS will be avoided).

#### 3.2.1 Covering

The most important node interaction is **covering**, which corresponds to discounting or entirely excluding a node due to a node higher in the lattice. Our model includes two types of covering: hard and soft.

**Hard covering** is based on the idea that, due to very similar counts, we can reasonably conclude that the presence of an $n$-gram in our statistics is a direct result of a subsuming $(n+i)$-gram. In Figure 2, e.g., if we have 143 counts of *keep* $*$ *under wraps* and 152 counts of *under wraps*, the presence of *keep* $*$ *under wraps* almost completely explains *under wraps*, and we should consider these two $n$-grams as one. We do this by permanently disabling any hard covered node, and setting the minLPR of the covering node to the maximum minLPR among all the nodes it covers (including itself); this means that longer $n$-grams with function words (which often have lower minLPR) can benefit from the strong statistical relationships between open-class lexical features in $n$-grams that they cover. This is done as a preprocessing step, and greatly improves the tractability of the iterative optimization of the lattice. Of course, a threshold for hard covering must be chosen: during development we found that a ratio of $2/3$ (corresponding to a significant majority of the counts of a lower node corresponding to the higher node) worked well. We also use the concept of hard covering to address the issue of pronouns, based on the observation that specific pronouns often have high LPR values due to pragmatic biases (Brooke et al., 2015); for instance, private state verbs like *feel* tend to have first person singular subjects. In the lattice, $n$-grams with pronouns are considered covered (inactive) unless they cover at least one other node which does not have a pronoun, which allows us to limit FS with pronouns without excluding them entirely: they are included only in cases where they are definitively formulaic.

**Soft covering** is used in cases when a single $n$-gram does not entirely account for another, but a turned-on $n$-gram to some extent may explain some of the statistical irregularity of one lower in the lattice. For instance, in Figure 2 *keep* $*$ *under* is not hard-covered by *keep* $*$ *under wraps* (since there are FS such as *keep* $*$ *under surveillance*, *keep it under your hat*, etc. ), but if *keep* $*$ *under wraps* is tagged as an FS, we nevertheless want to discount the portion of the *keep* $*$ *under* counts that correspond to occurrences of *keep* $*$ *under wraps*, with the idea that these occurrences have already been explained by the longer $n$-gram. If enough subsuming $n$-grams are on, then the shorter $n$-gram will be discounted to the extent that it will be turned off, preventing redundancy. This effect is accomplished by increasing the turned-off explainedness of *keep* $*$ *under* (and thus making turning on less desirable) in the following manner: let $c(\cdot)$ be the count function, and $a_1, ..., a_m$

be any turned-on nodes which are above $t$ in the lattice (covering nodes). Then, the cover$(t)$ score for a covered node $t$ is:

$$\text{cover}(t) = \max\left(0, \frac{c(t) - \sum_{i=1}^{m} c(a_i)}{c(t)}\right)$$

When applied as an exponent to a minLPR score, it serves as simple, quick-to-calculate approximation of a new minLPR with the counts corresponding to the covering nodes removed from the calculation. The cover score takes on values in the range 0 to 1, with 1 being the default when no covering occurs.

#### 3.2.2 Clearing

In general, covering prefers turning on longer, covering $n$-grams since doing so explains nodes lower in the lattice. Not surprisingly, it is generally desirable to have a mechanism working in opposition, i.e., one which views shorter FS as helping to explain the presence of longer $n$-grams which contain them, beyond the FS-neutral syntactic explanation provided by minLPR. **Clearing** does this by increasing the explainedness of nodes higher in the lattice when a lower node is turned-on. The basic mechanism is similar to covering, except that counts cannot be made use of in the same way—whereas it makes sense to explain covered nodes in proportion to the counts of their covering nodes (since the counts of the covered $n$-grams can be directly attributed to the covering $n$-gram), in the reverse direction this logic fails.

A simple but effective solution which avoids extra hyperparameters is to make use of the minLPR values of the relevant nodes. In the most common two-node situation, we increase the explainedness of the cleared node based on the ratio of the minLPR of two nodes, though only if the minLPR of the lower node is higher. Generalized to the (rare) case of multiple clearing nodes, we define clear$(t)$ as:

$$\text{clear}(t) = \prod_{i=1}^{m} \min\left(1, \frac{\text{minLPR}(t)}{\text{minLPR}(b_i)}\right)$$

where $b_i$ is the $i$th clearing node of $t$, i.e., turned-on nodes below $t$ in the lattice. We refer to this mechanism as "clearing" because it tends to clear away a variety of trivial uses of common FS which may have higher LPR due to the lexical and syntactic specificity of the FS. For instance, in Figure 2 if the node *keep* $*$ *under wraps* is turned on and has a minLPR of 8, then, if the minLPR of a node such as *keep* $*$ *under wraps for* is 4, clear$(t)$ will be 0.5. Like cover, clear takes on values in the range 0 to 1, with 1 being the default when no clearing occurs. Note that one major advantage with this particular formulation of clearing is that low-LPR nodes will be unable to clear higher LPR nodes above them in the lattice; otherwise, bad FS like *of the* might be selected as FS based purely to increase the explainedness of the many $n$-grams they appear in.

#### 3.2.3 Overlap

The third mechanism of node interaction involves $n$-grams which overlap in the corpus. In general, independent FS do not consistently overlap. For example, given that *be keep* $*$ *under* and *keep* $*$ *under wraps* often appear together (overlapping on the tokens *keep* $*$ *under*), we do not want both being selected as an FS, even in the case that both have high minLPR. To address this problem, rather than increasing the explainedness of turned-off nodes, we decrease the explainedness of the overlapping turned-on nodes—a penalty rather than an incentive which expresses the model's confusion at having overlapping FS. Let $oc(x, y)$ be the number of times the $n$-grams corresponding to nodes $x$ and $y$ overlapped in the corpus, and $o_1, ..., o_m$ refer to overlapping nodes, i.e., turned-on nodes which overlap with target node $t$ in the corpus. We define overlap$(t)$ as:

$$\text{overlap}(t) = \frac{c(t)}{c(t) - \sum_{i=1}^{m} oc(t, o_i)}$$

Overlap takes on values in the range 1 to $+\infty$, also defaulting to 1 when no overlaps exist. The effect of overlap is hyperbolic: small amounts of overlap have little effect, but nodes with significant overlap will effectively be forced to turn off.

### 3.3 Explainedness

In order to capture the current state of the model and provide an overall explainedness score, we introduce two indicator functions: $I_{on}(t)$ is 1 if node $t$ is on and 0 if $t$ is off ; $I_{\mathit{off}}(t)$ is the reverse. The objective function maximized by the model is then the explainedness (*expl*) across all the nodes of the lattice, $t_1, \ldots, t_{|T|} \in T$, which can be defined in terms of minLPR and the node interaction functions:

$$\begin{aligned} expl(T) = \prod_{i=1}^{|T|} & I_{on}(t_i) \cdot C^{-\text{overlap}(t_1)} \\ & + I_{off}(t_i) \cdot \text{minLPR}(t_i)^{-\text{cover}(t_i) \cdot \text{clear}(t_i)} \end{aligned} \quad (1)$$

When a node is off, its explainedness is the inverse of its minLPR, except if there are covering or clearing nodes which explain it by pushing the exponent of minLPR towards zero. When the node is on, its explainedness is the inverse of a fixed cost hyperparameter $C$, though this cost is increased if it overlaps with other active nodes. All else being equal, when minLPR$(t) > C$, a node will be selected as an FS, and so, independent of the node interactions, $C$ can be viewed as the threshold for the minLPR association measure under a traditional approach to MWE identification. There is no upper bound on $C$, but empirically, we have found values in the range $[3, 6]$ give reasonable results for the languages presented in this paper.

### 3.4 Optimization

The dependence of the explainedness of nodes on their neighbors effectively prohibits a global optimization of the lattice. Fortunately, though most of the nodes in the lattice are part of a single connected graph, most of the effects of nodes on each other are relatively local, and effective local optimizations can be made tractable by applying some simple restrictions. The main optimization loop consists of iterations over the lattice until complete convergence (no changes in the final iteration). For each iteration over the main loop, each potentially active node is examined in order to evaluate whether its current status is optimal given the current state of the lattice. The order that we perform this has an effect on the result: among the obvious options, good results are obtained through ordering nodes by frequency, which gives an implicit advantage to relatively common $n$-grams.

**Algorithm 1** Optimization algorithm. $T$ is an ordered list of the nodes in the lattice. Nodes (designated by $t$) contain pointers to the nodes immediately linked to them in the lattice. States (designated by $s$) indicate whether each node is ON or OFF. Explainedness values are indicated by $e$.

```
function LocalOpt($s_{start}, t, T_{rev}, T_{aff}$)
    $s_{start} \leftarrow$ Set($s, t$, ON)
    $Q \leftarrow$ EmptyQueue()
    $e_{best} \leftarrow 0$
    $s_{best} \leftarrow$ NULL
    Push($Q, s_{start}$)
    repeat
        $s_{curr} \leftarrow$ Pop($Q$)
        $e_{curr} \leftarrow$ CalcExpl($s_{curr}, T_{aff}$)
        for $t_{rev}$ in $T_{rev}$ do
            $s_{new} \leftarrow$ Set($s_{curr}, t_{rev}$, OFF)
            $e_{new} \leftarrow$ CalcExpl($s_{new}, T_{aff}$)
            if $e_{new} > e_{curr}$ then
                Push($Q, s_{new}$)
                if $e_{new} > e_{best}$ then:
                    $e_{best} \leftarrow e_{new}$
                    $s_{best} \leftarrow s_{new}$
    until IsEmpty($Q$)
    return $s_{best}$

FrequencySortReverse($T$)
$s \leftarrow$ InitializeAllOff($T$)
repeat
    $Changed \leftarrow$ FALSE
    for $t$ in $T$ do
        $T_{rev} \leftarrow$ GetRelevant($t, T$)
        $T_{aff} \leftarrow$ GetAffected($T_{rev}, T$)
        $s_{new} \leftarrow$ LocalOpt($s, t, T_{rev}, T_{aff}$)
        if $s_{new} \neq s$ then
            $s \leftarrow s_{new}$
            $Changed \leftarrow$ TRUE
until !$Changed$
```

Given the relationships between nodes, it is obviously not sufficient to consider switching only the present node. If, for instance, one or more of *be keep * under wraps*, *under wraps*, or *be keep * under* has been turned on, the covering, clearing, or overlapping effects of these other nodes will likely prevent a competing node like *keep * under wraps* from being correctly activated. Instead, the algorithm identifies a small set of "relevant" nodes which are the most important to the status of the node under consideration. Since turned-off nodes have no direct effect on each other, only turned-on nodes above, below, or overlapping with the current node in the lattice need be considered. Once the relevant nodes have been identified, all nodes (including turned-off nodes) whose explainedness is affected by one or more of the relevant nodes are identified. Next, a search is carried out for the optimal configuration of the relevant nodes, starting from an 'all-on' state and iteratively considering new states with one relevant node turned off; the search continues as long as there is an improvement in explainedness. Since the node interactions are roughly cumulative in their effects, this approach will generally identify the optimal state without the need for an exhaustive search. See Algorithm 1 for details.

Omitted from Algorithm 1 for clarity are various low-level efficiencies which prevent the algorithm from reconsidering states already checked or from recalculating the explainedness of nodes when unnecessary. We also apply the following efficiency restrictions, which significantly reduce the runtime of the algorithm. In each case, more extreme (less efficient) values were individually tested using a development set and found to provide no benefit in terms of the quality of the output lexicon:

- We limit the total number of relevant nodes to 5. When there are more than 5 nodes turned on in the vicinity of the target node, the most relevant nodes are selected by ranking candidates by the absolute difference in explainedness across possible configurations of the target and candidate node considered in isolation;
- To avoid having to deal with storing and processing trivial overlaps, we exclude overlaps with a count of less than 5 from our lattice;
- Many nodes have a minLPR which is slightly larger than 1 (the lowest possible value). There is very little chance these nodes will be activated by the algorithm, and so after applying hard covering, we do not consider activating nodes with minLPR $< 2$.

## 4 Evaluation

We evaluate our approach across three different languages including evaluation sets derived from four different corpora. In English, we follow Brooke et al. (2015) in using a 890M token filtered portion of the ICWSM blog corpus (Burton et al., 2009) tagged with the Tree Tagger (Schmid, 1995). To facilitate a comparison with Newman et al. (2012), which does not scale up to a corpus as large as the ICWSM, we also build a lexicon using the 100M token British National Corpus (Burnard, 2000), using the standard CLAWS-derived POS tags for the corpus. Lemmatization included removing all inflectional marking from both words and POS tags. For English, gaps are identified using the same POS regex used in Brooke et al. (2015), which includes simple nouns and portions thereof, up to a maximum of 4 words.

The other two languages we include in our evaluation are Croatian and Japanese. Relative to English, both languages have freer word order: we were interested in probing the challenges associated with using an $n$-gram approach to FS identification in such languages. For Croatian, we used the fhrWaC corpus (Šnajder et al., 2013), a filtered version of the Croatian web corpus hrWaC (Ljubešić and Klubička, 2014), which is POS-tagged and lemmatized using the tools of Agić et al. (2013). Similar to English, the POS regex for Croatian includes simple nouns, adjectives and pronouns, but also other elements that regularly appear inside FS, including both adverbs and copulas. For Japanese, we used a subset of the 100M-page web corpus of Shinzato et al. (2008), which was roughly the same size (in terms of token count) as the English corpus. We segmented and POS-tagged the corpus with MeCab (Kudo, 2008) using the UNIDIC morphological dictionary (Den et al., 2007). The POS regex for Japanese covers the same basic nominal structures as English, but also includes case markers and adverbials. Though our processing of Japanese includes basic lemmatization related to superficial elements like the choice of writing script and politeness markers, many elements (such as case marking) which are removed by lemmatization in Croatian are segmented into independent morphological units in the MeCab output, making the task somewhat different for the two languages.

Brooke et al. (2015) introduced a method for evaluating FS extraction without a reference lexicon or direct annotation of the output of a model. Instead, $n$-grams are sampled after applying the frequency threshold and then annotated as being either an FS or not. Benefits of this style of evaluation include replicability, the diversity of FS, and the ability to

|  | Contiguous |  | Gapped |  | $\kappa$ |  |
|---|---|---|---|---|---|---|
|  | FS | non-FS | FS | non-FS | Pre | Post |
| ICWSM | 169 | 702 | 29 | 916 | 0.52 | 0.84 |
| BNC | 49 | 403 | 8 | 475 | 0.51 | 0.84 |
| Croatian | 64 | 382 | 11 | 456 | 0.58 | 0.87 |
| Japanese | 102 | 337 | 9 | 438 | 0.49 | 0.82 |

Table 1: Statistics for test sets

calculate a true F-score. We use the annotation of 2000 $n$-grams in the ICWSM corpus from that earlier work, and applied the same annotation methodology to the other three corpora: after training and based on written guidelines derived from the definitions of Wray (2008), three native-speaker, educated annotators judged 500 contiguous $n$-grams and another 500 gapped $n$-grams for each corpus.

Other than the inclusion of new languages, our test sets differ from Brooke et al. (2015) in two ways. One advantage of a type-based annotation approach, particularly with regards to annotation with a known subjective component, is that it is quite sensible to simply discard borderline cases, improving reliability at the cost of some representativeness. To this end, we entirely excluded from our test set $n$-grams which just one annotator marked as FS. Table 1 contains the counts for the four test sets after this filtering step as well as Fleiss' Kappa scores before ("Pre") and after ("Post"). The second change is that for the main evaluation we collapsed gapped and contiguous $n$-grams into a single test set. The rationale is that the number of positive gapped examples is too low to provide a reliable independent F-score.

Our primary comparison is with the heuristic LPR model of Brooke et al. (2015), which is scalable to large corpora and includes gapped $n$-grams. For the BNC, we also benchmark against the DP-seg model of Newman et al. (2012) with recommended settings, and the LocalMaxs algorithm of da Silva and Lopes (1999) using SCP; neither of these methods scale to the larger corpora.[4] Because these other approaches only generate sequential multiword units, we use only the sequential part of the BNC test set for this evaluation. All comparison approaches have themselves been previously compared against a wide range of association measures. As such, we do not repeat all these comparisons here, but we do consider a lexicon built from ranking $n$-grams according to the measure used in our lattice (minLPR) as well as PMI and raw frequency. For each of these association measures we rank all $n$-grams above the frequency threshold and build a lexicon equal to the size of the lexicon produced by our model.

We created small development sets for each corpus and used them to do a thorough testing of parameter settings. Although it is generally possible to increase precision by increasing $C$, we found that across corpora we always obtained near-optimal results with $C = 4$, so to demonstrate the usefulness of the lattice technique as an entirely off-the-shelf tool, we present the results using identical settings for all four corpora. We treat covering as a fundamental part of the Lattice model, but to investigate the efficacy of other node interactions within the model we present results with overlap and clearing node interactions turned off.

## 5 Results

The main results for FS acquisition across the four corpora are shown in Table 2. As noted in Section 2, simple statistical association measures like PMI do poorly when faced with syntactically-unrestricted $n$-grams of variable length: minLPR is clearly a much better statistic for this purpose. The LPRseg method of Brooke et al. (2015) consistently outperforms simple ranking, and the lattice method proposed here does better still, with a margin that is fairly consistent across the languages. Generally, clearing and overlap node interactions provide a relatively large increase in precision at the cost of a smaller drop in recall, though the change is fairly symmetrical in Croatian. When only covering is used, the results are fairly similar to Brooke et al. (2015), which is unsurprising given the extent to which decomposition and covering are related. The Japanese and ICWSM corpora have relatively high precision and low recall, whereas both the BNC and Croatian corpora have low precision and high recall.

In the contiguous FS test set for the BNC (Ta-

[4] DP-seg is far too slow, and LocalMaxs, though faster, calculates counts for all $n$-grams in the corpus, which would require terabytes of RAM for the large corpora.

| | English | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ICWSM | | | BNC | | | Croatian | | | Japanese | | |
| | P | R | F | P | R | F | P | R | F | P | R | F |
| Countrank | 0.13 | 0.09 | 0.10 | 0.08 | 0.14 | 0.10 | 0.10 | 0.16 | 0.12 | 0.11 | 0.06 | 0.08 |
| PMIrank | 0.24 | 0.14 | 0.18 | 0.12 | 0.25 | 0.16 | 0.21 | 0.32 | 0.26 | 0.18 | 0.08 | 0.11 |
| minLPRrank | 0.49 | 0.31 | 0.38 | 0.25 | 0.44 | 0.32 | 0.36 | 0.45 | 0.40 | 0.47 | 0.21 | 0.29 |
| LPR-seg | 0.53 | 0.42 | 0.47 | 0.37 | 0.44 | 0.40 | 0.41 | 0.47 | 0.43 | 0.69 | 0.43 | 0.53 |
| Lattice $-cl,ovr$ | 0.47 | **0.53** | 0.50 | 0.31 | **0.60** | 0.41 | 0.32 | 0.66 | 0.43 | 0.49 | 0.61 | 0.54 |
| Lattice $-cl$ | 0.57 | 0.42 | 0.49 | 0.33 | 0.58 | 0.42 | 0.39 | 0.56 | 0.46 | 0.63 | 0.49 | 0.55 |
| Lattice $-ovr$ | 0.52 | **0.53** | 0.53 | 0.34 | **0.60** | 0.44 | 0.36 | **0.67** | 0.47 | 0.53 | **0.62** | 0.58 |
| Lattice | **0.67** | 0.44 | **0.54** | **0.38** | 0.58 | **0.46** | **0.44** | 0.56 | **0.49** | **0.76** | 0.49 | **0.59** |

Table 2: Results of FS identification in various test sets: Countrank = ranking with frequency; PMIrank = PMI-based ranking; minLPRrank = ranking with minLPR; LPRseg = the method of Brooke et al. (2015); “$-$cl” = no clearing; “$-$ovr” = no penalization of overlaps; “P” = Precision; “R” = Recall; and “F” = F-score. Bold is best in a given column. For ICWSM only, F-score difference relative to best baseline is significant at $p < 0.05$; the difference for all test sets together is significant at $p < 0.01$; all p-values based on the permutation test (Yeh, 2000).

| | P | R | F |
|---|---|---|---|
| PMIrank | 0.20 | 0.29 | 0.23 |
| minLPRrank | 0.34 | 0.45 | 0.39 |
| LPR-seg | 0.42 | 0.45 | 0.43 |
| LocalMaxs | **0.56** | 0.39 | 0.46 |
| DP-seg | 0.35 | **0.71** | 0.47 |
| Lattice | 0.46 | 0.61 | **0.53** |

Table 3: Results of FS identification in contiguous BNC test set; LocalMaxs = method of da Silva and Lopes (1999); DP-seg = method of Newman et al. (2012)

ble 3), we found that both the LocalMaxs algorithm and the DP-seg method of Newman et al. (2012) were able to beat our other baseline methods with roughly similar F-scores, though both are well below our Lattice method. Some of the difference seems attributable to fairly severe precision/recall imbalance, though we were unable to improve the F-score by changing the parameters from recommended settings for either model.

## 6 Discussion

Though the results across the four corpora are reasonably similar with respect to overall F-score, there are some discrepancies. By using the standard UNIDIC morpheme representation as the base unit for Japanese, the model ends up doing an extra layer of FS identification, one which is provided by word boundaries in the other languages. The result is that there are relatively more FS for Japanese: precision is high, and recall is comparably low. Importantly, the initial $n$-gram statistics actually reflect that Japanese is different: the number of $n$-gram types over length 4 is almost twice the number in the ICWSM corpus. One idea for future work is to automatically adapt to the input language/corpus in order to ensure a good balance between precision and recall.

At the opposite extreme, the low precision of the BNC is almost certainly due to its relatively small size: whereas the $n$-gram threshold we used here results in minimum counts of roughly 100 for the other three corpora, the BNC statistics include $n$-grams with counts less than 10. At such low counts, LPR is less reliable and more noise gets into the lexicon: the first column of Table 4 shows that the BNC is noticeably larger then the other lexicons, and the higher numbers in columns 2 and 3 (number of POS types and percentage of gapped expressions, resp.) are also indicative of increased noise. This could be resolved by increasing the $n$-gram threshold. It might also make sense to simply avoid smaller corpora, though for some applications a smaller corpus may be unavoidable. One idea we are pursing is mo-

| | | | | By Length (%) | | | |
|---|---|---|---|---|---|---|---|
| **Lexicon** | **Word types** | **POS types** | **Gapped (%)** | 2 | 3 | 4 | 5+ |
| English ICWSM | 202k | 16.8k | 18.3 | 46.7 | 30.1 | 15.0 | 7.5 |
| English BNC | 374k | 30.0k | 26.1 | 38.8 | 36.6 | 18.1 | 7.7 |
| Croatian | 172k | 6.7k | 7.9 | 55.0 | 31.4 | 10.3 | 3.2 |
| Japanese | 218k | 29.4k | 7.9 | 39.5 | 31.1 | 16.8 | 12.6 |

Table 4: Statistics for the lexicons created by our lattice method

difying the calculation of the LPR metric to use a more conservative probability estimate than maximum likelihood in the case of low counts.

We were interested in Croatian and Japanese in part because of their relatively free word order, and whether the handling of gaps would help with identifying FS in these languages. We discovered, however, that free word order actually results in *more* of a tendency towards contiguous FS, not less, a fact that is reflected in our test sets (Table 1) as well as the lexicons themselves (Table 4). Strikingly rare in Croatian, in particular, are expressions where the content of a gap is an argument which must be filled to syntactically complete an expression: it is English whose fixed-word-order constraints often keep elements of an FS distant from each other. The gaps that do happen in Croatian are mostly prosody-driven insertions of other elements into already complete FS. This phenomena highlights a problem with the current model, in that gapped and contiguous versions of the same $n$-gram sequence (e.g., *take away* and *take ∗ away*) are, at present, considered entirely independently. Alternatives for dealing with this include collapsing statistics to create a single node in the lattice, creating a promoting link between contiguous and gapped versions of the same $n$-grams sequence in the lattice model, or switching to a dependency representation (which, we note, requires very little change to the basic model presented here, but would narrow its applicability).

The statistics in Table 4 otherwise reflect the quantity and diversity of FS across the corpora, particularly in terms of the number of POS patterns represented in the lexicon. Looking at the most common POS patterns across languages, only noun-noun and adjective-noun combinations ever account for more than 5% of all word types in any of the lexicons. Though some of the diversity can of course be attributed to noise, it is safe to say that most FS do not fall into the standard two-word syntactic categories used in MWE work, and therefore identifying them requires a much more general approach like the one presented here.

Table 5 contains 10 randomly selected examples from each of the lexicons produced by our method. Among the English examples, most of the clear errors are bigrams that reflect particular biases of their respective corpora: The phrase *via slashdot* comes from boilerplate text identifying the source of a article, whereas *Maureen* (from *Maureen says*) is a character in one of the novels included in the BNC. The longer FS mostly seem sensible, in that they are plausible lexicalized constructions, though *primarily as ∗ and* from the BNC is not obviously an FS and likely the result of noise. Some FS are dialectal variants, for instance *license endorsed* refers to British traffic violations. More generally, the FS lexicons created by these two corpora are quite distinct, sharing less than 50% of their entries.

What is perhaps most striking about the non-English examples is how poorly some FS translate: many good FS in these languages become extremely awkward when translated into English. This is expected of course for idioms like *biti general poslije bitke* “be the general after the battle” (i.e., “hindsight is 20/20”), but it extends to other relatively compositional constructions like こう言う ∗ が続く “repeat occurrences of ∗ like this” and 前期比 “first half comparison”. This highlights the importance of focusing on FS when learning a language, since many FS do not translate well (though some do, particularly short ones in related languages). Though some of the errors seem to be the result of extra material added to an good FS, for instance *promet te-*

| | |
|---|---|
| English (ICWSM) | *heart ache*, *so ∗ have some time*, *part of the blame*, *via slashdot*, *any more questions*, *protein expression*, *work in ∗ bank*, *al-qaeda terrorist*, *continue discussions*, *I know a lot of people who* |
| English (BNC) | *go into decline*, *Maureen says*, *primarily as ∗ and*, *Peggy Sue*, *square ∗ shoulder*, *delivery system for*, *this ∗ also includes*, *license endorsed*, *point ∗ finger*, *highly ∗ asset* |
| Croatian | *negativno utjecati na* "negatively affects on", *jedan od dobrih poznavatelja* "one of the best connoisseurs of", *jasno ∗ je da* "it is clear to ∗ that", *promet teretnih vozila* "good vehicle traffic", *odvratiti pozornost* "divert attention", *biti general poslije bitke* "be the general after the battle", *popularni internetski* "popular internet", *izazvati kaos* "cause chaos", *austrijski investitor* "Austrian investor", *ideja o gradnji* "the idea of building" |
| Japanese | 高速 道路 整備 "highway construction", 年次 後期 "the second half of the fiscal year", 労働 者 派遣 事業 "temporary labor agency", こう 言う ∗ が 続く "repeat occurrences of ∗ like this", 風邪 っ 匹 "cold sufferer", ＤＨＣＰ サーバー "DHCP server", 前期 比 "first half comparison", 経営 事項 審査 "examination of administrative affairs", 自分 の 文章 "own writing", 深い 味わい "deep flavor" |

Table 5: 10 randomly selected examples from the final FS lexicon from each corpus. Lemmas have been converted to inflected forms where appropriate for readability.

*retnih vozila* "good vehicle traffic", most, again, are somewhat inexplicable artifacts of the corpus they were built from, like *austrijski investitor* "Austrian investor".

Since Zipfian frequency curves naturally extend to multiword vocabulary, our lexicons (and type-based evaluation of the them) are of course dominated by rarer terms. This is not, we would argue, a serious drawback, since in practical terms there is very little value in focusing on common FS like *of course* which manually-built lexicons already contain; most of the potential in automatic extraction comes from the long tail. However, we did investigate the other end of the Zipfian curve by extracting the 20 most common MWEs (including both strong and weak) from the Schneider et al. (2014b) corpus. In the ICWSM lexicon, our recall for these common terms was fairly high (0.75), with errors mostly resulting from longer phrases containing these terms "winning out" (in the lattice) over shorter phrases, which have relatively low LPR due to extremely common constituent words; for example, we missed *on time*, but had 19 FS which contain it (e.g. *right on time*, *show up on time*, and *start on time*). In one case which showed this same problem, *waste * time*, the lexicon did have its ungapped version, highlighting the potential for improved handling of this issue.

In Section 2, we noted than FS is generally a much broader category than MWE, which we take as referring to terms which carry significant non-compositional meaning. We decided to investigate the distinction at a practical level by annotating the positive examples in the ICWSM test set for being MWE or non-MWE FS.[5] First, we note that only 28% of our FS types were labeled MWE; this is in contrast to, for instance, the annotation of Schneider et al. (2014b) where "weak" MWE make up a small fraction of MWE types. Even without any explicit representation of compositionality, our model did much better at identifying MWE FS than non-MWE FS: 0.71 versus 0.34 recall. This may simply reflect, however, the fact that a disproportionate number of MWEs were noun-noun compounds, which are fairly easy for the model to identify.

Due to the lack of spaces between words and an agglutinative morphology, the standard approach to tokenization and lemmatization in Japanese involves morphological rather than word segmentation. In terms of the content of the resulting lexicon we believe the effect of this difference on FS extraction is modest, since much of the extra FS in Japanese

[5] The set was exhaustively annotated by two native-speaker annotators ($\kappa = 0.73$), and conflicts were resolved through discussion.

would simply be single words in other languages (and considered trivially part of the FS lexicon). However, from a theoretical perspective we might very much prefer to build FS for all languages starting from morphemes rather than words. Such a framework could, for instance, capture inflectional flexibilty versus fixedness directly in the model, with fixed inflectional morphemes included as a distinct element of the FS and flexible morphemes becoming gaps. However, for many languages this would result in a huge blow up in complexity with only modest increases in the scope of FS identification. Though it is indisputable that inflectional fixedness is part of the lexical information contained in an FS, in practice this sort of information can be efficiently derived post hoc from the corpus statistics.

Though we have demonstrated that competition within a lattice is a powerful method for the production of multiword lexicons, its usefulness derives less from the specific choices we have made in this instantiation of the model, and more from the flexiblity that such a model provides for future research. Not only do alternatives like DP-seg and LocalMaxs fail to scale up to large corpora, there are few obvious ways to improve on their simple underlying algorithms without compromising their elegance and worsening tractability. Fast and functional, the LPR decomp approach is nevertheless algorithmically ungainly, involving multiple layers of heuristic-driven filtering with no possibility of correcting errors. Our lattice method is aimed at something between these extremes: a practical, optimizable model, but with various component heuristics that can be improved upon. For instance, though the current version of clearing is effective and has practical advantages relative to simpler options that we tested, it could be enhanced by more careful investigation of the statistical properties of $n$-grams which contain FS.

We can also consider adding new terms to the exponents of the two parts of our objective function, analagous to the cover, clear, and overlap functions, based on other relationships between nodes in the lattice. One which we have considered is creating new connections between identical or similar syntactic patterns, which could serve to encourage the model to generalize. In English, for instance, it might learn that verb-particle combinations are generally likely to be FS, whereas verb-determiner combinations are not. Our initial investigations suggest, however, it may be difficult to apply this idea without merely amplifying existing undesirable biases in the LPR measure. Bringing in other information such as simple distributional statistics might help the model identify non-compositional semantics, and could, in combination with the existing lattice competition, focus the model on MWEs which could provide a reliable basis for generalization.

For all four corpora, the lattice optimization algorithm converged within 10 iterations. Although the optimization of the lattice is nevertheless several orders of magnitude slower than the decomposition heuristics of Brooke et al. (2015), the time needed to build and optimize the lattice is still only a fraction of the time required to collect the statistics for LPR calculation, and so the end-to-end runtime of the two methods is comparable. In the BNC, the end-to-end lattice method is several times faster than LocalMaxs (with a fraction of its memory usage), and an order of magnitude faster than DP-Seg.

Finally, though the model presented here was designed specifically for FS extraction, we note that it could be useful for related tasks such as unsupervised learning of morphological lexicons, particularly for agglutinative languages. Character or phoneme $n$-grams could compete in an identically structured lattice to be chosen as the best morphemes for the language, with LPR adapted to use phonological predictability (i.e., based on vowel/consonant "tags") instead of syntactic predictability. It is likely, though, that further algorithmic modifications would be necessary to target morphological phenomena well, and we leave this for future work.

## 7 Conclusion

We have presented here a new methodology for acquiring comprehensive multiword lexicons from large corpora, using competition in an $n$-gram lattice. Our evaluation using annotations of sampled $n$-grams shows that it is consistently outperforms alternatives across several corpora and languages. A tool which implements the method, as well as the acquired lexicons and test sets, are available at: `http://ANON.YMO.US`.[6]

[6] Removed for anonymity

## References

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57.

Vitor De Araujo, Carlos Ramisch, and Aline Villavicencio. 2011. Fast and flexible MWE candidate generation with the mwetoolkit. In *Proceedings of the ACL 2011 Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, pages 134–136, Portland, USA.

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, USA.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. *If you look at...*: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25:371–405.

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)*, pages 753–761, Dublin, Ireland.

Julian Brooke, Adam Hammond, David Jacob, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2015. Building a lexicon of formulaic language for language learners. In *Proceedings of the NAACL '15 Workshop on Multiword Expressions*, pages pp. 96–104, Denver, USA.

Lou Burnard. 2000. User reference guide for British National Corpus. Technical report, Oxford University.

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, USA.

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.

Trevor Cohen and Dominic Widdows. 2009. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405.

Matthieu Constant and Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16)*, pages 161–171, Berlin, Germany.

Joaquim da Silva and Gabriel Lopes. 1999. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Proceedings of the Sixth Meeting on Mathematics of Language (MOL6)*, Orlando, USA.

Nicole Dehé. 2002. *Particle Verbs in English: Syntax, Information, Structure and Intonation*. John Benjamins, Amsterdam, Netherlands/Philadelphia, USA.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics (in Japanese). *Japanese Linguistics*, 22:101–123.

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*, Cargèse, France.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

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.

Adele Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago/London.

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

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.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the 2008*

*Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 992–1001, Honolulu, USA.

Susan Hunston and Gill Francis. 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins.

Taku Kudo. 2008. MeCab: Yet another part-of-speech and morphological analyzer. `http://mecab.sourceforge.net/`.

Nidhi Kulkarni and Mark Finlayson. 2011. jMWE: A Java toolkit for detecting multi-word expressions. In *Proceedings of the ACL 2011 Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, pages 122–124, Portland, USA.

Jey Han Lau, Timothy Baldwin, and David Newman. 2013. On collocations and topic models. *ACM Transactions on Speech and Language Processing*, 10(3):10:1–10:14.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden.

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)*, pages 2077–2092, Mumbai, India.

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

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. 2012. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proceedings of ACL 2012 Student Research Workshop*, pages 1–6, Jeju Island, Korea.

Carlos Ramisch. 2014. *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer, Dordrecht, Netherlands.

Martin Riedl and Chris Biemann. 2015. A single word is not enough: Ranking multiword expressions using distributional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2430–2440, Lisbon, Portugal.

Martin Riedl and Chris Biemann. 2016. Impact of MWE resources on multiword recognition. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 107–111, Berlin, Germany.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing '02)*, pages 1–15, Mexico City, Mexico.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Assocation for Computational Linguistics*, 4:169–182.

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

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.

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, Reykjavík, Iceland.

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)*, pages 100–108, Pittsburgh, USA.

Keiji Shinzato, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. 2008. A large-scale web data collection as a natural language processing infrastructure. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2236–2241, Marrakech, Morocco.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a distributional memory for Croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789, Sofia, Bulgaria.

Thomas Wasow. 2002. *Postverbal Behavior*. CSLI Publications, Stanford, USA.

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

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

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 947–953, Saarbruecken, Germany.

# 3 Original decision letter and reviews

Starts on next page.

Dear ANONYMOUS and coauthors:

As TACL action editor for submission 1137, "Unsupervised Acquisition of Comprehensive Multiword Lexicons using Competition in an n-gram Lattice", I am writing to tell you that I am not accepting your paper in its current form, but due to its current strengths and potential, I encourage you to revise and submit it within 3-6 months.

You can find the detailed reviews below. My judgment is that the submission is not currently acceptable and that it might not be feasible to bring the submission to acceptable form within two months. However, I do think that with significant changes, which could be undertaken in the 3-6 month range, TACL would be very happy to reconsider a revised version.

If you do choose to revise and resubmit, please make use a *new* submission number, and follow the instructions in section "Revision and Resubmission Policy for TACL Submissions" at https://transacl.org/ojs/index.php/tacl/about/submissions#authorGuidelines. I am allowing you one to two additional pages in the revised version for addressing the referees' concerns.

Please understand that while we have endeavored to provide some guidance on how to revise the manuscript, we have NOT provided a complete list of modifications that guarantee acceptance; this is the distinguishing characteristic between the decision we have given your submission --- (c), rejection, but with encouragement to resubmit --- and the next higher level of evaluation, which is conditional acceptance ("(b)", in TACL terminology). The paper will be **reviewed afresh** should you choose to resubmit (possibly involving a change of action editor and reviewers), with **no guarantee of acceptance**, even if you make all the changes suggested.

Again, just to prevent misunderstandings, we repeat: **making all the changes suggested here does not guarantee subsequent acceptance**. A resubmission is treated as a new submission, and the subsequent review may identify different problems with the paper.

Please also note that if you do choose to revise and resubmit, TACL policy is, generally, to try not to give a (c) resubmission another (c), but rather, if the second revision does not meet the acceptance bar, to impose a rejection with a 1-year moratorium on resubmission. Thus, please be very thorough in revising any resubmission.

Thank you for considering TACL for your work, and, although you should take careful note of the caveats above, I do encourage you to revise and resubmit within the specified timeframe.

Noah Smith
University of Washington
nasmith@cs.washington.edu

------------------------------------------------------
------------------------------------------------------
....THE REVIEWS....
------------------------------------------------------
------------------------------------------------------

Reviewer A:

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:

2. Important questions were hard to resolve even with effort.

ORIGINALITY/INNOVATIVENESS: How original is the approach? Does this paper

break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
Note that a paper could score high for originality even if the results do not show a convincing benefit.
:
3. Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by properexperiments and are the results of the experiments correctlyinterpreted?:
4. Generally solid work, although there are some aspects of the approach or
evaluation I am not sure about.

MEANINGFUL COMPARISON: Does the author make clear where the presented system sits with respect to existing literature? Are the references adequate?:
3. Bibliography and comparison are somewhat helpful, but it could be hard
for a reader to determine exactly how this work relates to previous work or what its benefits and limitations are.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?:
4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:
3. Interesting but not too influential. The work will be cited, but mainly
for comparison or as a source of minor contributions.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?:
3. could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined; the training/evaluation data are not widely available.

IMPACT OF PROMISED SOFTWARE:  If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?:
3. Potentially useful: Someone might find the new software useful for their
work.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?:
4. Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Note: after you submit this review form, you'll need to answer a related but different question via a pull-down menu: how long would it take for the authors to revise the submission to be TACL-worthy?

:
        4. Worthy: A good paper that is worthy of being published in TACL.

Detailed Comments for the Authors:
        This paper presents a new approach for formulaic sequence identification
that represents them as lattices and uses the degree of subsumption and overlap to select among similar n-gram variants ("keep under wraps" vs "under wraps"). The proposal adopts the LPR measure of Brooke et al. (2015) to deal with contiguous and gapped sequences, but introduces 3 measures that reward or penalize possible competing variants (cover, clear and overlap) to determine which to select as formulaic sequence. The proposed approach tries to find, for a given set of similar overlapping candidates, a balance between the largest sequence (using the cover measure) that is also small enough to be reusable (using the clear measure) and that minimizes the number adjacent overlapping sequences (using the overlap measure) (keep * under vs. under wraps). Evaluation is done using 4 datasets in 3 languages, where for English two datasets of different sizes are available, to determine the impact of corpus size on the results. The test sets contain both contiguous and gapped candidates (1000 each for ICWSM and 500 for the other corpora). The results obtained show that the proposed method outperforms the results reported by Brooke et al. (2015) for English, with an increase in F-score especially when the 3 measures are take  into account. The results for using only some of the measures tend to produce a lower F-score with the exception of the Japanese corpus, where the use of overlap decreases the results.

The idea behind explainedness, a utility function, is to help determine the locally optimal combination of active nodes in the lattice that inhibit competing (redundant) nodes. The idea is quite neat, and could be seen as a form of Occam's Razor, where a balance needs to be found between specificity of the node that its generality/reusability.  However, the text needs to be revised to clarify some key points and improve readability.

First of all, as it stands, it does not provide enough details about the hypothesis, methodology and motivations for various decisions regarding the method, as detailed in what follows.

In the introduction, the proposed approach is claimed to be "effective, expandable, and tractable", however, it is not clear how each of these claims is evaluated in the paper (effective, expandable and tractable). The approach is also claimed to have the candidates compete for "the best (most parsimonious) explanation for statistical irregularities due to lexical affinity", but concepts like parsimony and lexical affinity need to be defined (what are they, how candidates display them, and how they are measured, ...).

The paper should explicitly state what the hypotheses are in relation to the utility functions, particularly in relation to the approach by Brooke et al. (2015), which is simpler than the proposed approach. Moreover, the state-of-the-art has several simpler formulations for the identification of general formulaic sequences of flexible sizes (e.g. Silva et al. 1999 for contiguous & non-contiguous cases and Villavicencio et al. 2007 for contiguous cases), and the text should justify why this particular approach is appealing, and advantageous in relation to the others. The addition of an error analysis would help highligh these points, helping to clarify the advantages of the particular approach especially in relation to Brooke et al.

The paper should also discuss the motivation for each of the node interaction functions for rewarding/penalizing competitive candidates in the lattice, including how they are expected to contribute to the overall performance of the approach and if the authors considered alternative functions or why these in particular, which linguistic intuition they wanted to capture, why define E0 in terms of the minLPR and E1 of a cost parameter, why they are introduced in the explainedness as exponents (why not using a linear dependence, for example. A suggestion for helping the presentation is for the authors to rewrite E0 and E1 into a single formula that combines the three measures in a utility function, as it would be easier to explain (how E0 and E1 compete, under which circumstances each of them wins), particularly with the  addition of an example. In addition, when presentating these functions, one problem is that although d0 and d1 are used to define E0 and E1 in section 3.2, they are only defined in later sections (sections 3.3.2 and 3.3.3). As a result the discussion appears too abstract as the functions have not been completely defined yet. There should be some explanation for their motivation/intuition, as it's not clear what they are trying to capture, how E0 and E1 would vary for FS and non-FS, etc. My suggestion is to first introduce and motivate the ideas behind them, and later define them together. For d1, there is also a difference in the number of arguments of "oc" in the text (oc(x,y)) and in the definition of d1 (oc(oi)). d0 is defined as a uniform combination of clear and cover, but it would be helpful to have a discussion of the effects of different weights for each of them for the different languages.

In terms of the methodology, a discussion is needed about the motivation for adopting Brooke et al.'s lexical predictability ratio (LPR), and for the modifications proposed, such as the use of minLPR instead of the product of the LPRs and why the focus is on the weakest link (was there any comparison between the use of minLPR and the product of LPRs, is it done for efficiency reasons, etc).  Moreover, the construction of the lattice needs to be explained, including how the nodes relate to one another. For instance, in figure 2, the solid lines indicate subsumption, but "keep everything under wraps" is linked with "be keep * under" via a subsumption relation, even though the latter contains elements that are not in the former (shouldn't it be an overlap link?).

In terms of the evaluation, the distinction between contiguous and gapped test sets, which are later merged into a single test set per language when presenting the results, is not helpful. Instead, maintaining the distinction would help clarify the advantages of the proposed approach in each language, particularly in relation to the error analysis. Additionally, for the contiguous n-grams in these languages a comparison with LocalMaxs could be done. Moreover, the authors need to add the statistical significance of the results in tables 2 and 3 as the approaches often differ only in the second decimal place. In Table 2, it would be helpful to have also the results for only the cover function (Lattice-cl-ovr), to determine how much the other functions contribute. A figure with the top/bottom candidates proposed by the approach be helpful.

Add the list in an appendix for English and explicitly explain the lists for Croatian and Japanese. How many expressions and how many cases they cover.

The paper also needs to add a list of existing MWE lexica or lexical resources with MWEs available to the community (e.g. those listed in the MWE community website, or available in WordNet or other resources).

Stylistically the text would also benefit from a revision that splits sentences that are too long (there are several throughout the text). For instance, the last sentence in the first paragraph of the introduction is 9 lines long: 'We present an effective, expand- able, and above all tractable new approach to com- prehensive multiword lexicon acquisition that aims to find a middle ground between standard MWE ac- quisition approaches based on association measures (Ramisch, 2014), and more sophisticated statistical models (Newman et al., 2012) which fail to scale to the large corpora which are the main sources of the distributional information in modern NLP systems.').

Moreover, some passages are copied from Brooke et al. (2015), but lack the citation, such as:

*"Other mea- sures 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." (page 2) and

*"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" (page 3)

There are also other instances, and these need to be rephrased.

For the annotations, the text mentions 3 annotators, but doesn't specify whether they were different for each of the 3 languages. When comparing the languages, the paper also mentions that "free word order actually results in more of a ten- dency towards contiguous FS, not less.", but it is unclear where this is shown.

Joaquim Ferreira da Silva, Gaël Dias, Sylvie Guilloré, José Gabriel Pereira Lopes: Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. EPIA 1999: 113-132

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, Carlos Ramisch:Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. EMNLP-CoNLL 2007: 1034-1043

Minor comments:

* “accessible by analogy (e.g., glass limb or government ambiguity)” by analogy with what? Specify and add the original expression

* “define the explainedness of a node in terms of two functions” Which two functions?

* In section 3.2 can the authors explain how two valid but similar FSs would be treated  (e.g. keep * under wraps/surveillance)?

* "We also use the concept of hard covering to ad- dress the issue of pronouns, based on the observation that pronouns often have high LPR values". This sentence is too vague. Explain using an example.

* How was the 2/3 threshold for hard covering determined? Why also have soft covering?

* "has been turned on, the covering, blocking, or overlap- ping effects of these other nodes" --> blocking shoud be clearing

* For the efficiency restrictions explain if other values have been tested for them. How did they affect the overall performance? A table detailing the behavior with different values would be helpful.

* "which was roughly the same size (in terms of token count) as the English corpus" specify which English corpus

* In Table 1 the values are less than 2000 (for ICWSM) and 1000 for the other corpora. Add also the initial numbers for each type of FS/non-FS.

REVIEWER CONFIDENCE:

4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my

ratings.

--------------------------------------------------------

--------------------------------------------------------

Reviewer B:

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:

4. Understandable by most readers.

ORIGINALITY/INNOVATIVENESS: How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
Note that a paper could score high for originality even if the results do not show a convincing benefit.
:

3. Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by properexperiments and are the results of the experiments correctlyinterpreted?:

4. Generally solid work, although there are some aspects of the approach or evaluation I am not sure about.

MEANINGFUL COMPARISON: Does the author make clear where the presented system sits with respect to existing literature? Are the references adequate?:

4. Mostly solid bibliography and comparison, but there are a few additional references that should be included. Discussion of benefits and limitations is acceptable but not enlightening.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?:

4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:

3. Interesting but not too influential. The work will be cited, but mainly for comparison or as a source of minor contributions.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?:

2. would be hard pressed to reproduce the results. The contribution depends on data that are simply not available outside the author's institution or consortium; not enough details are provided.

IMPACT OF PROMISED SOFTWARE:  If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?:

1. No usable software released.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?:

1. No usable datasets submitted.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Note: after you submit this review form, you'll need to answer a related but different question via a pull-down menu: how long would it take for the authors to revise the submission to be TACL-worthy?
:
3. Ambivalent: OK but does not seem up to the standards of TACL.

Detailed Comments for the Authors:
// GENERAL

This is nice paper addressing the challenge of acquiring a lexicon of MWE (or rather, Formulaic Sequences (FS), which is a related, though not identical set) with the end goal of potentially improving both language processing technology and investigations/proposals concerning the mental representation of the lexicon. The gist of the contribution is the proposal to construct a complete weighted lattice of potentially related (covered/subsumed/overlapping) sequences that represent legitimate FS candidates, and, by turning lattice arcs on (is-an-FS) or off (not-an-FS) deliver an algorithm that finds a configuration that selects the set of FS that best explain the original data.

// STRENGTHS

The paper addresses the important problem of MWE lexicon acquisition, that, as the authors suggest, despite many years of research, still awaits a suitable (theoretically appropriate as well as empirically viable) solution.

The paper is well written, in the sense that it is well structured, well phrased, the English is flawless, and it paints a clear structure of the argument. I particularly liked the thorough review of related work related formal terms, terms that others would have possibly simply lump under "MWE" without further scrutiny. I also like the proposal both in terms of the idea (constructing a lattice and finding an optimal score by turning on/off mutually related arcs) and the implementation (taking into account covering/clearing/overlaps relations to affect the selected arcs).

The algorithmic solution that is proposed is greedy and exploits local properties of the lattice, which is a reasonable solution given the pressing need for scalability. Assuming that the solution is indeed scalable and empirically viable -- I am wondering if the same idea could be used in the context of other tasks (morphological composition in morphological lattices, preferred expressions in lattices of proposed machine translations, etc). It would be nice if the author could comment on the applicability of the technical solution in other domains (if there is any).

// WEAKNESSES / POINTS FOR IMPROVEMENT

All that said, I have three main concerns about the paper: Concerning the theoretical/linguistic framework within which the discussion is situated, concerning (insufficient) formalisation of the algorithmic solution, and finally, concerning the proposed method of evaluation. I believe many of these concerns could be addressed with relative ease. I will review these matters in details, in turn:

(1) theoretical/ linguistic discussion:

In section 2, the authors go, in great detail, through related work concerning how MWE are defined, classified, and in general what is the theoretical object of this investigation (FS). What is painfully missing in this discussion, is the relation of MWEs, and in particular, of their specific FS formulation, to two major component of NL grammar: (i) syntax, and (ii) morphology.

In the case of syntax, much research in theoretical linguistics and in language technology has proposed the idea that idiomatic expressions (of which MWEs are a special case) are the rule, rather than the exception, in NL grammar. Among the theoretical proposal one can find versions of Construction Grammar (Goldberg 1996) where language knowledge is composed of complete constructions, rather than primitive units (words) and in technological proposals one can find the idea of Data-Oriented Parsing (Remko Scha, Rens Bod), wherein each syntactic subtree (with potential slots for substitution) is an FS-like element that can potentially be used and reused for analysis and generation. And so: while the relation of the proposed FS to existing theories of strongly lexicalized fixed expressions on the one end is clear, it is unclear what is the relation  larger idiomatic syntactic constructions with lexicalized or unlexicalized slots on the other end -- and this relation should be made clear, for the work to be properly situated.

The case of morphology is far less clear, however it is equally important, and it has to do with the basic question of what is the unit that constructs the FS (or the lattices themselves in empirical terms). Are these words? morphemes? lemmas? POS? inflections? In the case  of words, what is the relation of inflected sequences forms / inflected idioms (kicked his bucket vs kicked her bucket, for instance) -- are they the same? related? overalapping? subsumed?. As it stands, the discussion of morphology comes out only later as an afterthought -- where the construction of the lattice in Japanese (in the evaluation section) involves some morphological segmentation whereas the lattice construction for the other languages doesn't. It is thus unclear if the algorithm finds sequences that are of the same type, or are comparable for that matter, across those languages. The selected FS may be incomparable in terms of effective length (morpheme sequences tend to be longer), the status of function words vs. inflection in FS, and so on. The authors should put forth clearly what is the status of morphology inside their general theory (and in terms of their regular expressions), and then later revisit the empirical ramifications of their decision -- for instance what happens when moving between typologically different languages.

(2) the algorithmic formulation:

The author discuss at length the method, the construction of the lattice, and possible relation between arcs (covering clearing and overlapping) and their related scores. The objective function however and the respective optimisation algorithm are discussed informally in passing. I would much prefer a more precise formalisation of the objective function and greedy algorithm, possibly with a running example of several states of progress of the algorithm over a sample lattice. This would also help to clarify what "locality properties" are at play here, which justify a greedy solution that is locally optimised. As a more general note, I think it should be made clear if the target of the lattice construction is a single lattice for the entire text, optimized at once, or multiple lattices, reflecting different aspects of the vocabulary, optimized separately.  As a minor note, there are two many forward references in the text, IMO, which hinder understandability. Example "where d1 and d2 are ... as detailed in the next section". I would prefer the details closer to where they are relevant.

(3) empirical evaluation:

It is unclear to me why comparing the collected lexicon using the proposed method to a sample set of ngrams that respond to a certain frequency threashold is a sound comparison-- isnt the frequency threashold just a simpler and less sophisticated baseline for MWE extraction? It appears that you use this method not as a baseline, but as an actual method to extract gold standard for comparison with your method. I understand that others have used such evaluation methods before -- but I would need much more convincing in order to accept the claim of efficacy. If I misunderstood, and you do not use it as a gold for comparison, or you use a different gold, please clarify.

// OVERALL

I think this is a good paper which, after addressing the various comments I listed above, is certainly worthy of a TACL publication.

REVIEWER CONFIDENCE:
4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

------------------------------------------------------

------------------------------------------------------

Reviewer C:

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:
3. Mostly understandable to me with some effort.

ORIGINALITY/INNOVATIVENESS: How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
Note that a paper could score high for originality even if the results do not show a convincing benefit.
:
4. Creative: An intriguing problem, technique, or approach that is substantially different from previous research.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by properexperiments and are the results of the experiments correctlyinterpreted?:
4. Generally solid work, although there are some aspects of the approach or
evaluation I am not sure about.

MEANINGFUL COMPARISON: Does the author make clear where the presented system sits with respect to existing literature? Are the references adequate?:
4. Mostly solid bibliography and comparison, but there are a few additional
references that should be included. Discussion of benefits and limitations is acceptable but not enlightening.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?:
4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:

4. Some of the ideas or results will substantially help other people's ongoing research.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?:

3. could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined; the training/evaluation data are not widely available.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?:

4. Useful: I would recommend the new software to other researchers or developers for their ongoing work.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?:

4. Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Note: after you submit this review form, you'll need to answer a related but different question via a pull-down menu: how long would it take for the authors to revise the submission to be TACL-worthy?
:

3. Ambivalent: OK but does not seem up to the standards of TACL.

Detailed Comments for the Authors:

Synopsis:

This paper presents a model over n-gram types from a corpus that aims to detect formulaic sequences (FS), which are n-grams that are presumably memorized/prepackaged/entrenched (roughly a superset of multiword expressions, which are the ones with idiosyncratic linguistic properties). The concept of FS in the literature and its relationship to the multiword expression literature is discussed at length.

A lattice model based on n-gram counts is proposed and argued to be a scalable approach to FS detection. The model is structured as a graph whose nodes are n-grams (contiguous or gapped) that might be worthy of listing in the lexicon as formulaic. An optimization procedure decides which nodes to activate (i.e., mark as formulaic) based on functions of corpus counts and the graph structure. In particular, nodes whose n-grams are related by a minimal edit are linked, and different kinds of inhibition relations are defined over these links to discourage redundancy in the lexicon. The formulas in the model make heavy use of the Lexical Predictability Ratio (LPR) from prior work, which is a way of measuring specificity of a word to a surrounding lexical (as opposed to grammatical) context.

To facilitate empirical evaluation (over several corpora/languages), n-grams are sampled and manually annotated as FS or not; inter-annotator agreement is high. In comparison with simple baselines and previously published FS extraction techniques, the lattice method appears to be most successful with respect to F-score.

Evaluation:

The lattice model is, as far as I am aware, novel and creative, with a clear intuitive appeal: it makes it possible to characterize the lexicon of stored n-grams as finding the longest n-gram sequences that occur with higher-than-expected frequency. A couple of broad parallels come to mind: Eisner’s (EMNLP 2002) PCFG model which is parametrized in terms of a graph structure; and n-gram language models (if the formulaic n-grams had probabilities, the parameters could be chosen to assign a high probability to the training corpus while maintaining sparsity, i.e., deactivating many of the n-grams).

Where I had trouble was with the details of the model. First of all, much of it seemed to be ad hoc; even accepting the notion of LPR from prior work, there are many heuristic (or at least not obviously natural) choices made to operationalize the general lattice idea mathematically. Second, even if the model is reasonable, it is not entirely clear how the different parts fit together into an optimization problem. There is mention of a parameter C; and there is a notion of nodes being activated or not (I don’t think notation is introduced for these binary variables, though it would be appropriate). Are these the only things that have to be tuned? Is the calculation of explainedness deterministic based on these? There is a greedy search to optimize explainedness of the nodes in the lattice, subject to various heuristics. Presenting a formal algorithm for the optimization may clarify things considerably.

I think it would also be worth discussing how the lattice model relates back to the cognitive/psychological literature. Would you predict (or are you aware of any evidence) that node interactions such as covering, clearing, and overlapping have an analogue in human language processing? Do any of the design decisions intentionally sacrifice cognitive plausibility for engineering reasons?

Some specific suggestions:

- After the equation for LPR, it would be good to give the intuition: choosing a context span such that w_i is more likely given its context words than its context POS tags.

- Inter-annotator agreement is surprisingly high. Could you elaborate on Wray’s criteria for FS that were applied by annotators?

- I would like to see qualitative analysis of the sequences that were identified as formulaic (or not) by annotators/the method. E.g.: How many of them are syntactic constituents? Are there a wide variety of grammatical kinds of expressions? Are high-frequency as well as low-frequency sequences detected? What is the distribution in terms of length? What kinds of gapped sequences are there? It would be nice to see some examples.

- Related to the previous point: While noting the distinction between FS and MWEs, is it perhaps worth measuring recall against MWE datasets? Most MWEs apart from proper names should count as FS, right?

Typographical issues:

- The ovals in Figure 1 and the edges in Figure 2 show up for me on screen (OS X Preview) but not when printed.
- “pairing down” -> paring down
- “what is typically considered MWE” -> are…MWEs
- “could be useful addition” -> *a* useful addition
- "When the minLPR metric is 1 (crresponding to the case of the syntactic content predicting": content -> context?

REVIEWER CONFIDENCE:

4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.