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.