Catching Idiomatic Expressions in EFL essays

Michael Flor
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
mflor@ets.org

Beata Beigman Klebanov
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
bbeigmanklebanov@ets.org

Abstract

This paper presents an exploratory study on large-scale detection of idiomatic expressions in essays written by non-native speakers of English. We describe a computational search procedure for automatic detection of idiom-candidate phrases in essay texts. The study used a corpus of essays written during a standardized examination of English language proficiency. Automatically-flagged candidate expressions were manually annotated for idiomaticity. The study found that idioms are widely used in EFL essays. The study also showed that a search algorithm that accommodates the syntactic and lexical flexibility of idioms can increase the recall of idiom instances by 30%, but it also increases the amount of false positives.

1 Introduction

An idiom is an expression whose meaning cannot be derived from the usual meaning of its constituents. As such, idioms present a special learning problem for non-native speakers of English (Cooper, 1998), especially learners of English as foreign language (EFL). Understanding of idiomatic expressions can be important, for example, in academic settings, where presentation of ideas often involves figurative language (Littlemore et al., 2011). Even more encompassing is the notion that "natural use of idioms can overtly demonstrate participation in a realm of shared cultural knowledge and interests, and so help a learner gain social acceptance" (Boers and Lindstromberg, 2009). Indeed, it has been claimed that accurate and appropriate use of idioms is a strong distinguishing mark of the native-like command of the language and might be a reliable measure of the proficiency of foreign learners (Cowie et al., 1984).

The present research is informed by the idea that estimation of the use of idiomatic expressions in student essays might be utilized as yet another indicator of proficiency in English. For practical text-analysis applications (e.g. web-based services), and for use in large-scale assessments, such estimation would require automatic tools. Such tools might use a two-step approach: find candidate expressions in text and then verify that they are indeed idiomatic. We have conducted a large-scale study to examine the feasibility of the first step – finding a variety of idiom-candidate expressions in student essays. A wide-coverage extended search algorithm was used to flag candidate expressions and manual annotation was used for verification.

Prior computational work on detection of idioms concentrated on methods of discrimination – is a given expression compositional/idiomatic or not (or to what degree). For purposes of evaluation, such research always relied on manually curated sets of candidate expressions. Our current work is complementary, our question is: how can we automatically obtain a great variety of idiom-candidate expressions, in unrestricted context.

The rest of this paper is structured as follows. Section 2 presents related work on idioms and EFL. Section 3 outlines the complexities of idiom detection. Section 4 describes our approach to detecting candidate idioms in essays. Section 5 describes the corpus and the annotation study. Results and additional experiments are presented in section 6.

2 Idioms and EFL

Applied linguistic research has focused on EFL students' knowledge, comprehension and pro-

duction of idioms. Cooper (1999) investigated idiom comprehension with non-native English speakers from diverse backgrounds, and found that subjects used a variety of strategies for comprehension. Laufer (2000) investigated avoidance of English idioms by EFL university students, using a fill-in translation test, and found that lower English proficiency was associated with greater avoidance of English idioms. Tran (2013) investigated knowledge of 50 idioms collected from the lists of frequently used English idioms and found poor idiomatic competence among EFL students in Vietnam. Multiple factors contribute to *figurative competency*, such as learners' proficiency levels, types of idioms, learners' vocabulary knowledge, and similarity of idioms between foreign and native language (Alhaysony, 2017; Na Ranong, 2014; de Caro, 2009; Irujo, 1986).

Researchers have also looked at figurative language that EFL learners encounter in their educational environments and materials (e.g. textbooks, lectures, etc.). Liu (2003) conducted a corpus-based study of the spoken American English idioms encountered most frequently by college students and provided suggestions for improving the development of idiom teaching and reference materials, including improving the coverage of idiom variants. Littlemore et al. (2011; 2001) investigated the range of difficulties that non-native speakers of English experience when encountering metaphors[1] in British university lectures, including non-understanding (failure to interpret) and misunderstanding (incorrect interpretation).

A complementary line of research focuses on the EFL students' use of metaphors in language production. Littlemore et al. (2014) analyzed the use of metaphors in 200 exam essays written by EFL students, at different levels of English proficiency. They found that metaphor use increases with proficiency level, and even suggested that descriptors for metaphor use could be integrated in the rating scales for writing. Beigman Klebanov and Flor (2013) investigated the use of metaphors in 116 argumentative essays and found moderate-to-strong correlation between the percentage of metaphorically used words in an essay and the writing quality score. Notably, both studies used a small number of essays and conducted an exhaustive manual analysis of metaphoric expressions.

3 Idiom identification

Syntactic and lexical flexibility are two of the issues dealt with at length in the linguistic and psycholinguistic literature on idioms (Glucksberg, 2001; Nunberg et al., 1994). Idioms can vary from being fully syntactically flexible to not at all. Although, traditionally, idiomatic expressions had been considered as 'fixed expressions' (Alexander, 1978), researchers have demonstrated that idioms allow a lot of variation, including adjectival and adverbial modification, quantification, negation, substitution, passivization and topicalization. Glucksberg (2001) illustrates the flexibility of idiomatic expressions, using the idiom "*don't give up the ship*", which has a wide range of variations:

1. Tense inflection: *He gave up the ship.*

2. Number inflection: *Cowardly? You wont believe it: They gave up all the ships!*

3. Passivization: *The ship was given up by the city council.*

4. Adverbial and adjectival modification: *After holding out as long as possible, he finally gave up the last ship.*

5. Word substitution: *Give up the ship? Hell, he gave up the whole fleet!*

It has been long noted that many idioms allow for application of various kinds of modifiers, which often insert words and phrases around or even into the core idiomatic phrase (Ernst, 1981). Linguists have proposed different theories and taxonomies for idiom modification (McClure, 2011; Glucksberg, 2001; Nicolas, 1995), while psycholinguistic experiments demonstrated the flexibility of idiom recognition mechanisms (Hamblin and Gibbs, 1999; McGlone et al., 1994; Gibbs and Nayak, 1989; Gibbs et al., 1989). Researchers who focused on computer-aided identification of idiomatic expressions in texts have noted the need to account for idiom flexibility (Bond et al., 2015; Minugh, 2006; Moon, 1998).

---

[1]On the close relation between idioms and metaphors, see Gibbs et al. (1997)

In this respect, it is important to mention one very common sub-type of idiomatic expressions: idioms that are not fully lexically specified. Such idioms, e.g. "*be the apple of one's eye*", include slots that must be filled in context, thus involving modification and discontinuity of the lexical components of the idiom, posing an additional challenge for automatic detection.

3.1 Automated detection of idioms

In computational linguistics, idiom detection systems fall into one of two paradigms (Muzny and Zettlemoyer, 2013): type classification, where a decision is made whether an expression (out of any context) is always/usually idiomatic or literal (Shutova et al., 2010; Gedigian et al., 2006; Widdows and Dorow, 2005), and token classification, where each occurrence of a phrase, in a specific context, can be idiomatic or literal (Peng et al., 2014; Li and Sporleder, 2009; Sporleder and Li, 2009; Fazly et al., 2009; Katz and Giesbrecht, 2006).

Early work on idiom detection involved small sets of expressions (Fazly and Stevenson, 2006), and focused on specific types of syntactic constructions (such as verb + complement, e.g. "*stir excitement*","*play with fire*") (Shutova et al., 2010; Li and Sporleder, 2009; Diab and Bhutada, 2009; Diab and Krishna, 2009). More recent research on detection of non-compositional word combinations has shown a proliferation of approaches, but much work still focuses on acontextual classification (Hashimoto and Tsuruoka, 2016; Cordeiro et al., 2016; Ramisch et al., 2016; Yazdani et al., 2015; Salehi et al., 2014; Salehi and Cook, 2013; Kiela and Clark, 2013; Reddy et al., 2011). Recent work on detection of idiom instances in context (Gharbieh et al., 2016; Salton et al., 2016; Peng et al., 2014) focused only on Verb+Noun constructions, using the same dataset (Cook et al., 2008). A notable exception is the work of Feldman and Peng (2013), which is not limited by the type of syntactic construction.

4 Procedure for identifying idiom-candidates in essays

Our approach to identifying idiomatic expressions in texts is motivated by three factors.

First, we aim for broad coverage, so as to identify as many different idioms as possible. Second, we aim at identifying idiomatic expressions in context, in real-life texts. Third, our focus is on learner language, in essays written by non-native learners of English. We assume that most of the idioms that might be found in such texts are very well known idioms that are listed in various dictionaries. Our approach to idiom detection proposes two phases: candidate detection followed by verification. We compiled a large listing of idiomatic expressions that we want to detect. The idea is to automatically identify such expressions in texts, as candidate-idioms, and then apply verification algorithms that would confirm/reject the candidate expressions as being an idiom in the given context. In this paper we report on our initial results with the first part of this approach - detecting candidate-idiom expressions in student essays.

4.1 A collection of idioms

For our collection, we use Wiktionary as a resource. Wiktionary has a facility for contributors to tag definitions as idiomatic. The English Wiktionary was used in some previous computational work on idioms (Salehi et al., 2014), as it has rather broad coverage for idioms (although it is far from being complete (Muzny and Zettlemoyer, 2013)). We collected all English expressions that were tagged as idiomatic, from the English Wiktionary of October 2015. That initial list totaled about 8,000 entries. From that list, we eliminated several classes of expressions. First, we eliminated all single-word expressions, (e.g. *backwater*), since we are interested in idiomatic phrases. Next, we eliminated verb-particle constructions and prepositional verbs (such as *whisk away* and *yell at*). Finally, we eliminated expressions that are common greetings (e.g. *good evening*) or conventional dialogic expressions (e.g. *how do you do*). The resulting list contains 5,075 English idiomatic expressions. The list is of course extensible and more idioms can be added in the future.

4.2 The algorithm

Our algorithm for detecting candidate idiom expressions involves checking whether any of the listed idioms occur in a text. Since id-

iomatic expressions can exhibit considerable flexibility with inflectional and syntactic-form variations, a broad-coverage search algorithm must take such variation into account. This is achieved by enriched representation and flexible algorithmic matching.

Our initial Wiktionary-based list of 5,075 expressions contains only canonical forms of idioms. Using an in-house morphological toolkit, we automatically enrich the representation of an idiom entry by including all inflectional variants to the idiom's content words. The automatic expansion is not part-of-speech sensitive. For example *"melting pot"* is expanded to *"{melting, melt, molten, melts, melted, meltings} {pots, pot, potted, potting}"*.

The next step is to mark optional elements in the idiom representation: determiners, prepositions and a set of other common function words (see appendix for the full list), as well as possessive *"'s"*, and punctuation like commas and hyphens. An idiom should be matched even if such elements are missing in the text. For example, with inflectional expansion and with marking of optional elements, the idiom *"give the royal treatment"* becomes *"{give, given, gave, giving, gives} [the,a,an] {royal, royals} {treatment, treatments}"*. The need for optional elements stems from the notion that writers, especially EFL writers, often omit articles and prepositions, or use erroneous ones (Dale et al., 2012).

The third step is the treatment of idioms that are not fully lexicalized, for example *"pour one's heart out"* or *"knock someone's socks off"*. We pre-fill the slots with a set of pronouns that might occur in such position. For idioms that include a possessive slot, we substitute the canonical *"someone's"* with possessive pronouns. For example, *"knock someone's socks off"* becomes *"{knocked, knock, knocking, knocks} [my, your, his, her, our, their, one, someone] ['s] {sock, socked, socking, socks} off"*. For other idioms, the substitution list uses non-possessive pronouns. For example, in canonical expressions like *"bite off more than one can chew"*, *"one"* is substituted with *"{i, you, he, she, we, they, one, someone,somebody, me, him, her, us, them}"*. Reflexive pronouns in canonical idiom forms (e.g. *"let oneself go"*)

are expanded to a set of reflexives *"{myself, oneself, yourself, yourselves, himself, herself, itself, ourselves, themselves}"*. All automatically added pronouns are treated as optional elements. This treatment does not fill the slots with non-pronominal material (names and full noun phrases), but that is compensated with the skip-words-algorithm (see below).

The automated enrichment described above is performed only once, when we transform the list of canonical idioms into an enriched search-specification format. Some idioms allow insertion of various modifiers over the core components, for example *"kick the proverbial bucket"*, *"pay little attention"*. To detect such variant instances, we provide some flexibility to the search algorithm. Essentially, the search algorithm must match all the non-optional elements of an idiom, in sequence. Flexibility is achieved when the algorithm is allowed to match the core components, in order (as specified by the enriched representation), but they don't have to be consecutive. The algorithm may allow up to k unmatched words between the first and last elements of an idiom. This enables detection of idioms with unspecified modifiers and intervening insertions. The value of k is a settable parameter.

Note that the algorithm has two separate skip strategies. On the one hand, there are optional elements in the idiom search-specification, such as determiners or pronouns. This means that not all components of an idiom have to be matched in order to spot a potential idiom-instance. On the other hand, the algorithm can skip over tokens in the text, to allow for intervening material. The combination of these two approaches allows to find instances of lexically underspecified idioms. For example, the idiom *"change one's mind"* is expanded to *"{changes, changing, change, changed} [my, your, his, her, our, their, one, someone] ['s] {minds, mind, minding, minded}"*, and the algorithm can identify *"changed the people's minds"* in a text, because the pronouns are optional and '*the*' and '*people*' are skippable.

The approach outlined above was implemented with a tokenizer, a sentence-boundary detection module and an indexing module. Since we are using a tokenizer, the idiom-

Annotation category	Explanation
Idiomatic use	choose this option if you think that the sentence indeed contains an instance of the idiom
Literal Use	choose this option if you think that the expression is correct, but it is used in a literal and not idiomatic sense
Wrong Expression	choose this option if you think that the system picked up a wrong expression, not an intended one
Need More Context	choose this option if you feel that you need more context to decide

Table 1: Classification categories for the idiom annotation study.

search specifications are token-oriented, which allows for very simple specification of patterns (e.g. all the examples above). The sentence detector allows restricting the search only within sentences (and never across sentences). For each sentence in each text under consideration, we need to check whether any of our 5,075 enriched expressions is present in the sentence. Naive search would amount to matching against 5,075 expressions. Indexing allows for a faster solution. The enriched dictionary of idioms is indexed by keywords (non-optional idiom components) when it is loaded to memory. Each text (essay) is also indexed, on-the-fly, when loaded for processing. The indices are cross-compared, and the algorithm attempts to find only those idioms whose keywords appear in the index of the current text.

One limitation of the above approach is the constraint of sequential matching (even with skips). Some idioms are flexible enough to allow for passivization or topicalization (Glucksberg, 2001), variations that invert the word order (especially for idioms involving a verb + direct object, e.g. *the ship was given up by the city council*). Extending our algorithm to handle such cases is left for future work.

It should be stressed that the approach outlined above identifies idiom-candidates, i.e. it finds, in texts, expressions that are likely to be instantiations of stock idioms. However, the current algorithm does not perform any verification - it does not attempt to confirm that the detected expressions are actually idioms in context. Adding such capabilities is subject of continuing research.

5 Data and annotation

We conducted a study in which our flexible algorithm was applied to a large set of essays written by EFL students. Candidate-idioms were automatically marked and later manually annotated.

5.1 Data

We used the publicly available corpus of essays, the ETS Corpus of Non-Native Written English (Blanchard et al., 2014, 2013). This corpus consists of essays written for the TOEFL® iBT test. The test is used internationally as a measure of academic English proficiency, among other purposes, to inform admissions decisions for students seeking to study at institutions of higher learning where English is the language of instruction. The corpus contains about 12,000 essays, sampled from eight prompts (i.e. eight different discussion topics), along with score levels (low/medium/high) for each essay. Each prompt poses a proposition and asks examinees to write an argumentative essay, stating their arguments for or against the proposition.

For our present work, we sampled 3,305 essays from this corpus, selecting (a) only among essays that received medium or high score; and (b) only among essays that had at least one candidate idiom match (using the algorithm with maximum skip $k = 4$). The sampled data set has 1,111,618 words; essay length varies from 143 to 801 words, with an average of 336.

5.2 The annotation study

In total, our algorithm identified 5,704 expressions as candidate-idiom instances, in the 3,305 essays. All those expressions were then annotated, using the following setup. For each candidate-idiom expression, the whole sentence in which that expression occurred was automatically extracted from the essay, and all such sentences were collected in a spread-

sheet file. For each extract, we provided the full sentence, what idiom (canonical form) was tentatively detected, and what were the first and last words of the detected instance. For each candidate-expression, the annotator had to pick one out of four classification options (see Table 1).

All annotation was performed by a single annotator, a native speaker of American English, contracted through a commercial linguistic service provider. The annotator was given an explanation of how the data was pre-processed, and was encouraged to consult the Wiktionary entries for the canonical stock expressions. Upon completion of a training session with 100 instances, the annotator was given 300 new candidate instances. This set of 300 items was also annotated by the first author. We had exact agreement in 285 cases out of 300, which is 95% (Cohen's kappa 0.92). The annotator then proceeded to annotate the rest of the 5K+ candidate instances. The first author also adjudicated the disagreed cases from the 300-items set, and twenty-one instances that the annotator marked as 'Need More Context' in the rest of the data.

6   Results

Out of 5,704 instances marked by our algorithm, the annotation study confirmed 1,302 cases as idiomatic uses, 693 cases were found to be literal uses, and 3,709 cases were classified as wrong expressions.

It should be noted that since the annotation was performed only on the automatically flagged candidate instances, it is quite possible that essays in our data set contain even more idioms: a) undetected instances (e.g. due to word order inversions, insertions larger than $k = 4$, etc.), and b) instances of idioms that are not on our current list.

The 1,302 attested idiom instances in our data belong to 294 types (canonical forms). Table 2 lists some of the most common idioms found in the essays. Thus, out of 5,075 idioms types in our dictionary, we found attested instances for $294/5,075 = 5.8\%$. This demonstrates that argumentative essays written to TOEFL prompts have quite a rich variety of idiomatic expressions. Notably, the idioms were not concentrated in just a few es-

says. Out of 3,305 essays, 1,017 essays (30%) had at least one verified idiom instance.

Idiom (canonical form)	Count
pay attention	112
matter of fact	84
other than	54
long run	46
find oneself	37
come to mind	36
side effect	35
day-to-day	34
change one's mind	32
again and again	30
great deal	28
jack of all trades	23
rush hour	22
open doors	21

Table 2: Instance counts for fourteen most frequent idioms found in student essays in the corpus.

The majority (65%) of the automatically marked candidates were classified as 'Wrong Expression' (WE). Such instances are misdetected by our algorithm when the mandatory content words of an idiom-specification do occur in text, but are not part of the sought-for expression, or are even parts of unrelated expressions. See examples in Table 3.

Ideally, we would like our algorithm to mark as candidates only expressions that might be idioms or literal uses, so that some verification algorithm might then distinguish among them. The proliferation of wrong expressions complicates this outlook. In order to check how the quality of marked candidate instances is affected by our skip algorithm, we conducted two additional experiments.

6.1   Additional experiments

We applied the candidate-idiom detection algorithm to the 3,305 essays, using different values of the max-skip-tokens parameter k, from 0 to 4. With $k = 0$, no intervening words are allowed within an idiom. Notably, $k = 4$ was used in the annotation study, so all candidate expressions marked in runs with smaller values of k are proper subsets of the annotated data. The results are presented in Figure 1A.

Predictably, increasing the value of k allows to detect more idioms, but it also leads to the

Canonical form	Sentence with candidate	Status
long run	Because, such advertisements are neither wise and profitable options for firms in the *long run* nor legal in many countries.	Idiom
grass roots	we have to understand the content from the *grass root* level of that matter.	Idiom
try one's hand	Thereby we have stories of some 60-70 year old *trying their hands* at trekking or a cross-country run.	Idiom
draw a line	When do we *draw the line* to where we should stop gaining any new knowledge?	Idiom
draw a line	Suppose if a student is thought in class to *draw lines*, boxes...	Literal
great deal	Some people even offer a *great deal*, but you have to pay in advance, and in the end you do not even get a product.	Literal
leave home	And also the most of us *leave home* eary in the morning and come back home late in the night.	Literal
well-oiled	People already realize *well the oil* will be run out in a short time.	WE
come to life	So can you disagree with above statement after *coming across Faradays life*?	WE
any more for any more	The *more you do, the more* you learn, and life become more interesting.	WE

Table 3: Examples of candidate-idiom expressions in context and their annotations.

increase in the number of candidates that are literal uses, and an increase in the number of wrongly-marked expressions (false positives). The largest increase is observed in transition from zero to just one allowed intervening word. The number of detected idioms increases by 222 instances (22%), while the number of literal uses increases by 79 instances (13%). At the same time, the number of wrong expressions increases dramatically from 153 to 2214 (more than a 1300%).

As we raise the value of k further, the amount of added idiomatic instances decreases (3.7% added at $k = 2$, 2% at $k = 3$ and 0.7% at $k = 4$). The amount of added literal uses also decreases (1.3%, 0.7%, 0.4%). The amount of added WE instances decreases slowly (25%, 17%, 14.8%), hundreds of WE instances are added for each increment of k. This suggests that $k = 4$ might be a practical limit for our current approach, since wrong expressions become increasingly dominant in the output.

The largest number of wrong expressions is produced by the idiom "*any more for any more*": 683 at $k = 1$, rising to 998 when $k = 4$. Since 'any' and 'for' are optional, the algorithm flags any sequence of 'more ... more' with up to k intervening words. Other idioms that generated more than 100 WE instances

(at $k = 4$) are "*day of days*" (157), "*well and good*" (134), "*more like it*" (124). No literal or idiomatic use of those expressions was found.

Overall the skip-enabled search shows considerable promise. With no skip, the algorithm found 1,000 idiom instances in texts. With skip $k = 4$, the algorithm found 1,302 instances, an increase of 30%. To illustrate the usefulness of the skip-enabled search, we list some extended forms of idioms that were detected. For "*pay attention*": *researchers should pay their attention on the specific subject; if Einstein had not paid specific attention to...; pay particular attention*. For "*change one's mind*": *...people change their mind; you might change your mind; the customer change his mind after...; advertisements can change consumer's mind about products*.

In a second experiment we also varied the values of k, but this time we switched all the optional (function) words in idiom specifications to being mandatory. Thus, for example, for "*draw a line*", a determiner in the middle is now mandatory – one of {*the,a,an*} should be matched for an instance to be flagged. (Punctuation and "*'s*" remain optional.) The results are presented in Figure 1B.

The general trends observed in the previous experiment are still present: as the number

of allowable insertions rises, more idiom instances are detected, but also more literal uses and more misdetected expressions; the increment decays with larger k.

Next we compare between the results of the two experiments (each bar in Figure 1A vs. a corresponding bar in Figure 1B). When function words in the patterns are mandatory, the number of detected idioms is reduced by 0.6% at $k = 0$, 3.6% at $k = 1$, 5.4% at $k = 2$, 6.5% at $k = 3$ and 6.7% at $k = 4$ (from 1,302 to 1,214). There is also some reduction in the number of detected literal-use instances (6.2% at $k = 4$). The strongest reduction is in the number of misdetected expressions: 70% at $k = 4$ (3,709 to 1,090) and 74% at $k = 1$. Some such reduction might have been expected: with all mandatory components, the idiom patterns are stricter, and so less irrelevant material fits into them. However, the magnitude of the reduction is impressive, as it demonstrates that function words in idioms can be very useful for filtering out irrelevant material.

Still, with function words being non-optional, we loose about 6.7% of idioms. Here are some corpus examples of idiom instances that are detected when optional components are allowed, but are not detected otherwise. For 'pain in the neck': "...but it's always a pain of neck to decide whether going with a tour guide or by themselves"; here the student used a wrong preposition of. For 'seize the day': "...young people tend to seize each day because even in his early age an human being is fully aware..."; here the student used the unexpected determiner each, but not any from the 'mandatory' set.

7 Conclusions

We presented a large-scale investigation of the use of idiomatic expressions in argumentative essays written by non-native English speakers.

We described a search procedure for automatic detection of candidate phrases in essay texts. The procedure was developed to address multiple demands - provide wide coverage (with an extensible dictionary with thousands of idioms) and address the flexibility of idiomatic expressions (via lexical enrichment and skip-steps in the search algorithm).

In an annotation study, candidate-idiom instances were automatically marked and then manually classified as idiomatic, literal, or wrong (misidentified) expressions. The study revealed that stock idiomatic expressions are quite common in EFL student essays and that a rather rich variety of English idioms is used.

Our study has confirmed the importance of tending to the syntactic and lexical flexibility of English idiomatic expressions. Allowing optional components in idioms and lexical insertions in text, increases recall of idiom instances by 30% relative to a baseline.

The flexible candidate-detection algorithm also flags a lot of irrelevant material, especially when more intervening words are allowed within an idiom. We have shown that consideration of function words in idioms can
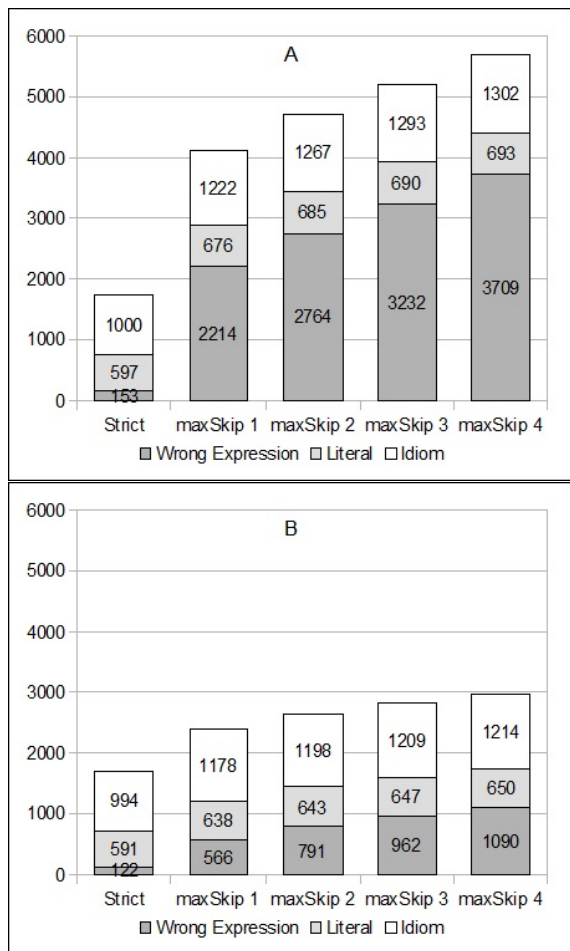

Figure 1: Counts of Idiom, Literal Use and Wrong Expression instances marked in essays, as a function of the number of allowable intervening words in candidate detection. Panel A: with optional words in idioms; Panel B: all words in idioms are mandatory.

help reduce the amount of false positives. We are working on integrating those findings towards an improved algorithm.

References

Richard Alexander. 1978. Fixed expressions in English: a linguistic, psycholinguistic, sociolinguistic and didactic study. *Lingustic Agency, University of Trier*, 26:171–188.

Maha Alhaysony. 2017. Strategies and difficulties of understanding English idioms: A case study of Saudi university efl students. *International Journal of English Linguistics*, 7(3):70–84.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia. Association for Computational Linguistics.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. *TOEFL11: A Corpus of Non-Native English. Research Report ETS RR13-24*. Educational Testing Service, Princeton, NJ, USA.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2014. *ETS Corpus of Non-Native Written English, Catalog No. LDC2014T06*. Linguistic Data Consortium, Philadelphia, PA, USA.

Frank Boers and Seth Lindstromberg. 2009. *Optimizing a Lexical Approach to Instructed Second Language Acquisition*. Palgrave MacMillan, UK.

Francis Bond, Jia Qian Ho, and Dan Flickinger. 2015. Feeling our way to an analysis of English possessed idioms. In *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar*, pages 61–74, Stanford, CA, USA. CSLI Publications.

Edith Eliana Roberto de Caro. 2009. The advantages and importance of learning and using idioms in English. *Cuadernos de Lingstica Hispnica*, 14:121–136.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22. European Language Resources Association (ELRA).

Thomas C. Cooper. 1998. Teaching idioms. *Foreign Language Annals*, 31(2):255–266.

Thomas C. Cooper. 1999. Processing of idioms by L2 learners of English. *TESOL Quarterly*, 33(2):233–262.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997, Berlin, Germany. Association for Computational Linguistics.

Anthony Paul Cowie, Ronald Mackin, and Isabel R. McCaig. 1984. *Oxford Dictionary of Current Idiomatic English, vol. I-II. General Introduction*. Oxford University Press, Oxford, UK.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. Hoo 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada. Association for Computational Linguistics.

Mona Diab and Pravin Bhutada. 2009. Verb noun construction mwe token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 17–22, Singapore. Association for Computational Linguistics.

Mona Diab and Madhav Krishna. 2009. Handling sparsity for verb noun MWE token classification. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 96–103, Athens, Greece. Association for Computational Linguistics.

Thomas Ernst. 1981. Grist for the linguistic mill: Idioms and extra adjectives. *Journal of Linguistic Research*, 113(5):51–68.

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

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 337–344. Association for Computational Linguistics.

Anna Feldman and Jing Peng. 2013. Automatic detection of idiomatic clauses. In *Computational Linguistics and Intelligent Text Processing*, pages 435–446. Springer.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 41–48, New York City, New York. Association for Computational Linguistics.

Waseem Gharbieh, Virendra C. Bhavsar, and Paul Cook. 2016. A Word Embedding Approach to Identifying Verb-Noun Idiomatic Combinations. In *Proceedings of the 12th Workshop on Multiword Expression*, pages 112–118. Association for Computational Linguistics.

Raymond W. Gibbs, Josephine M. Bogdanovich, Jeffrey R. Sykes, and Dale J. Barr. 1997. Metaphor in idiom comprehension. *Journal of Memory and Language*, 37:141–154.

Raymond W. Gibbs and Nandini P. Nayak. 1989. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, 21(1):100–138.

Raymond W. Gibbs, Nandini P. Nayak, and Cooper Cutting. 1989. How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, 28(5):576–593.

Sam Glucksberg. 2001. *Understanding Figurative Language: from metaphors to idioms*. Oxford University Press, New York, NY.

Jennifer L. Hamblin and W. Gibbs, Raymond. 1999. Why you cant kick the bucket as you slowly die: Verbs in idiom comprehension. *Journal of Psycholinguistic Research*, 28(1):25–39.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 205–215, Berlin, Germany. Association for Computational Linguistics.

Suzanne Irujo. 1986. A piece of cake: Learning and teaching idioms. *ELT Journal*, 40(3):236–242.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1427–1432, Seattle, Washington, USA. Association for Computational Linguistics.

Batia Laufer. 2000. Avoidance of idioms in a second language: the effect of l1-l2 degree of similarity. *Studia Linguistica*, 54(2):186–196.

Linlin Li and Caroline Sporleder. 2009. A cohesion graph based approach for unsupervised recognition of literal and non-literal use of multiword expressions. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 75–83, Suntec, Singapore. Association for Computational Linguistics.

Jeannette Littlemore. 2001. Metaphor as a source of misunderstanding for overseas students in academic lectures. *Teaching in Higher Education*, 6(3):333–351.

Jeannette Littlemore, Tina Krennmayr, James Turner, and Sarah Turner. 2014. An investigation into metaphor use at different levels of second language writing. *Applied Linguistics*, 35(2):117–144.

Jeannette. Littlemore, Phyllis Trautman Chen, Almut Koester, and John Barnden. 2011. Difficulties in metaphor comprehension faced by international students whose first language is not English. *Applied Linguistics*, 32(4):408–429.

Dilin Liu. 2003. The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly*, 37(4):671–700.

Scott McClure. 2011. Modification in non-combining idioms. *Semantics and Pragmatics*, 4(7):1–7.

Matthew S. McGlone, Sam Glucksberg, and Cristina Cacciari. 1994. Semantic productivity and idiom comprehension. *Discourse Processes*, 17(2):167–190.

David Minugh. 2006. The filling in the sandwich: Internal expansion of idiomatic expressions. In Roberta Facchinetti, editor, *Corpus Linguistics 25 Years on*, pages 205–224. Rodopi, Amsterdam, NL.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Clarendon Press, Oxford, UK.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA. Association for Computational Linguistics.

Sirirat Na Ranong. 2014. Idiom comprehension and processing: The case of Thai EFL learners. *Journal of English Studies*, 9:51–97.

Tim Nicolas. 1995. Semantics of idiom modification. In Martin Everaert, Erik-Jan van der Linden, Andrew Schenk, and Robert Schreuder, editors, *Idioms: Structural and Psychological Perspectives*, pages 233–254. Lawrence Erlbaum, Hillsdale, NJ, USA.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 266–275, Atlanta, Georgia, USA. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Detecting non-compositional mwe components using wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1792–1797, Doha, Qatar. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sententional distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association of Computational Linguistics*, pages 194–204, Berlin, Germany. Association for Computational Linguistics.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, Beijing, China. Coling 2010 Organizing Committee.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Huong Quynh Tran. 2013. Figurative idiomatic competence: An analysis of EFL learners in Vietnam. *Language Education in Asia*, 4(1):23–38.

Dominic Widdows and Beate Dorow. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 48–56, Ann Arbor, Michigan. Association for Computational Linguistics.

Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon, Portugal. Association for Computational Linguistics.

Appendix

The list of words that were defined as optional in idiom specifications: Determiners: *a, an, the, any, some*; Wh-words: *what, who, whom, whose, how, when, why, where*; Auxiliary verbs: *can, can't, cannot, could, couldn't, may, might, should, do, does, did, done, don't, doesn't, didn't*; Be forms: *be, been, was, wasn't, were, weren't, ain't, am , is, are, isn't, aren't*; Common prepositions: *in, on, of, off, at, as, to, for, from, down, up, it, and, or, with*; Pronouns: *i, me, my, you, your, he, his, him, she, her, hers, we, our, ours, us, they, them, their, theirs*; Demonstratives: *there, here, this, that, these, those*; Other: *but, yet, and, or, so, s, 's, one, someone, somebody, thus, such, ever, never, no, not, none*.