A Gensim Adaptation for Idiom Recognition on Large Corpora

Chandler Jones**
cjones87@calpoly.edu
California Polytechnic State University SLO
San Luis Obispo, California, USA

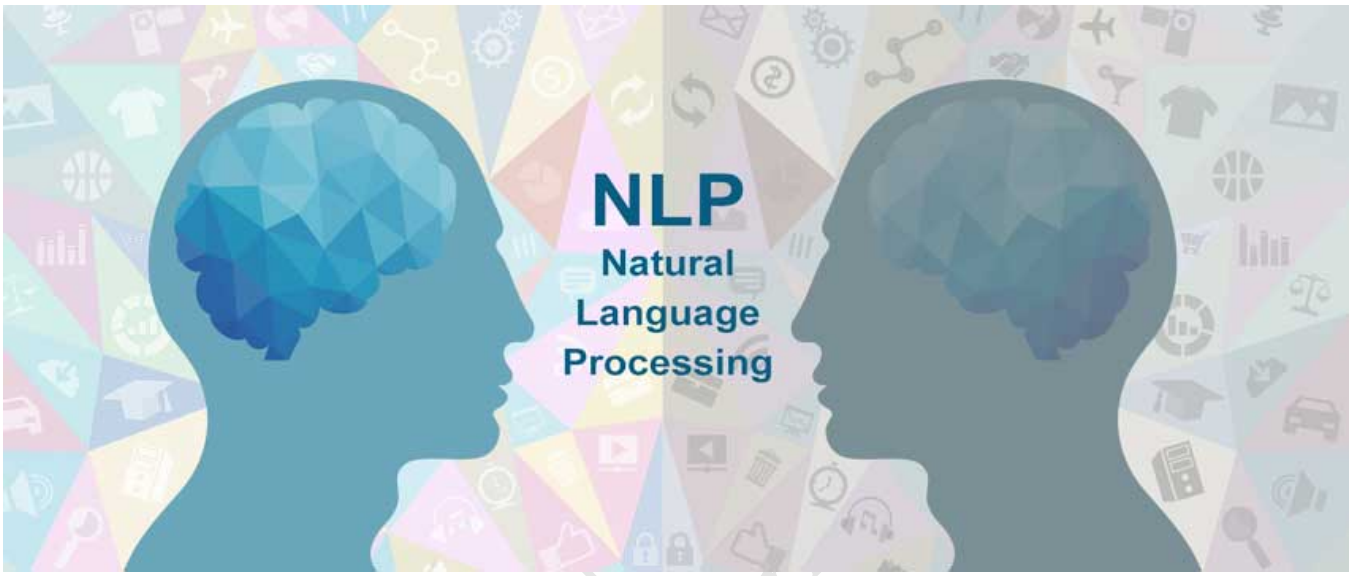

Figure 1: Graphic of NLP composition

ABSTRACT

Idiom recognition has remained a challenge in the Natural Language Processing field. Idioms, and the sub-strings which comprise them are so varied in type, frequency, and usage; that NLP programs often fail to distinguish them from standard text. Idioms come in many forms which may occur as rarities between long dialogues of legal text, or common occurrences in daily informal communication. While their usages can sometimes be parsed from the literal translation of the contained words, many times idioms have illogical or perpendicular meanings to the subtext of the phrase. Further, idioms morph in their usages, few have standard form, and others are often misused or combined with imperfect language usages.

In this thesis, we implement a parser to extract and identify idiom usages in large corpora. We use the largest known database of complied idioms to test identification on first the portion of the Brown Corpus contained within NLTK, and then on the four major sections of the COCA corpus to identify the prevalence of idioms in different forms of language. The parser uses a state of the art adaptation of the Gensim model's Multi-word Expression tool to accurately identify collocated idiom usages. The goal in this research is an initial foray into the field to expand the detection of idioms, and learn in which settings idioms have the most popular usages. We find that the prevalence of idioms across genres to occur at the rate of roughly 3 instances per 1000 sentences.

CCS CONCEPTS

• **Software and its engineering** → *Search-based software engineering*; • **Human-centered computing** → *Empirical studies in HCI*; • **Applied computing** → **Document searching**.

KEYWORDS

Gensim, Collocation, Multi-word Expressions, text tagging

ACM Reference Format:
Chandler Jones. 2021. A Gensim Adaptation for Idiom Recognition on Large Corpora. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The issue of idiom recognition remains a challenging prospect for Natural Language Processing. In fact, the prevalence of idioms in text or vocal language is varied in type, frequency, and meaning. An informal rule for idiom usage in conversational english is to, 'Speak

*Research Completed under the direction of Dr. Foaad Khosmood.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

idiomatically unless there is a good reason not to do it.' [3]. Idioms present a complex case for foreign language learners to parse, and often leave the learner with a severe lack of understanding despite the ability to translate each word individually [4], [11]. As idioms are often created, modified, and mixed on the fly, there remains no central repository of Idioms in English, let alone other foreign languages[10]. Despite these characteristics, their form generally follows grammatical rules, and as such cannot be parsed out as a singular phrase. The use of idiomatic language often integrates the language of specific groups. Ingroup language can be defined by specific idiom usages. Speakers may find that some locals use certain idioms while foregoing others, and other groups use the others and forego the first group altogether. In this way, idioms can be used to quickly identify the ingroup of the speaker[1].

For Natural Language Processing, the focal point of idioms becomes a near insurmountable problem. Without the presentation of nonverbal language to rely on, idiom usages remain undetected. And yet, idioms generally have definitive meanings. Their usages are particular, and are generally not used without cause, but to convey a generally agreed upon idea. Confoundingly, idioms can be used literally. *"Add fuel to the fire"* can mean both to invigorate a situation, or to literally add wood to a fire [12]. Parsing out this meaning without a central repository is a yet unconquered feat in the field of natural language processing.

Idioms themselves fall into a category of language called **Multi-Word Expressions, (MWEs)**. An overview of the **MWE** field can be found here [2]. A MWE can be thought of as a collection of words which together can be thought of as having a single definition [5].

If they can be identified, the incorporation of MWE knowledge has been shown to improve task accuracy for a range of NLP applications including dependency parsing (Nivre and Nilsson 2004), supertagging (Blunsom and Baldwin 2006), sentence generation (Hogan et al. 2007), machine translation (Carpuat and Diab 2010), and shallow parsing (Korkontzelos and Manandhar 2010) [5].

1.1 Defining Idioms

MWEs form four major categories[20][5]:

1.1.1 Fixed. Fixed phrases do not allow morphosyntactic variation or internal modification (*in short, by and large, such as*). These are often the simplest idioms, which present in a singular form barring any mistakes of the speaker. They are generally three words or less and whose words cannot be modified without losing the core of the definition.

1.1.2 Semi-Fixed. These phrases can undergo modification in the verb tenses sometimes but not always. (*"Kicked the bucket"*, but not *"Kicked the pail"*)

1.1.3 Syntactically Flexible. Syntactically flexible—undergo syntactic variation such as inflection (e.g., *"{look, looked, looks} up to"* or *"{wrote, writes, written} down*).

1.1.4 Institutionalized Phrases. These fully compositional phrases that are statistically idiosyncratic (*traffic light, Secretary of State*).

Unfortunately there are some such idioms similar in nature, yet which are not emphasized in these major categories. These idioms are generally resistant to change, yet creative speakers may modify them with the expectation that the listener interprets the meaning from prior knowledge of the correct meaning of the idiom phrase. (i.e. *Top Dog* could reasonably be replaced with *Bottom Dog*)

Similarly, another form of idioms are created but not mentioned when unwitting speakers mistakenly modify the subject or nominative of one idiom, with the predicate of another. An example...*I don't trust him farther than you can bat an eye.* or *Don't look a charlie horse in the mouth.* [10] The author notes that this is not an exhaustive list of the forms of idioms in the english language, but may extend to MOST of the idioms available.

The last subsection of idiom research worth mentioning here is that of the Aspectual class. Jackendoff [13] observes that, while idioms have been traditionally treated by grammarians as a relatively marginal phenomenon, there are probably as many of them as there are adjectives, and theories of linguistic structure and processing had therefore better pay heed to them [19]. Of specific note in [19], is that idioms fall under specific Aspectual classes, different from their literal translations. Consider, firstly, the verb phrase 'paint the town red', which is often used idiomatically and means, according to the Longman Dictionary of Idioms (Longman 1979), "have a very enjoyable time, esp. in a lively and noisy manner". Mary and her friends painted the town/shed red/green for/in a few hours. Compare the left side of the parenthesis for the idiomatic aspectual class, and the right side for the literal aspectual class.

1.2 Types of Idiom Phrases

Using a rudimentary scanning technique of a list of idioms, it seems plausible that idioms can be broken further into verb phrase based idioms, and another subset, to be defined further soon.

1.2.1 Verb Phrase Based Idioms.

(*give it a rest, get on someone's nerves*, or *fall from grace*)

1.2.2 Noun Phrase Based Idioms.

(*pins and needles*, or *Three score and ten*)

1.2.3 Prepositional Phrase Based Idioms.

(*on the edge of one's seat*, or *weak at the knees*)

1.3 Idiom Detection using Natural Language Processing

Idiom detection in language remains a problem for machine language without significant adaptation from human users. Some combination of Natural Language processing tools may yet identify unlabeled idioms. Until then, using an extensive repository may in fact be the only reasonably accurate solution. Previously developed tools such as NLP tools, such as lemmatizers, and part of speech taggers may be used to generate the associations necessary to at least identify possible idioms. In order to attempt idiom recognition using advanced methods, there first needs to be a baseline level. This thesis presents a method with which to achieve that baseline. It offers the first of it's kind research into idioms on modern English corpora.

2 RELATED WORKS

There were a number of early attempts to classify Idioms by hand [12][6][3][17].

An expansion on the work of idiom identification begun to be developed under the field of **Multi-Word Expressions (MWE's)**. An overview of the **MWE** field can be found here [2]. This field of study is expansive, and includes phrases of any sort which come to have a single definition formed by a collection of individual words. This stretches far beyond the specifics of idioms, and includes such notable phrases as *field of dreams*, or *President of the United States (POTUS)*. Much of the work presented on MWEs has been completed in languages other than English, such as in French Arabic [14], or Japanese [15]. These foreign studies show promising results with Supervised Context Free Parsing Models, and dependency pattern recognition; respectively.

In the years since the manual era, Part of Speech Tagging, Collocation Extraction, Verb-Particle Extraction, n-gram classification, Context-Free Parsing Model and Machine Learning have each in their time been used to Extract/Classify MWE's [2]. The trade offs between these methods are in their leveraging of human vs computing power, and emphasis on accuracy or precision. The methods which at their core rely on a dictionary or repository of MWEs tend to provide fast, accurate results. Due to the incomplete nature of such dictionaries, these methods have the significant drawback in the high likelihood of missing terms. Similarly, statistical methods, which rely on analysis of word structure, tend to promote extreme computing times for marginal results with a high rate of false positives.

According to the Multi-Word Expression survey [5],To our knowledge, only two previous studies considered MWEs in the context of statistical parsing. Nivre and Nilsson (2004) converted a Swedish corpus into two versions: one in which MWEs were left as tokens, and one in which they were grouped (words-with-spaces). They parsed both versions with a transition-based parser, showing that the words-with-spaces version gave an improvement over the baseline.

Previously, work to solve the idiom problem has centered around the direct parsing of word for word phrases from idiom repositories. Due to the nature of idioms discussed above, these methods are inherently incomplete. An extension of this method to include variations upon standard idioms may increase the effectual size of idiom recognition/tagging software.

Jon Doughty wrote extensively upon the process of Idiom extraction using Natural Language Processing in his prior thesis "QUANTIFYING THE IMPACT OF IDIOM RECOGNITION IN GENRE CLASSIFICATION" [9]. This early work provides the framework which this study will be based upon. In his Thesis, Doughty uncovered idiom recognition and the tools necessary to build a parser. His work culminated by building a repository for future researchers, parsing idioms from raw text, and a classification framework utilizing idioms as features. This parser developed in this thesis is able to recognize a subset of idioms with 95% accuracy [2].

There have been previous attempts to gather idioms and other lexical metaphors together into one location [8][10]. Each is smaller than the 9,700 Jon Doughty was able to compile. The work by Dr. Farber [18] accentuates the need to identify mixed metaphors/idioms, with prefixes from one origination, and suffixes from a different altogether.

There have been claims [7] that MWE classification is a not important for machine recognition of language. There are claims that these phrases do not significantly most language, and should therefore not be optimized for at present. However, Pecina [18] notes that, by replacing individual Part of Speech tokens with respective MWE tokens, a parser was able to change PoS classes in 7.2% of cases throughout a corpus of 376,007 sentences. These tokens could mean the difference between understanding and confusion in human machine interactions.

3 THE MODEL

This thesis presents a method to parse and identify idioms in large corpora using an n-gram verification method. The model is sourced on GitHub for future use by other researchers [16]. It can be used with any properly formatted repository of multi-word phrases.

3.1 Accessing the JSON Repository

The JSON Repository of the Doughty Corpus provides dictionary-like access to over 9000 idioms, their variations, sources of the idiom, definition, and Part-of-speech. For this study, the JSON file was parsed, and both the idioms & their variations were stored in a working variable for later recognition. The researchers recognize data was lost with this method, but accept that for the results found.

```
{
    "variations": [],
    "idiom": "",
    "sources": [""],
    "entry": [
      {
        "usages": [[]],
        "definition": "",
        "pos": ""
      }
    ],
    "id": 1,
    "confidence": 1
}
```

3.2 Restructuring for Study

Once the idioms were gathered into a list of separate strings, code was developed to sort each idiom into a form where it could be recognized by a corpus parser. In this way, a dictionary was developed as seen in the three examples below, where the first word of the idiom is retained as the key of the top level dictionary, with 'connector' and 'end' as the two values contained within. These themselves are dictionaries, and contain the individual forms of idioms whose first word is the key in the top level dictionary. With all connector words stored as their own list, this retains the separate keys and will be used within the code to ensure an idiom is actually found, as opposed to just encountering a bag-of-words jumble of idiom-like words.

abandon : 'connector': [[]], 'end': [['ship']]

about : 'connector': [[], [], []], 'end': [['time'], ['to'], ['turn']]

above : 'connector': [['and'], ['and', 'beyond', 'the', 'call', 'of'], [], ['one', "'s"], ['the'], ['the'], ['the'], [], []], 'end': [['beyond'], ['duty'], ['board'], ['bend'], ['curve'], ['law'], ['salt'], ['water'], ['yourself']]

3.3 Phrase Identification Model

In this research the Gensim MWE phrase-identifier was used as the base code[19]. For the purposes of this research, much of the Gensim model's code which worked well for identifying 2-3 word phrases which occured multiple times in the same corpus, did not work for phrases which would only occur 1-2 times in the sample corpora. For this reason, it was only possible to retain the logic the program used to identify a phrase it had already tagged as a MWE. After restructuring it work with the dictionary structure shown above, the code would generally assess each word in a corpus for a match within the top-level of the dictionary. If a match was found, the code would simply parse the rest of the sentence to verify a match with the sub-level connector and end words. If a diverson from baseline idiom structure was detected, the program reverts the potential idiom string to its base tokens. If, however, it does detect an idiom, the program returns the sentence with the idiom delimited by underscores ("_") as a single token (e.g. "pencil_in").

3.4 Asserting Idiom Structure

4 EXPERIMENTAL DESIGN

This thesis aims to use a previously compiled idiom repository test the compiled Idiom repository on first the Brown corpus, and then the Corpus of Contemporary American-English (COCA) using Natural Language Processing Techniques. In addition, it will attempt to add newly found idiom variations to Jon Doughty's JSON repository of idioms.

4.1 Testing on Brown Corpus

With the test code functioning, it was applied to the portion of the brown corpus contained within the NLTK toolkit. A slight modification was made to present the brown corpus to the test code sentence by sentence, and to count each idiom found, as they occur.

The important results are as follows:

Table 1: Brown Corpus Outputs

Idiom	Rate of Occurrence
pencil in	2
such as	2
...All others...	272

Of the 274 individual idioms within this section of the Brown Corpus, only two are repeated. While *Pencil in* would be considered an idiom by any standard, it is arguable if *Such as* should retain that same classification. This could be easily removed from the idiom corpus as it may not be considered a true idiom, but more of a relational Multi word Expression. Additionally, it is important to note that given the NLTK brown corpus' wordcount of 1161192 individual words; the brown corpus achieves a 0.02% idiom occurence rate. This is noted to compare later to the modern COCA corpus. However, neither of these findings do not detract from the main result of the test, that many idioms are indeed being found by the code in a relatively small amount of processing time.

5 RESULTS

5.1 Modifications for the COCA

The COCA dataset is arranged by type of corpus Fiction, Academic, Spoken, News, and by year 1990-2010. To process this large amount of data, the program was rearranged to pre-process each year of a single type into tokenized sentences using NLTK's sentence tokenizer, and further each sentence into word tokens, again using NLTK's word tokenizer. By splitting the COCA into four major computation groupings by type, the chances of a failure during computation was reduced. Computations for each type took roughly 45 minutes to produce, outputting the rate each idiom in the Doughty Idiom Repository was identified.

5.2 Results within the COCA

When tested on the COCA data, interesting results can be found. With the nominal output being a dictionary containing idioms matched, and frequency, many plots can be made. Displayed below are some such examples this tool can provide. In Figure 2, below, the average idiom occurrence per sentence is displayed on the y axis, with each major category of the COCA on the x axis. The average rate of idiom usage in the News/media sources occur at 0.31% over the 20 year period (1990-2010) the COCA provides data. This may be of use to researchers who wish to generate accurate sounding news media using AI. A specific number for contemporary American English is important to achieve the aesthetic of computerized language humans have evolved to be comfortable with.

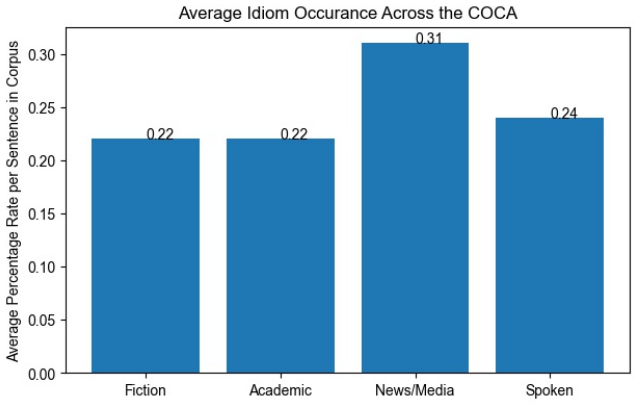

Figure 2: Idiom Usage Across the COCA

Of further note, these results establish a baseline for the COCA corpus, which continuing researchers may be capable of overcoming, by increasing the idioms present in the repository. The results above show that the idioms we tested occurred at roughly 10x the

rate in Contemporary American English as opposed to the mid-century English present when the Brown corpus was developed. This coincides with the findings on the next page, showing a rise in idiom usage as the years go on. This may also show that the relative turnover rate of idioms is high. Individual idioms may go in and out of style often, further differentiating their complexities and complicating idiom discovery, repository maintenance, and tagging.

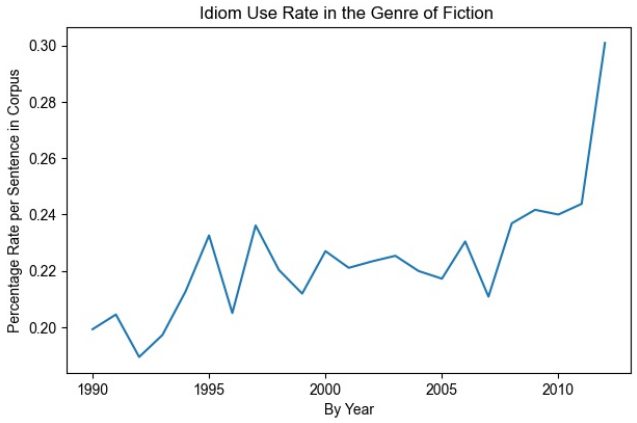

Figure 3: Results within the COCA Fiction Genre

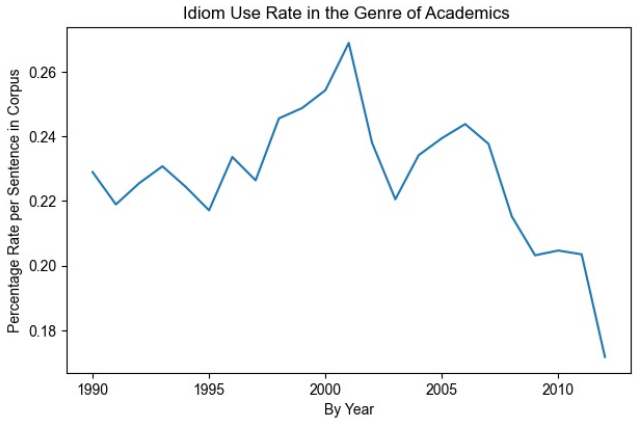

Figure 4: Results within the COCA Academic Genre

In the above figures, the results separated by type and by year are shown. While this data may not be enough to draw full correlations, initial review shows that a preliminary overton window can be developed for the field, stating no less than 1 idiom per 1000 sentences, and probably no greater than ten times our maximum 4.2 idioms per 1000 should be considered if developing a natural language AI bot to sound human-like. We suppose this large margin of error on the upward bound for variations deviating from the standard forms of idioms put into the system.

Questions remain, such as, why is there such a drastic deviation

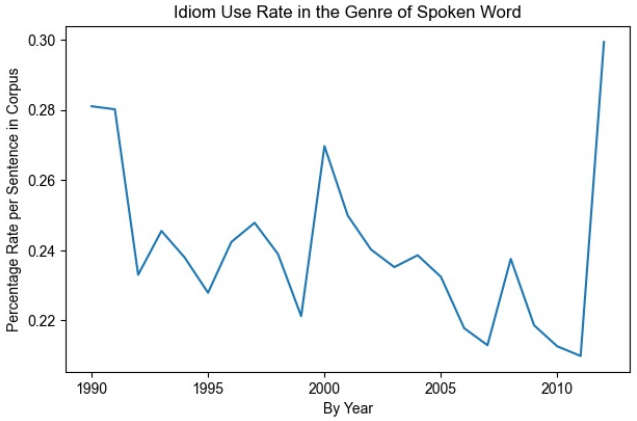

Figure 5: Results within the COCA Spoken Word Genre

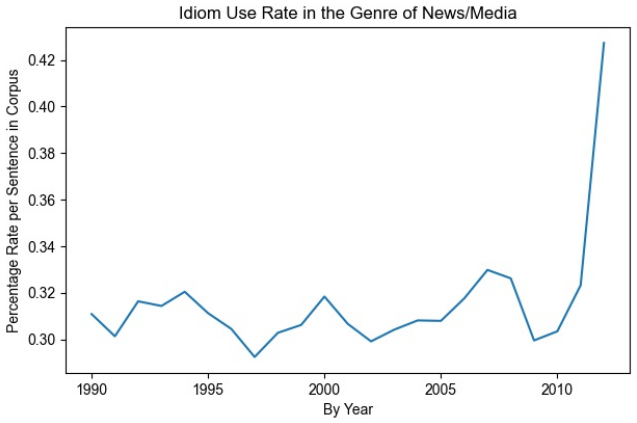

Figure 6: Results within the COCA News/Media Genre

within the News/Media section within the Corpus of Contemporary American English? More data must be analyzed to determine whether this is statistically significant, if the trend has remained to this day, or if in fact that section of the corpus simply matched well to the idiom repository by chance.

Of larger concern, is that this program can be successfully used to identify the subset of multi-word expressions within the English language, called idioms. The repository remains present to be built upon for increased scale and recognition, and the code presented in this paper proves to be a suitable testing ground for part-of-speech tagging idiomatic language in the English language.

6 CONCLUSION

Future work in this area may begin by drastically increasing the variations within the Doughty Idiom Repository, namely by lemmatizing verb forms and generalizing the pronouns present in many of the idioms (e.g. [{Striking, Struck} out on {his, her, one's} own]).

With these additions, the incidence rate of idioms may be proven to be far higher than the initial estimates presented above.

ACKNOWLEDGMENTS

A special thanks to Dr. Foaad Khosmood for his continual support in this research, as well as the practical tools, knowledge, and wisdom he provided during this research. Thanks also to Jon Doughty, for his detailed research into idioms, and the wonderful repository he has left as a memento to the field. With any luck it will be the strong base that is continued to be built upon for English language idioms.

REFERENCES

[1] Souha Ayed. 2008. *Avoidance of idioms: An ethnic group identity issue?* Ph.D. Dissertation. Concordia University.
[2] Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing* 2 (2010), 267–292.
[3] Cristina Cacciari. 1993. The place of idioms in a literal and metaphorical world. *Idioms: Processing, structure, and interpretation* 27 (1993), 55.
[4] Anna B Cieślicka. 2015. Idiom acquisition and processing by second/foreign language learners. (2015).
[5] Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* 43, 4 (2017), 837–892.
[6] J Cooper Cutting and Kathryn Bock. 1997. That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blendsß. *Memory & cognition* 25, 1 (1997), 57–71.
[7] Marie-Catherine de Marneffe, Sebastian Padó, and Christopher D Manning. 2009. Multi-word expressions in textual inference: Much ado about nothing?. In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer)*. 1–9.
[8] Ricarda Dormeyer, Ingrid Fischer, and Martina Keil. 1998. A database for verbal idioms. In *euralex*, Vol. 98. 99–109.
[9] Jon Doughty. 2017. Quantifying the Impact of Idiom Recognition in Genre Classification. (2017).
[10] David J. Farber. 1997. Farberisms. (1997).
[11] Claudia Felser, Leah Roberts, Theo Marinis, and Rebecca Gross. 2003. The processing of ambiguous sentences by first and second language learners of English. *Applied Psycholinguistics* 24, 3 (2003), 453–489.
[12] Ingrid Fischer and Martina Keil. 1996. Parsing Decomposable Idioms. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1* (Copenhagen, Denmark) *(COLING '96)*. Association for Computational Linguistics, USA, 388–393. https://doi.org/10.3115/992628.992696
[13] Sheila Glasbey. 2008. Aspectual composition in idioms. In *Recent advances in the syntax and semantics of tense, aspect and modality*. De Gruyter Mouton, 71–88.
[14] Spence Green, Marie-Catherine de Marneffe, and Christopher D Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics* 39, 1 (2013), 195–227.
[15] Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. 2006. Japanese idiom recognition: Drawing a line between literal and idiomatic meanings. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. 353–360.
[16] Chandler Jones. 2021. Gensim Idiom Adaptation. https://github.com/chandler150/Idiom-Parser.
[17] Mark Lee and John Barnden. 1999. Mixing metaphors. *arXiv preprint cs/9904004* (1999).
[18] Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, Vol. 2008. Citeseer, 54–61.
[19] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.
[20] Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *International conference on intelligent text processing and computational linguistics*. Springer, 1–15.