Like Finding a Needle in a Haystack: Annotating the American National Corpus for Idiomatic Expressions

Laura Street, Nathan Michalov, Rachel Silverstein, Michael Reynolds, Lurdes Ruela, Felicia Flowers, Angela Talucci, Priscilla Pereira, Gabriella Morgan, Samantha Siegel, Marci Barousse, Antequa Anderson, Tashom Carroll, and Anna Feldman

Montclair State University
1 Normal Avenue, Montclair, New Jersey, 07043, USA
{streetl1, michalovn1, silversteir2, feldmana}@mail.montclair.edu

Abstract
This paper presents the details of a pilot study in which we tagged portions of the American National Corpus (ANC) for idioms composed of verb-noun constructions, prepositional phrases, and subordinate clauses. The three data sets we analyzed included 1,500-sentence samples from the spoken, the non-fiction, and the fiction portions of the ANC. This paper provides the details of the tagset we developed, the motivation behind our choices, and the inter-annotator agreement measures we deemed appropriate for this task. In tagging the ANC for idiomatic expressions, our annotators achieved a high level of agreement (> .80) on the tags but a low level of agreement (< .00) on what constituted an idiom. These findings support the claim that identifying idiomatic and metaphorical expressions is a highly difficult and subjective task. In total, 135 idiom types and 154 idiom tokens were identified. Based on the total tokens found for each idiom class, we suggest that future research on idiom detection and idiom annotation include prepositional phrases as this class of idioms occurred frequently in the nonfiction and spoken samples of our corpus.

1. Introduction
Researchers have been investigating idioms and their properties for many years. According to traditional approaches, an idiom is “in its simplest form...a string of two or more words for which meaning is not derived from the meanings of the individual words comprising that string” (Swinney and Cutler, 1979: 523). As such, the meaning of to kick the bucket (which is ‘to die’) cannot be obtained by breaking down the idiom and analyzing the meanings of its constituent parts, to kick and the bucket. In addition to being influenced by the principle of compositionality, the traditional approaches are also influenced by theories of generative grammar (Flores d’Arcais, 1993: 80-82; Langlotz, 2006: 15-16). The properties that traditional approaches attribute to idiomatic expressions are also the properties that make them difficult for generative grammars to describe. For instance, idioms can be syntactically ill-formed (e.g., by and large), resistant to grammatical transformations (e.g., the bucket was kicked by him ≠ ‘he died’), impervious to lexical substitutions (e.g., to kick the pail ≠ ‘to die’), and semantically ambiguous without context (e.g., Charles kicked the bucket and was buried last night vs. Charles kicked the bucket through the window).

In recent years, post-generative approaches to idioms have argued that idioms, like all linguistic expressions, are a type of construction. As Goldberg (2003: 219) states: “Constructions are pairings of form and function, including morphemes, words, idioms, partially lexically filled and fully general linguistic patterns.” To be more precise, a construction is any linguistic pattern where an aspect of its form or function is unpredictable based on its component parts or other known constructions. When conducting research on idiomatic expressions, a constructionist approach is advantageous because it seeks to describe language in general. That is to say, it does not disregard unusual linguistic phenomena (such as idioms) because they are ‘peripheral’ rather than part of the ‘core’ of language. Constructionist approaches are also beneficial because they claim that constructions are acquired based on their frequency in the input and this works well with a corpus-based approach where data and frequency counts are used to capture properties of linguistic phenomena such as collocations, multi-word expressions, etc. For a more detailed discussion of the approach we take to idioms in this paper, please refer to (Goldberg, 2003).

<table>
<thead>
<tr>
<th>Constructions</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morpheme</td>
<td>pre-, -ing</td>
</tr>
<tr>
<td>Word</td>
<td>avocado, anaconda, and</td>
</tr>
<tr>
<td>Complex Word</td>
<td>dare-devil, shoo-in</td>
</tr>
<tr>
<td>Idiom</td>
<td>going great guns</td>
</tr>
<tr>
<td>Idiom (Partially Filled)</td>
<td>jog (someone’s) memory</td>
</tr>
<tr>
<td>Covariational Conditional</td>
<td>the more you think about it, the less you understand (Form: The Xer, the Yer)</td>
</tr>
<tr>
<td>Ditransitive</td>
<td>he gave her a Coke, he baked her a muffin (Form: Subj [V Obj1 Obj2])</td>
</tr>
<tr>
<td>Passive</td>
<td>the armadillo was hit by a car (Form: Subj aux VPpp (PPby))</td>
</tr>
</tbody>
</table>

Table 1. Examples of constructions (adapted from Goldberg 2003: 220)

Previous work on automatic idiom classification has typically been of two types: those which make use of type-based classification methods (Lin, 1999; Baldwin et al. 2003; Fazly and Stevenson, 2006; Bannard, 2007; Fazly et al., 2009) and those which make use of token-based classification methods (Birke and Sarkar,
English with which we are most familiar. Furthermore, while the ANC Second Release has already been annotated for word and sentence boundaries, parts-of-speech, and noun and verb chunks, it does not yet include any annotations for figurative language.¹

The ANC Second Release includes data from a variety of domains ranging from government, academic, and technical writing to journal articles, fiction, blogs, magazines, and spoken interviews. Although the corpus as a whole will not be considered balanced until it is complete,² we selected a sample that we believe represents a wide range of idiomatic language. Approximately one third of our sample data comes from the written nonfiction domain, one third comes from transcribed spoken narratives, and one third comes from written fiction. In a trial run of the study, the annotators identified and tagged idioms in a 111-sentence sample (1644 tokens) according to the tagset elaborated below. After the trial run, the annotators discussed the numbers and types of idiomatic expressions they found and re-evaluated the tagset, ultimately deciding to use the same tagset for the pilot study as they had used for the trial run. For the pilot study, the annotators tagged a 4,500-sentence sample in the same manner as the trial run. The annotators worked individually, but three or four were assigned to each sample. The task was completed in roughly six to ten hours.

To expand upon the work of previous studies and to facilitate future work in the area of automatic idiom detection, we developed a tagset comprised of three tags. Our tagset included tags for idioms composed of verb-noun constructions (VNCs), prepositional phrases (PPs), and subordinate clauses (SCs). We included less prototypical structures such as PPs and SCs because one of our goals was to illustrate the many cases of figurative language that are missed by approaches that focus exclusively on idioms containing select verbs.

<table>
<thead>
<tr>
<th>Tags</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;PP&gt; … &lt;/PP&gt;</td>
<td>at death’s door, to a fault</td>
</tr>
<tr>
<td>&lt;VN&gt; … &lt;/VN&gt;</td>
<td>pass the buck, learn the ropes</td>
</tr>
<tr>
<td>&lt;SC&gt; … &lt;/SC&gt;</td>
<td>when hell freezes over, if the price is right</td>
</tr>
</tbody>
</table>

Table 2. The tagset with examples

In an effort to make the tagging process more efficient and reliable, we limited our tagset to a very small number of tags and gave our annotators the following guidelines: 1.) Always tag for the narrowest scope possible, 2.) Use the same syntax throughout the annotation process, and 3.) Do not include punctuation such as periods or quotation marks.

¹ The ANL Second Release includes data from a variety of domains ranging from government, academic, and technical writing to journal articles, fiction, blogs, magazines, and spoken interviews. Although the corpus as a whole will not be considered balanced until it is complete, we selected a sample that we believe represents a wide range of idiomatic language. Approximately one third of our sample data comes from the written nonfiction domain, one third comes from transcribed spoken narratives, and one third comes from written fiction. In a trial run of the study, the annotators identified and tagged idioms in a 111-sentence sample (1644 tokens) according to the tagset elaborated below. After the trial run, the annotators discussed the numbers and types of idiomatic expressions they found and re-evaluated the tagset, ultimately deciding to use the same tagset for the pilot study as they had used for the trial run. For the pilot study, the annotators tagged a 4,500-sentence sample in the same manner as the trial run. The annotators worked individually, but three or four were assigned to each sample. The task was completed in roughly six to ten hours.

To expand upon the work of previous studies and to facilitate future work in the area of automatic idiom detection, we developed a tagset comprised of three tags. Our tagset included tags for idioms composed of verb-noun constructions (VNCs), prepositional phrases (PPs), and subordinate clauses (SCs). We included less prototypical structures such as PPs and SCs because one of our goals was to illustrate the many cases of figurative language that are missed by approaches that focus exclusively on idioms containing select verbs.

<table>
<thead>
<tr>
<th>Tags</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;PP&gt; … &lt;/PP&gt;</td>
<td>at death’s door, to a fault</td>
</tr>
<tr>
<td>&lt;VN&gt; … &lt;/VN&gt;</td>
<td>pass the buck, learn the ropes</td>
</tr>
<tr>
<td>&lt;SC&gt; … &lt;/SC&gt;</td>
<td>when hell freezes over, if the price is right</td>
</tr>
</tbody>
</table>

Table 2. The tagset with examples

In an effort to make the tagging process more efficient and reliable, we limited our tagset to a very small number of tags and gave our annotators the following guidelines: 1.) Always tag for the narrowest scope possible, 2.) Use the same syntax throughout the annotation process, and 3.) Do not include punctuation such as periods or quotation marks.
3. Measuring Inter-Annotator Agreement & Reliability

To ensure reliability of the results two alpha scores were calculated for each of the three sections using the following variation of Krippendorff's Alpha:\(^1\)

\[
\alpha = 1 - \frac{(n-1)\sum O_{ck} - \sum \frac{n_c(n_c-1)}{n(n-1)}}{D_e}
\]

Where Do represents observed disagreement among the annotators, De is the disagreement which would be expected to occur if tags had been selected by chance. \(\sum\) indicates summation of all data of the given type, \(n\) represents the number of items associated with the given subscript (where no subscript is present \(n\) indicates the total number of tags), and subscripts relate to the coincidence matrices which are described in more detail below.

Krippendorff’s Alpha was used because it was appropriate for our data set. It is specifically designed to analyze nominal data sets generated by multiple annotators, including instances where one or more annotators did not provide a rating for some of the units in the data set. While Fleiss’ Kappa and Cohen’s Kappa are commonly used to analyze inter-annotator agreement, they were inappropriate in the current study as Cohen’s Kappa can only be used with two annotators and Fleiss’ Kappa cannot be used when data sets contain units that are not rated by all of the annotators.

The first alpha evaluated annotator reliability in the detection of idioms within the corpus and accounted for annotator reliability regarding the status of a phrase as idiomatic or not. The second alpha determined reliability in the categorization of detected idioms. In cases where annotators disagreed on the scope of idioms and overlapping tags were noticed, these tags were grouped together for alpha scoring and noted for annotators. Given the size of the corpus and the fact that tag lists were obtained by machine search, it is possible that not all of the overlapping tags were noticed, which could have affected the preliminary alpha score. Given that subsequent modifications were made manually by the annotators, however, such artifacts are not present in the final corpus and all tags agree in all relevant domains.

The specific methodology of the alpha scoring was to form two tables of annotator responses. The first table (Table 3) was used to generate the first alpha score. It contains all the phrases (units) that were tagged as idioms along with their corresponding tags. The second table (Table 4) was used to generate the second alpha score. It contains all the phrases (units) that were tagged as idioms along with their corresponding tags. (Note: \(\cdot\) indicates that the annotator did not mark the phrase as being an idiom.) The final row records the total number of tags for a given phrase, and the final column records the number of each type of tag a given idiom received. Mu is the total number of tags and is present only in the second table as the reliability of idiom categorization can only be determined in cases where annotators have marked a statement as a type of idiom. In the first case, all annotators will have marked any given phrase as idiomatic or not, so Mu will always equal the number of annotators and its inclusion in the table would be redundant.

![Table 3. Table of phrases marked as idioms](image)

<table>
<thead>
<tr>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator A</td>
<td>Idiom</td>
<td>Idiom</td>
<td>Non-Idiom</td>
</tr>
<tr>
<td>Annotator B</td>
<td>Non-Idiom</td>
<td>Idiom</td>
<td>Non-Idiom</td>
</tr>
<tr>
<td>Annotator C</td>
<td>Idiom</td>
<td>Idiom</td>
<td>Idiom</td>
</tr>
</tbody>
</table>

![Table 4. Presents tags assigned to each given phrase.](image)

<table>
<thead>
<tr>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator A</td>
<td>VN</td>
<td>SC</td>
<td>PP</td>
</tr>
<tr>
<td>Annotator B</td>
<td>VN</td>
<td>SC</td>
<td>PP</td>
</tr>
<tr>
<td>Annotator C</td>
<td>VN</td>
<td>PP</td>
<td>PP</td>
</tr>
<tr>
<td>Number of values in unit (Mu)</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Data within the tables were then used to generate coincidence matrices of the form:

\[
\begin{matrix}
1 & 0 & \ldots \\
0 & 1 & 0 \\
\vdots & \ddots & \ddots \\
0 & 0 & 1 \\
\end{matrix}
\]

![Figure 1. Sample matrix used to calculate the number of tag pairs that occur.](image)

Where Ock is determined using the following formula:

\[
O_{ck} = \sum_{U} \frac{\text{Number of c-k pairs in unit } u}{m_{u} - 1}
\]

---

\(^1\) Information regarding Krippendorff’s Alpha can be found at Klaus Krippendorff’s website: http://www.asc.upenn.edu/usr/krippendorff/dogs.html
For example, if a given phrase was tagged as a VN by two annotators and a PP by a third, there would be 6 possible pair combinations of the given tags: 2 VNPP, 2PPVN, and 2VNVN pairs. Note that in calculating ck pairs, the ordering matters and PPVN is not the same as VNPP. The number of pairs of each type is then normalized by dividing it by one less than the total number of tags for the unit. In the given example each pair would be divided by 2 (3 tags - 1) to create a value of 1VNPP, 1PPVN, and 1VNVN. The matrix then displays the total number of normalized tag pairs of each type. Matrices figures are then used in the alpha formula.

\[
\begin{array}{c|ccc|c}
\text{PP} & \text{VN} & \text{SC} & \text{Total} \\
\hline
\text{PP} & 2 & 4 & 1 & 7 \\
\text{VN} & 4 & 10 & 3 & 17 \\
\text{SC} & 1 & 7 & 3 & 11 \\
7 & 17 & 11 & 35 \\
\end{array}
\]

Figure 2. An example of how the matrix appears for tag pairs

The alpha score is obtained by calculating the difference of observed disagreement over expected disagreement, or the amount of agreement that would be expected if tags were assigned at random, subtracted from one (1 – (Do/De)) so that perfect annotator agreement would result in \( \alpha = 1 \) (1 – (0/1) = 1) and if annotator agreement were equal to chance \( \alpha = 0 \) (1 – (1/1) = 0). Specifically the form of the alpha for nominal data used here is obtained by subtracting the total possible number of agreeing tag pairs given the annotations made (\( \sum \text{occ} \)) from the actual number of agreeing tag pairs made by the annotators (\( \sum \text{occ} \)), and then multiplying by one less than the total number of tags in agreement (n-1). This number is then divided by the difference obtained by subtracting the number of possible tag pairs given the number of each tag type (again (n-1)\( \sum \text{occ} \)) from the maximum possible tag pair agreement for the overall number of tags (n(n-1)). Put more generally, the formula can be seen as calculating reliability in the following way:

\[
(\text{# of actually agreeing tag pairs}) - (\text{# of possible pairs given tags}) \\
(\text{Maximum # agreeing pairs given tags}) - (\text{# of possible pairs given tags})
\]

An alpha of \( \geq .80 \) is considered reliable.

4. Results & Discussion

4.1 Inter-Annnotator Agreement

Alpha scores for each section were as follows:

**Spoken:**
- Tag agreement: 1.0
- Idiom agreement: -0.34

**Non-Fiction:**
- Tag agreement: 0.85
- Idiom agreement: -0.199

**Fiction:**
- Tag agreement: 0.83
- Idiom agreement: -0.28

For each section of the annotated sample, alpha scores for tag agreement exceeded the .80 threshold of reliability, indicating that annotators agreed reliably and significantly above chance upon the types of idioms tagged. Tag agreement was most notable in the section of the sample drawn from spoken narratives, where a 1.0 alpha score indicates 100% agreement among annotators with regard to the type of tag. That is, when an idiomatic expression such as drawing a blank occurred, all annotators agreed that the idiom constituted a VN construction. Annotators also noted a few constructions for which they did not have tags, in particular suggesting that future research include a tag for verb + prepositional phrase constructions (VPPCs) such as walking on air.

Unlike the alpha scores for tag agreement, the alpha scores for idiom agreement were all very low, indicating that the level of agreement among annotators regarding which phrases were actually idioms was no better than chance. In fact, there were many instances in which only one annotator tagged a particular construction as an idiomatic expression. One possible explanation for the low inter-annotator agreement could be, as some studies suggest (Gibbs 1984), that there is no clear psychological distinction between literal and figurative expressions, so annotating for idioms is a rather difficult and subjective task, and the level of disagreement our annotators demonstrated on literal-idiom expressions is to be expected. The fact that we did not give our annotators a preset list of idioms to look for (as is usually done in corpus-based research on idioms) might have made the task more difficult as well.

Because the inter-annotator agreement for each idiom list was low, each group of annotators held a meeting following the analysis of the results to create a finalized list of idioms to be used as the official idioms for the corpus. The annotated corpus is currently available on Montclair State’s server and will soon be available on the following website as well: http://netdrive.montclair.edu/~street11/index.htm. We hope that the tagset and the methodology we have developed for this pilot study can serve as the basis for developing a more complete tagset for annotating the entire ANC.

4.2 Type and Token Counts

In total, 135 idiom types were found in the corpus with a total of 154 idiom tokens appearing. The following tables provide a breakdown of the idioms found within each section of the corpus. Since some types were found in more than one portion of the corpus, the types given sum to 140 rather than 135.
Variations between fiction, agreement (< .00) on what his class of here were so few instances of SCs, we a novel, text identifying idiomatic and metaphorical expressions counted as an idiom. These findings support the claim that the tags but a low level of a annotators achieved a high level of agreement (> .80) on tagging the ANC for idiomatic expressions, our non-included 1,500 subordinate clauses. The three data verb Corpus (ANC) were tagged for idioms composed of annotators noted higher instances of them while tagging. particular, verb + prepositional phrase constructions prioritize other types of constructions over SCs. In another major finding was that 79% of the idioms tagged were VNCs (n = 122). What was surprising, however, was that VNCs and PPs occurred in much closer numbers in both the nonfiction (n = 10, n = 14) and the spoken samples (n = 19, n = 12). PPs also occurred roughly 4.3 times more in the nonfiction and spoken samples than in the fiction sample. These findings indicate that many idioms are missed when researchers annotate texts solely for VNCs—particularly if those texts come from genres other than fiction. One possible explanation for why VNCs outnumber PPs in the fiction domain is that novel figurative language, which tends to occur more frequently in fiction, is likely to take the form of a VNC rather than a PP. More research needs to be done, however, to explain why VNCs occur more frequently than PPs in this domain. Finally, since there were so few instances of SCs, we recommend that future research on idiom detection prioritize other types of constructions over SCs. In particular, verb + prepositional phrase constructions (VPPCs) appear to be a good candidate as several of our annotators noted higher instances of them while tagging.

5. Conclusion

In a pilot study, portions of the American National Corpus (ANC) were tagged for idioms composed of verb-noun constructions, prepositional phrases, and subordinate clauses. The three data sets we analyzed included 1,500-sentence samples from the spoken, the non-fiction, and the fiction portions of the ANC. In tagging the ANC for idiomatic expressions, our annotators achieved a high level of agreement (> .80) on the tags but a low level of agreement (< .00) on what counted as an idiom. These findings support the claim that identifying idiomatic and metaphorical expressions in a text is a highly difficult and subjective task. In total, we identified 135 idiom types and 154 idiom tokens. Based on the total tokens found for each idiom class, we suggest that future research on idiom detection and idiom annotation include prepositional phrases. This class of idioms occurred frequently in the nonfiction and spoken portions of our corpus.

6. Appendix

The final list of idioms for all three groups of annotators:

- at that point
- across history
- are spreading out
- as melted architecture
- at face value
- at the top of his lungs
- be a tall order
- be able to bring
- be better off
- be carried out for anyone
- be totally out of character
- behind-the-scenes
- being saddled with a wrong name
- being torn apart
- blurring the edges
- bugs me
- by the same token
- by the wayside
- call him on it
- calls to mind
- came across a passage
- came pat
- camp up
- can swing that
- can't figure out
- claim any deep roots
- come to terms with it
- come up with all this stuff
- conjures up an image
- conquered his fear
- cooling down
- count on your continued partnership
- cracked it
- cried out on that topic
- dawned on me
- deal with it
- depend on me
- didn't get it
- drawing a blank

<table>
<thead>
<tr>
<th>Fiction Types</th>
<th>Fiction Tokens</th>
<th>Nonfiction Types</th>
<th>Nonfiction Tokens</th>
<th>Spoken Types</th>
<th>Spoken Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>VNC 10</td>
<td>10</td>
<td>PP 16</td>
<td>19</td>
<td>SC 14</td>
<td>14</td>
</tr>
<tr>
<td>PP 85</td>
<td>93</td>
<td>14</td>
<td>14</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>SC 0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td><strong>∑</strong> 88</td>
<td>96</td>
<td>25</td>
<td>25</td>
<td>27</td>
<td>33</td>
</tr>
</tbody>
</table>

Table 5. Type and token counts for each sample

<table>
<thead>
<tr>
<th>∑ Tokens</th>
<th>VNC</th>
<th>PP</th>
<th>SC</th>
<th><strong>∑</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>122</td>
<td>29</td>
<td>3</td>
<td>154</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Total token counts for each idiom type
fell on the vermicelles
fess up to how things had developed
figure out
find your way around
fished around
foot a bill
for anything in the world
freeze with horror
from point to point
from the bottom of my heart
from time to time
gave himself back
get off his butt
get off my back
get over it
get round the issue
get through it
get your goat
give you a hand
go on about that
go to great lengths
go up
going my own way
grabbed my attention
grew out of a popular mass movement
had a gift
had my share
have a clue
have a foot to stand on
have his memory in common
have the same connection to him
having such an impact
hear about it
held our attention
higher in the feudal chain
hit me
hit me
if I don't make it
in a related vein
in the right direction
in trouble
into boys
is a world away
keeled over
keep them from selling it
keep up to his example
keep your head high
land us in trouble
learned first hand
left for dead
7. References


