Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 279–287
Marseille, 11–16 May 2020

MAGPIE: A Large Corpus of Potentially Idiomatic Expressions

Hessel Haagsma, Johan Bos, Malvina Nissim
CLCG, University of Groningen
Oude Kijk in 't Jatstraat 26, Groningen, The Netherlands
{hessel.haagsma, johan.bos, m.nissim}@rug.nl

Abstract

Given the limited size of existing idiom corpora, we aim to enable progress in automatic idiom processing and linguistic analysis by creating the largest-to-date corpus of idioms for English. Using a fixed idiom list, automatic pre-extraction, and a strictly controlled crowdsourced annotation procedure, we show that it is feasible to build a high-quality corpus comprising more than 50K instances, an order of a magnitude larger than previous resources. Crucial ingredients of crowdsourcing were the selection of crowdworkers, clear and comprehensive instructions, and an interface that breaks down the task in small, manageable steps. Analysis of the resulting corpus revealed strong effects of genre on idiom distribution, providing new evidence for existing theories on what influences idiom usage. The corpus also contains rich metadata, and is made publicly available.

Keywords: idiom corpus, crowdsourcing, idiomatic expressions, language resources

1 Introduction

Idiomatic expressions are an established research topic within NLP, but progress has been hampered by a lack of large annotated corpora. Existing corpora cover less than 5,000 instances of less than 100 different idiom types, which means they do not provide a good training and testing ground for methods aimed at capturing the phenomenon of idiom as a whole. Idioms are a very large, constantly changing, fuzzy-boundaried category of expressions, which cannot be completely captured in a corpus by its very nature. However, the bigger and more varied the corpus, both in number of different idiom types and the number of idiom instances, the more likely it is that conclusions drawn from this corpus are valid for 'idioms' as a whole.

Larger corpora have indeed benefits from both the descriptive and computational linguistic perspectives. For the first, they allow for better evaluation of assumptions about idiomatic expressions regarding their distribution, frequency, variation, and more. For natural language processing purposes, they allow for more reliable, more fine-grained, and more representative evaluation of idiom disambiguation tools, and for better training of such tools, given the possibility of using more data-greedy machine learning tools.

Due to the desired size of the corpus, we could not rely on the tried and tested method of using expert annotators to label the potentially idiomatic expressions. Instead, we used crowdsourcing to get the annotations from non-experts. While this is an established and reliable method for large-scale annotation, crowdsourcing has not been applied to idioms specifically, likely because of the significant difficulty of the task and the resulting difficulty of instructing crowdworkers to perform the task reliably. We were however encouraged by related work by Kato et al. (2018), who use crowdsourcing for a similar annotation task (verbal multiword expressions, including some idioms), showing that annotators agree in approximately 67% of cases.

In this paper, we describe the whole corpus creation, namely the selection of idiom types and data to annotate, the crowdsourcing setup, procedure and difficulties, as well as a detailed analysis of the resulting corpus.

Contributions This paper describes the construction of the, to our knowledge, largest available corpus of idioms.[1] In addition to the practical advantage deriving from creating this resource, we we also able to answer the following research questions:

1. Is crowdsourcing a suitable method for large-scale, high-quality annotation of a large variety of potentially idiomatic expressions?
2. To what extent can existing assumptions and theories about the distribution and frequency of idioms be verified using a corpus of this size?
3. How does genre interact with sense distributions and annotation difficulty?

Terminological clarification The ambiguity of phrases like *wake up and smell the coffee* poses a terminological problem. Usually, these phrases are called *idiomatic expressions*, which is suitable when they are used in an idiomatic sense, but not so much when they are used in a literal sense. Therefore, we propose a new term: *Potentially Idiomatic Expressions*, or *PIEs* for short. The term *potentially idiomatic expression* refers to those expressions which can have an idiomatic meaning, regardless of whether they actually have that meaning in a given context. So, *see the light* is a PIE in both 'After another explanation, I finally saw the light' and 'I saw the light of the sun through the trees', while it is an idiomatic expression in the first context, and a literal phrase in the latter context.

2 Related Work

In this section, we discuss existing idiom corpora and the variety of approaches used in their creation. The idiom corpora contain both literal and idiomatic occurrences of idiomatic expressions, and are labelled with their meaning. The four biggest of these are discussed here in detail.

[1] The corpus is publicly available at `github.com/hslh/magpie-corpus`

Name	Types	Instances	Senses	Syntax Types
VNC-Tokens	53	2,984	3	V+NP
Gigaword	17	3,964	2	V+NP/PP
IDIX	52	4,022	6	V+NP/PP
SemEval-2013	65	4,350	4	unrestricted

Table 1: Overview of existing corpora of potentially idiomatic expressions and sense annotations for English. The syntax types column indicates the syntactic patterns of the idiom types included in the dataset.

There are four sizeable corpora of idiom annotations for English, an overview of which is presented in Table 1. The table includes the number of different idiom types in the corpora (i.e. different expressions, such as *sour grapes* and *speak of the devil*), the number of PIE instances, the number of different senses annotated (e.g. *idiomatic*, *literal*, and *unclear*), and the 'syntactic type' of the expressions covered. Syntactic type means that, in some cases, only idiom types following a certain syntactic pattern were included, e.g. only verb-(determiner)-noun combinations such as *hold your fire* and *see stars*.

In general, there is large variation in corpus creation methods, regarding PIE definition, extraction method, annotation schemes, base corpus, and PIE type inventory. Depending on the goal of the corpus, the amount of deviation that is allowed from the PIE's dictionary form to the instances can be very little (Sporleder and Li, 2009), to quite a lot (Sporleder et al., 2010). The number of PIE types covered by each corpus is limited, ranging from 17 to 65 types, often limited to one or more syntactic patterns. The extraction of PIE instances is usually done in a semi-automatic manner, by manually defining patterns in a text or parse tree, and doing some manual filtering afterwards. This works well, but an extension to a large number of PIE types (e.g. several hundreds) would also require a large increase in the amount of manual effort involved. Considering the sense annotations done on the PIE corpora, there is significant variation, with Cook et al. (2008) using only three tags (*idiomatic*, *literal*, *unclear*), whereas Sporleder et al. (2010) use six (*idiomatic*, *literal*, *both*, *unclear*, *embedded*, *meta-linguistic*). As for our approach, we allow a large amount of deviation from the PIE's dictionary form, use a large amount of almost 2,000 PIE types with no restriction on syntactic pattern, automatic pre-extraction, and five sense labels.

As far as we know, crowdsourcing has not been utilised for creating an idiom corpus before. However, there is closely related work by Kato et al. (2018), who create a corpus of verbal multiword expressions, a group which includes idioms as well, but is a lot broader, incorporating particle verbs, collocations, and other types of set phrases. Kato et al. extract all instances of a set of MWE types taken from Wiktionary from part of the OntoNotes corpus (Hovy et al., 2006). Since simple extraction based on words can yield a lot of noise, i.e. non-instances, they refine those extractions based on the gold-standard part-of-speech tags and parse trees that are present in the OntoNotes corpus. Most novel, however, is their use of crowdsourcing for distinguishing between literal equivalents of MWE phrases like *get up* in 'He gets up early' and actual MWE instances like in 'He gets up a hill'. They frame the task as a sense annotation task, asking crowdworkers to label instances as either literal, non-literal, unclear, or 'none of the above'. Using this procedure, they create a corpus of 7,833 verbal MWE instances, of 1,608 different types.

3 Method

Overall, the procedure for building the corpus is as follows: we select a set of idioms from a dictionary (Section 3.1), and use a system to extract all forms of this idiom type from a base corpus (Section 3.2). The resulting candidates are annotated using crowdsourcing, and the obtained annotations are aggregated and converted into a corpus of sense-annotated PIEs (Section 4).

3.1 Collecting Idioms from Dictionaries

We use three electronic dictionaries: Wiktionary,[2] the Oxford Dictionary of English Idioms (ODEI, (Ayto, 2009)), and UsingEnglish.com (UE).[3] Taking the intersection of the idiom inventories of these three dictionaries yields the highest precision but only a limited number of 591 expressions. Therefore, we compromise a little bit on precision to get a higher number of expressions. We use all expressions which occur in ODEI and in either UE or Wiktionary, i.e. $(Wiktionary \cap ODEI) \cup (UE \cap ODEI)$. This yields a list of 2,007 idiom types.

3.2 Candidate Extraction from Corpora

As base corpus for our idiom annotation, we select the British National Corpus (BNC Consortium, 2007). This corpus has many benefits, such as its size (large enough to get sufficient idiom instances), its variety (many different genres are included, which hopefully leads to more varied idiom types and forms), and its standardised format (no noisy data, even in the transcribed spoken part). The fact that it only contains British English also means that it aligns well with the main source of idioms, ODEI, which is focused on British idioms.

We use the pre-extraction system described in Haagsma et al. (2019) to select candidate phrases for annotation. The pre-extraction system provides wide coverage, with the capacity to extract both morphological and syntactic variants of the PIE. It is robust against the most common modifications of the PIE, e.g. through word insertions (*spill all the beans*), passivisation (*the beans were spilled*), and article variation (*spill beans*). To achieve both high-precision and high-recall extraction, the combination of the regular parser-based system and the case-sensitive no-intervening-words inflection-based system is used, which achieved an F1-score on PIE extraction of almost 91%. We value a balance between precision and recall over maximising recall, since a lower precision would lead to annotators being overloaded with non-PIEs, complicating the annotation task. Applying this combined system to the whole of the BNC yields just over 200,000 instances. Some heuristics

[2] en.wiktionary.org

[3] www.usingenglish.com/reference/idioms

are applied to increase the quality of this set of instances and to whittle down its size. Instead of using the original idiom list, we manually refine it, based on the following criteria:

- Limit the number of false extractions that would be triggered by some expressions. These are cases with little lexico-semantic content like *to go*, *at it*, and *have a go*.
- Obey our working definition of idiom: *a conventionalised multiword expression, which is to some extent lexically fixed and semantically non-compositional.*

Filtering these instances out leaves 1,967 types, which greatly decreases the number of instances. Indeed, by excluding 40 idiom types, the number of annotation candidates drops by over 80,000. This means that the excluded types were highly frequent, due to an extreme number of false extractions (e.g., *to go*).

Finally, we consider downsampling highly frequent idiom types. The reason for this is that these types tend to be frequent either because they have a large number of false extractions, or because they occur very frequently, but only in their literal sense. Moreover, we want our corpus to be useful for generalised idiom disambiguation, that is, idiom disambiguation that works for all idiom types. This means that we prioritise having more idiom types over having a huge number of instances per type, which would only benefit per-type classifiers. Limiting the number of instances per type allows us to annotate more different types with the same annotation budget. We settle on a maximum of 200 instances per type as a compromise between type coverage and overall corpus size, leaving us with 72,713 instances for annotation. The type coverage of this set is excellent, with 1,781 of the 1,967 types occurring at least once (90.54%) and 1,153 types occurring at least 10 times (58.62%).

In addition to the BNC data, we use a smaller additional corpus: the Parallel Meaning Bank (PMB, Abzianidze et al. (2017)). The PMB is a multilingual corpus with many annotation layers, resulting in a fine-grained meaning representation for each text in the corpus. We include this corpus for two reasons. For one, its multilingual and parallel nature allows for a future extension of the idiom detection and sense annotations to translated texts. Secondly, detecting and annotating idioms in this corpus allows for research on how to represent the meaning of idioms in a deep semantic framework.

Pre-extraction from the PMB is done in a separate step from the BNC extraction, but the annotation procedure is the same. Pre-extraction was done using the same selection of idiom types as earlier, but with one additional filtering step. Because the PMB contains many very short documents, there might not be enough context to disambiguate the PIEs. Therefore, we extract only PIEs with at least 50 characters of context, i.e. the document should be at least 50 characters longer than the span of the PIE. The resulting set of PIEs contains 2,560 instances across 598 types. No downsampling was applied. Eventually, the total number of instances to be crowd-annotated amounts to 75,273.

4 Annotation Procedure

We make use of the FigureEight crowdsourcing platform for annotation.[4] We see crowdsourcing as the only feasible method for creating a corpus of this size, in contrast to existing corpora, which were all created by experts. Crowdsourcing does pose some challenges, especially for a relatively hard task like this one. Given the set of candidate instances from the pre-extraction system, the challenge is to get high-quality sense annotations at manageable costs of time and money using the FigureEight platform. Therefore, we strive to make the task as easy as possible for crowdworkers.

The basic setup of the task is three-tiered (see Figure 1). Given a highlighted PIE instance in context, the annotators are asked to make a three-way decision: whether the instance is used idiomatically, literally, or in a way that does not fit the binary distinction. By only asking about non-binary senses in a subquestion, the initial decision is kept as simple as possible. The dictionary definition of the idiom, extracted from the ODEI, is available on mouse-over of the highlighted text below the question. This is only displayed on-demand, in order to prevent information overload.[5] This constitutes the first tier of the annotation.

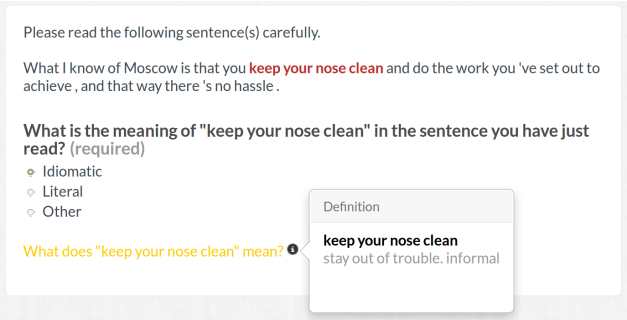

Figure 1: Annotation interface presented to crowdworkers, with definition tooltip displayed.

The second tier of annotation is only triggered when the 'Other'-sense is selected in the first tier (see Figure 2). In this case, annotators are asked whether the instance is a false extraction ('Not an instance of ...'), whether the given context is insufficient to interpret the instance ('Unclear') or whether it can be interpreted but simply does not fit the binary sense distinction ('Non-standard usage'). Since this latter category is a 'miscellaneous'-category that can contain anything unforeseen, it triggers the third tier. The third tier is a plain text input asking annotators to describe the instance's usage or meaning in the context. For example, this category contains idioms used meta-linguistically ('The origin of the expression *bite the bullet* is...') and instances occurring as part of different idioms ('We *saw the*

[4] www.figure-eight.com. These experiments were funded by an "AI For Everyone Challenge" grant awarded to us by FigureEight.

[5] During one round of pilot annotations, crowdworkers were asked to indicate whether they were unfamiliar with the idiom type in question. This was almost never indicated, but annotation quality increased dramatically after adding in definitions nonetheless.

light at the end of the tunnel', when annotating the PIE *see the light*) which in this context is used as part of *light at the end of the tunnel*).

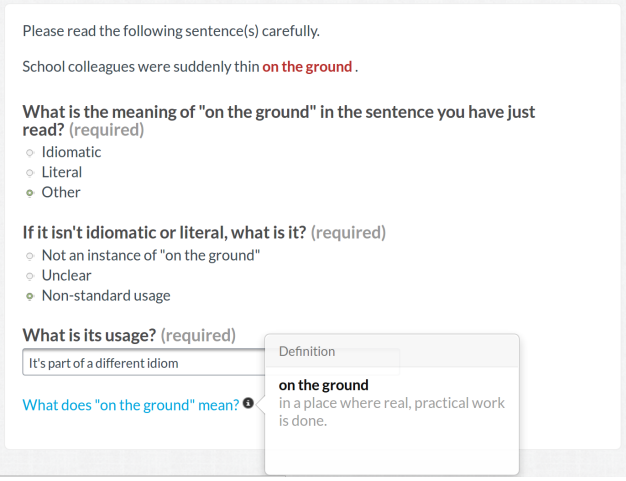

Figure 2: Screenshot of the annotation interface presented to crowdworkers, with definition tooltip displayed and all subquestions triggered.

4.1 Selection of Crowdworkers

The annotation procedure itself stayed constant throughout the annotation of the corpus, but the way crowdworkers were selected and tested was subject to experimentation and variation. Based on pilot rounds of annotation, our initial setup was as follows.

Annotators are presented with instructions which describe the nature of the task, which steps to go through at each annotation question, an overview of the possible answers, examples for each possible answer, and a section highlighting difficult cases (e.g., PIEs as part of proper names). After reading the instructions, annotators are presented with a set of six test questions, in the first phase of work (called quiz mode). If they answer at least 70% of these questions correctly, they can start the actual annotation phase (called work mode). Here, annotators are presented with six instances at a time, of which one is a test question. During work mode, their accuracy on test questions should remain above 70% for them to be able to continue working. Accuracy is used rather than other agreement metrics, since it is the only metric available on the platform. The test questions have gold standard annotations and were taken from the PIE Corpus described in Haagsma et al. (2019). Additional test questions were added later by taking items with high-agreement crowdsourced annotations and manually checking them for validity and suitability.

Although the use of test questions in general is a useful quality assurance method, we find that it is very much dependent on the selection of questions. The questions should be a good representation of the possible cases and answers in the data, balancing simple and difficult questions. Including too many hard questions excludes quality crowdworkers, while having only easy questions does not expose crowdworkers to the more challenging cases. We do this by enforcing a label distribution[6] among test questions and excluding test questions which are ambiguous. That is, we leave in difficult cases, but only if they are unambiguously so and clearly covered by the instructions. In addition, we provide reasons for why the test questions have a certain label. This ensures that, if crowdworkers get a question wrong, they get clear feedback on why that is the case and they can learn from this for the remainder of the task. If crowdworkers provide the wrong answer, they can also provide feedback on why they think their initial answer should be correct. Based on this feedback, the set of test questions is updated and problematic questions are filtered after every batch of annotations.

We limit the pool of crowdworkers to largely monolingual English-speaking countries, so we include only the United States, United Kingdom, Canada, Ireland, Australia, and New Zealand. Crowdworkers on the FigureEight platform are assigned level 1, 2, or 3 status, indicating the quality of their work, level 3 being the highest. We experimented with allowing different levels of crowdworkers to participate in the task and initially settled on using only level 2 and level 3 crowdworkers, since this greatly reduced the number of nonsensical contentions on test questions and the failure rate in quiz mode. Initially, no maximum number of annotations per annotator was set, but to prevent concentration loss and over-reliance on a single annotator, we settled on a limit of 500 per batch. Annotators were paid 4 cents (USD) per annotation (3, 3.5, and 5 cents per annotation were also tried).

Pilot annotations showed that, for many cases, having three annotations is sufficient and leads to 100% agreement. However, we also found that for a minority of challenging cases, agreement is very low, usually $< 50\%$. That is, we see a clear two-way split between straightforward cases (100% agreement) and ambiguous cases ($< 50\%$ agreement). As such, we collect more annotations for these low-agreement cases, until agreement is over 70%, up to a maximum nine annotations.

4.2 Managing Untrustworthy Crowdworkers

Initially, this setup yielded good results. We got a large number of annotations quickly with high accuracy. However, after a certain number of annotation batches, annotation quality degraded quickly and dramatically. We noticed a large influx of US-based crowdworkers achieving 100% accuracy on test questions, annotating the maximum number of instances unrealistically quickly, with implausible answer distributions (e.g., exactly 33.3% literal, 33.3% idiomatic, 33.3% other, whereas we expect ca. 70% idiomatic, 25% literal, 5% other) and very low agreement on those annotations. Manual inspection of the annotations showed that labels were assigned randomly, free text entries were nonsensical and the annotations were unusable. As such, we stopped annotation and concluded that the quality controls were being circumvented by (a group of) untrustworthy crowdworkers.

[6]The final label distribution is 25% false extractions, 32% idiomatic, 37% literal, and 6% other, but small variations on this have also been used.

Multiple attempts were made to prevent these workers from participating in the task. An additional constraint was placed on annotators by enforcing a specific answer distribution. We experimented with both strict limits, which erroneously excluded too many honest crowdworkers, and lax limits, e.g. *idiomatic* labels between 20% and 80%, *literal* labels between 10% and 60%, and *other* labels between 0% and 40%, which did not sufficiently filter out untrustworthy workers. We also tried not using the same test questions across multiple batches, adding noise to the test questions (machine-readable noise, invisible to humans), excluding USA-based workers, and excluding known untrustworthy workers from working on our tasks. However, none of these approaches yielded lasting results and annotation quality did not increase sufficiently.

Ultimately, the use of an open pool of crowdworkers was abandoned. The initial preference for using an unrestrained set of workers was based on maximising the number of different annotators for the data in order to ensure a diverse set of annotators (and thus, annotations) and to increase annotation throughput. However, this was outweighed by the problems described earlier, and we settled on using a manually selected set of known, reliable, trusted crowdworkers. These workers were selected by going through the list of workers with high accuracy scores on already-annotated batches, manually inspecting their annotations (both labels and text entries) and classifying them as either trusted, sincere but inaccurate, or untrustworthy crowdworkers. Only crowdworkers classified as 'trusted' were allowed to work on subsequent batches. As for untrustworthy crowdworkers, they were excluded from working on this task and their previous annotations were discarded as unreliable. This meant that part of the already-annotated data did not have sufficient annotations. To compensate for this, additional annotations on this data were done by the set of trusted crowdworkers, in order to not let the remaining good annotations on this data go to waste.

4.3 Crowd-Annotation Results

Using a set of 54 known and trusted crowdworkers, annotation quality increased drastically and remained stable throughout, as did the complexity of the free text explanations. Although annotation throughput was clearly lower with a small set instead of a large pool of annotators, several thousands of instances could still be annotated per day. Note that trusted crowdworkers still followed the same setup as previously, i.e., they still had to pass quiz mode and maintain a minimum level of accuracy throughout work mode. Pay was increased to 5 cents per annotation, the maximum of 500 annotations per crowdworker was maintained, and up to 7 annotations were gathered for difficult instances. In order to maintain a high standard within the set of trusted crowdworkers, their annotations were compared to the aggregated (majority agreement) label for the annotated instances after each batch. A high overlap indicates agreement with other workers, whereas low overlap reveals outliers. Generally, agreement between individual annotators and the majority label was at least 80%, indicating high reliability of annotations, but avoiding complete uniformity.

Listing 1: Example of a data point in the corpus for the PIE *larger than life*.


```
{"confidence": 0.8245609268758899,
  "context": [
    "I am undone -- `",
    "It was a deep voice of great
       beauty even when as now , she
       was over - emphasizing .",
    "Dear Vicky -- larger than life (
       too large for little life ... )
       ",
    "She had sat up and was pulling her
        fingers through the tangled
       forest .",
    "' Oh dear God -- I must take your
       name in vain ."],
  "document_id": "FPH",
  "genre": "W fict prose",
  "id": 316,
  "idiom": "larger than life",
  "judgment_count": 9,
  "label": "i",
  "label_distribution": {
    "?": 0.0,
    "f": 0.0,
    "i": 0.8245609268758899,
    "l": 0.17543907312411014,
    "o": 0.0},
  "non_standard_usage_explanations": []
    ,
  "offsets": [[13, 19], [20, 24], [25,
     29]],
  "sentence_no": "1078",
  "split": "training"}
```


Eventually, with the combination of funds and rows available, over 75% of the data was annotated. That is, of the 75,273 extracted instances, 56,622 were annotated. This constitutes a random subset of the full pre-extracted data, so we assume conclusions drawn based on the annotated data hold for the full dataset as well.

5 The MAGPIE Corpus

The annotated data is converted into a practically usable corpus by excluding those instances on which the majority of annotators agree that it is a false extraction from the data. Aggregation of annotations, i.e. selecting the majority label and assigning a confidence score, is done on the annotation platform, so those labels and scores are used. An example data point is presented in Listing 1.

Given that this is an annotated corpus of PIEs of unprecedented size, in addition to being an excellent resource for training disambiguation models, we we can gather substantial insights from it on idioms' characteristics and behaviour. In this analysis, we look into various aspects: annotation (dis)agreement, distribution of idiom types, sense distributions across types, composition of the 'other'-category, and influence of genre.

Please note that we carry out our analysis on the whole corpus. However, if one were interested only in the most reliable portions of the corpus, i.e. the cases all annotators agreed upon, different confidence thresholds can be set, as shown in Table 2. Even when including only instances with perfect annotation agreement, almost 4/5th of the corpus is included. This indicates that, although annotators found the task quite difficult, this difficulty originates from just a small portion of cases.

Threshold ($\geq$)	# Instances	% Instances
0.10	56,622	100.00
0.30	56,491	99.77
0.50	55,742	98.45
0.70	52,866	93.37
0.80	45,007	79.49
0.90	44,488	78.57
1.00	44,488	78.57

Table 2: An overview of corpus size using different confidence threshold values.

5.1 Distribution of Sense Labels

Given the full corpus of 56,622 annotated instances, our first point of interest is the overall distribution of sense labels. Table 3 provides an overview of label distributions for the corpus and its subcorpora. Noteworthy is the distribution, which is far from balanced. The 'other' and 'unclear' labels are rare, as expected, with only 436 'other' instances and seven 'unclear' instances in the whole corpus. The two major labels, 'idiomatic' and 'literal', are not equally common either, with idiomatic instances being around 2.5 times as frequent as literal instances. This is not wholly surprising, given that we include many idiom types which are unlikely to ever be used literally. Moreover, we excluded types which yielded an overwhelming amount of only literal instances and we downsampled the most frequent types, which were likely mostly literal as well. Nevertheless, there is a significant amount of literal instances, making the corpus suitable for PIE disambiguation experiments.

Section	Inst.	%idiom	%literal	%other	%uncl.
BNC-written	47,766	71.96	27.29	0.75	0.01
BNC-spoken	6,498	63.50	35.58	0.88	0.05
PMB	2,358	64.12	34.90	0.98	0.00
Total	56,622	70.66	28.55	0.77	0.01

Table 3: Basic statistics of the unfiltered annotated corpus and its subcorpora, including label distributions.

In addition to the overall picture, we see a clear difference between the written part of the BNC on the one hand and the other two subcorpora on the other. In the spoken and PMB data, literal instances are more frequent than in the written data, at the cost of idiomatic instances. Moreover, unclear instances occur more in the spoken data. This can be explained by the fact that transcribed spoken data is inherently more noisy than edited, written data.

5.2 Distribution in Idiom Types

Having 70% idiomatic instances and 30% literal instances by itself does not make a challenging PIE disambiguation task. If it were the case, for example, that 70% of idiom types are always idiomatic, and 30% are always literal, without any types having both usages, the disambiguation task would be trivial. As such, we need to look at the interaction between the label distributions and idiom types. How many types always have one sense? How many are relatively balanced?

The annotated data contains 1,756 different idiom types, with an average of 32.24 instances for each type. Table 4 shows the distribution of instances per type, i.e. how many types have only a single instance, and how many have over a 100 instances. Note that the number of instances per type was limited to a maximum of 200. The table shows that the vast majority of idiom types, 1,430 of 1,756, occur less than 50 times. This fits with the general idea that idioms are individually rare, but frequent as a group.

Frequency	Idioms	Example
1	126	*in apple-pie order*
2–5	372	*greasy spoon*
6–10	264	*hit the right note*
11–20	288	*big hitter*
21–50	380	*true to form*
51–100	148	*for Africa*
100–200	178	*at the crossroads*

Table 4: Overview of number of idiom types in each frequency band, with examples.

There is a large amount of variation in the label distributions of the 1,756 idioms in the data, ranging from all idiomatic to all literal. Given the overall predominance of idiomatic senses, it is no surprise that many idioms occur exclusively in their idiomatic sense. In fact, 1,035 out of 1,755 idiom types (58.94%) are purely idiomatic. Of the remaining 721 types, only 17 (0.97%) occur only in their literal sense, meaning that 704 types (40.09%) are ambiguous to some extent.

However, truly ambiguous idiom types, i.e. those with a label distribution close to 50/50, are rare: only 81 types (4.61%) fall in the 40–60% idiomatic range. Looking at high-frequency types (> 100 instances) only, it becomes even clearer that, even among ambiguous types, most are clearly leaning towards the literal or idiomatic side: 14 are in the 40–60% range, while 38 are in the 10–30% range. Looking at specific types, the most 'balanced' ones with a significant number of instances (> 10) are *cold feet* with 28 instances, 50% idiomatic, and *go all the way* with 168 instances, 50.60% idiomatic.

This is in line with previous findings, such as those by Cook et al. (2008). They select 53 idiom types, explicitly choosing those for which they deem both a literal and an idiomatic interpretation as 'possible'. Still, after this

pre-selection they define a 'skewed' subset of the data containing 25 expressions which were too imbalanced for their purposes. Even then, the remaining 28 expressions are generally strongly imbalanced, with only four falling in the 40–60% range.

5.3 The Black Sheep

For 'other'-labels, we also collected annotators' explanations for selecting that label. A closer inspection of these reveals what makes a PIE fall out of the binary split between 'idiomatic' and 'literal', and which subcategories make up the 'other' instances. There are 436 instances with 'other' as the majority label. Although these instances have a low average confidence score of 0.5748, manual inspection reveals that the labels are generally correct and reliable. By applying some simple heuristics, we can get an overview of the frequency of different subgroups in this category. For example, we feel it is safe to assume that any instance with free text explanations containing the word 'linguistic' are used meta-linguistically. Based on such heuristics, we find that, out of a total of 436, at least 41 (9.40%) are (part of) titles and names, which we consider non-PIEs and should have been annotated as such (Example (1)), at least 332 (76.15%) are PIEs part of bigger and/or different idiom types (Example (2)), and at least 21 (4.82%) are meta-linguistic usages (Example (3)).

(1) George was recording '**Under Lock and Key**', Steve was doing 'Eat 'Em And Smile' in David Lee Roth's band [..] and the list went on from that. (*under lock and key* - BNC - doc. C9J - sent. 1504)

(2) GRAHAM Taylor was down **to the bare bones** today when only 14 of his England squad took part in his first Bisham training session [..] (*to the bone* - BNC - doc. K2D - sent. 288)

(3) When gentle-folk rented horse-drawn transport from Mr Hobson, they were never allowed to select their own horse and carriage, [..] hence the expression '**Hobson's choice**', meaning no choice at all (*Hobson's choice* - BNC - doc. B11 - sent. 697)

5.4 Influence of Genre

A benefit of using the BNC is that it has genre information. For example, a document having the genre label `W fict prose`, indicates that it is (part of) a written work of fictional prose. An additional advantage of these labels is that they are not atomic, but structured: every part of the label forms a category of labels. For example, we could select all documents with genres starting with `W` to get all written documents, all `W fict` documents to get all kinds of written fiction, and so on. For an overview of genre encoding in the BNC, see Lee (2000).

We are interested in two main things related to genre: the distribution of PIE labels across genres and the overall frequency of PIEs and idioms in each genre. At the highest level of genre labels, distinguishing written from spoken language, there are only small differences. PIEs and idioms are almost equally frequent, while written PIEs are somewhat more likely to be idiomatic (72%) than those in spoken language (63%).

At a more fine-grained level, there are 34 different genres. Here, we see much larger differences appear. For example, in transcribed parliamentary language PIEs are idiomatic 87% of the time, whereas only 38% of PIEs in the also transcribed 'live' demonstration genre are idiomatic. Generally, we find technical and instructional language to have the lowest percentage of idiomatic PIEs, whereas the highest percentages are found with speech and writing which is persuasive in nature, such as political discussions and debates. The likely underlying reason is that the genres with more literal PIEs talk about concrete, physical things (as in a demonstration or instruction), whereas the genres with more idiomatic PIEs focus on rhetoric and abstract ideas, leading to the invocation of PIEs in their non-literal sense. The high frequency of idioms in persuasive and rhetorical language corroborates the statements by McCarthy (1998) that idioms are used for commenting on the world, rather than describing it, and Minugh (2008), who finds that idioms are used most often by those with some authority, especially when conveying 'received wisdom'.

However, this genre distinction is still quite crude. For example, all academic texts are grouped together under `W ac`. When this genre is split out further, a clear split emerges between the texts on exact sciences on the one hand, which rarely use idiomatic PIEs, and humanities, law and social sciences on the other hand, which use idioms much more frequently. Perhaps, this difference too is caused by the focus on concrete, physical things in one area and a focus on rhetoric and abstract ideas in another. Our findings nicely complement those by Simpson and Mendis (2003), who also look at the frequency of idioms in different academic disciplines. Their initial expectation is that 'soft sciences' use more idioms than 'hard sciences', but they find no significant difference in their data. However, they use transcribed spoken data of different academic contexts, whereas we use written papers. As such, a possible explanation for their incorrect expectation is that it was based on written data from the different disciplines.

The balance between idiomatic and other usages of PIEs does not mean much if the overall number of PIEs and/or tokens in the genre is very small. Therefore we also look at the frequency of PIEs relative to the number of tokens. The average frequency across the corpus is 560 PIEs per million tokens, of which 70.95% are idiomatic, so 397 idiomatic PIEs per million tokens. These frequencies differ greatly between genres, with `W ac` averaging only 230 PIEs per million, and `W newsp` averaging 830 PIEs per million. More extreme values are found, but only with small genres such as `W admin` and `S sermon`.

If we exclude such genres, looking only at those having at least a million tokens, we get the overview in Table 5. Now, `W news script`, which contains newsreaders' autocue data, contains the most PIEs per million tokens, 896, almost nine times as frequent as its opposite, `W ac: medicine`, academic texts on medicine, which has only 103 PIEs per million tokens. For idiomatically used PIEs, the proportions are similar.

Finding `W news script` and `W newsp other` with

Genre	%Id.	#PIEs	K-Tok.	f-PIE	f-Id.
W news script	75.86	1,077	1,202	896	680
W newsp other	80.94	4,712	5,559	848	686
W fict prose	65.40	13,250	15,662	846	553
W pop lore	73.59	5,797	7,356	788	580
S conv	56.30	2,849	3,979	716	403
W newsp brdsht	79.54	2,116	3,010	703	559
W biography	72.52	2,107	3,523	598	434
S meeting	76.60	748	1,342	557	427
W religion	78.31	627	1,125	557	436
W nonAc: tech engin	87.07	642	1,212	530	461
W misc	66.03	4,416	9,190	480	317
W hansard	87.24	478	1,167	409	357
W nonAc: nat science	63.89	1,022	2,527	404	258
W nonAc: soc science	73.46	1,458	3,683	396	291
W commerce	81.75	1,452	3,789	383	313
W nonAc: hum. arts	71.65	1,330	3,724	357	256
W ac: hum. arts	78.78	1,046	3,338	313	247
W nonAc: pol law edu	78.18	1,219	4,502	271	212
W ac: soc science	75.78	1,160	4,765	243	184
W ac: polit law edu	79.00	1,081	4,671	231	183
W ac: nat science	36.25	160	1,122	142	52
W ac: medicine	59.18	147	1,433	103	61

Table 5: Distribution of (idiomatic) PIEs across genres. *K-Tok.* is the number of tokens in each genre, in thousands. *f-PIE* is the frequency of PIEs per million tokens. *f-Id.* is the frequency of idiomatically used PIEs per million tokens.

the highest idiom frequencies aligns nicely with previous work on idiom frequencies. For example, Moon (1998) notes that the frequency of idioms in spoken language has been overestimated relative to written language. She suggest this may be caused by the high frequency of idioms in scripted speech, such as in fiction, film, and television, a category which also covers `W news script`. As for `W newsp other` (and `W newsp brdsht`, which has the third-highest fIdiom), it has been noted that journalistic writing is a particularly rich source of idioms (Moon, 1998; Fazly et al., 2009; Grégoire, 2009).

More generally, we see that academic texts (`W ac`) have the lowest frequency of PIEs, followed by non-academic non-fiction (`W nonAc`), i.e. texts whose main purpose is instruction, information, and education. PIEs are most frequent in news, prose fiction, conversations, and popular magazines (`pop lore`), i.e. texts whose main purpose is entertainment. However, spoken conversations (`S conv`) do not fit this category neatly, even if they have a similar PIE frequency. We have no clear explanation for its high PIE frequency, but we do note that it stands out of the group of news, prose, and magazines if we look at the frequency of idiomatically used PIEs rather than PIEs overall.

6 Conclusions

We have built the largest corpus of sense-annotated PIEs to date using a crowdsourced annotation approach, and we make it available to the community to enable progress in idiom disambiguation and evaluation. Based on a high-accuracy list of idioms created by combining idiom dictionaries, and by using a strictly controlled crowdsourced annotation procedure, we created a high-quality corpus containing a total of 56,622 PIE instances.

Based on the lessons learned during the creation of the corpus and the findings resulting from analysing its contents, we are now in a good position to answer the research questions posed in Section 1. First, we consider whether crowdsourcing is a suitable method for large-scale, high-quality annotation of a large variety of potentially idiomatic expressions. The answer to this is a qualified yes: crowdsourcing is suitable, but the procedure has to be set up carefully to yield reliable results. We found that the most important characteristics were the procedure for selecting crowdworkers, dealing with untrustworthy annotators (cf. Buchholz and Latorre (2011)), the writing of instructions which are both clear and comprehensive, and the development of an interface that breaks down the task in small, manageable steps. The fact that, taking care of these requirements, yielded a corpus which is an order of magnitude bigger than previous ones, both in number of types and instances, with a high-level of inter-annotator agreement, confirms the potential of crowdsourcing for complex linguistic tasks.

As for the second question, which covers the relation between a corpus of this size and its potential for insight in idioms' behaviour and distribution, the answer is varied. On the one hand, the fact that the BNC is the base corpus means it has rich genre information, which provides insight into the matters of idiom usage in spoken vs. written language, and differences between academic disciplines. On the other hand, what we have done here in terms of analysis is limited to a relatively high-level view of idiom distributions and frequencies. As such, the true potential for linguistic insight of the corpus remains for further investigation, which likely includes more in-depth inspection of specific questions and manual encoding of specific idiom characteristics, such as a more fine-grained classification of variation types than we have used here.

Finally, we look at previously under-researched matters regarding idiom: the influence of genre on the proportion of literal vs. idiomatic PIEs. Overall, we have found that almost all types are strongly skewed towards either idiomatic or literal usage, and that truly balanced types are rare. This implies that a corpus properly representative of idiom as a linguistic phenomenon will necessarily include many skewed types, which contrasts strongly with previous corpus-building approaches, which all explicitly aimed to include mostly balanced idioms. The strongest effect, however, is that of genre on the proportion of literal and idiomatic PIEs. The percentage of idiomatic PIEs ranges from over 85% in some genres to less than 40% in others, showing a clear pattern: technical, instructional language discussing concrete, physical topics have more literal PIEs, while argumentative, rhetorically rich language touching on more abstract concepts contains more idiomatic PIEs.

7 Acknowledgements

This work was funded by the NWO-VICI grant *Lost in Translation – Found in Meaning* (288-89-003). The crowdsourcing effort was made possible by FigureEight, as part of their AI For Everyone Challenge competition.

8 Bibliographical References

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain, April. Association for Computational Linguistics.

John Ayto, editor. (2009). *From the horse's mouth: Oxford dictionary of English Idioms*. Oxford University Press, Oxford; New York, 3rd edition.

Buchholz, S. and Latorre, J. (2011). Crowdsourcing preference tests, and how to detect cheating. In *INTERSPEECH-2011*, pages 3053–3056.

Cook, P., Fazly, A., and Stevenson, S. (2008). The VNC-Tokens dataset. In *Proceedings of the LREC Workshop: Towards a shared task for Multiword Expressions*, pages 19–22.

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

Grégoire, N. (2009). *Untangling Multiword Expressions: A study on the representation and variation of Dutch multiword expressions*. Ph.D. thesis, Universiteit Utrecht.

Haagsma, H., Nissim, M., and Bos, J. (2019). Casting a wide net: Robust extraction of potentially idiomatic expressions. arXiv:1911.08829.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

Kato, A., Shindo, H., and Matsumoto, Y. (2018). Construction of large-scale English verbal multiword expression annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Lee, D. (2000). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. In *Proceedings of the Fourth International Conference on Teaching and Language Corpora*, Language and Computers, pages 245–292. Brill — Rodopi.

McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge University Press, Cambridge.

Minugh, D. (2008). The college idiom: Idioms in the COLL corpus. *ICAME Journal: Computers in English Linguistics*, 32:115–138.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press, Oxford.

Simpson, R. C. and Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly*, 37(3):419–441.

Sporleder, C. and Li, L. (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, March. Association for Computational Linguistics.

Sporleder, C., Li, L., Gorinski, P., and Koch, X. (2010). Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

9 Language Resource References

BNC Consortium. (2007). *The British National Corpus*. Distributed by Bodleian Libraries, University of Oxford, Version 3 (BNC XML Edition).