The Analysis of PP Idioms

Ken Litkowski
CL Research
9208 Gue Road
Damascus, MD 20872 USA
ken@clres.com

July 23, 2018

Abstract

The Preposition Project (TPP) used a sense inventory of single-word and phrasal prepositions in the *Oxford Dictionary of English* (ODE). The inventory was created from entries labeled as prepositions and from phrases under other entries having preposition signature definitions. TPP provided the behaviors for each sense using 20 fields, particularly characterizing the preposition complement. Some prepositions were given **treatments** summarizing the senses, some times mentioning some common idiomatic complements (without further detailed analysis). Recently, Schneider et al. (2016) describing an annotation effort to give supersenses and identified a difficulty in characterizing some idiomatic phrases (e.g., *by far* and *for free*). In commenting on such idioms, we suggested that looking up the definitions for such phrases in ODE might help to identify the appropriate supersense.

Such idiomatic phrases raise the question of whether they have been properly incorporated in the TPP sense inventories. This question requires two tasks: (1) identifying a comprehensive list of such idioms and (2) determining how to incorporate the idioms into the current sense inventories. From the 120,000 dictionary entries in ODE, we have identified 5269 multiword expressions (MWEs) that are entries that might be preposition-phrase (PP) idioms. These MWEs are entries that either begin (2484) or end (2785) with prepositions in the TPP set of entries. We describe how these were identified, their presence and absence in TPP corpora, how to integrate these idioms into TPP (possibly expanding corpora with these idioms), the difficulties in annotating the PP idioms, and describing future work.[1]

[1]Working Paper 18-01. Damascus, MD: CL Research. Contains tasks under study

1 Introduction

The sense inventory for The Preposition Project (Litkowski (2002); TPP, Litkowski (2005)) was initially obtained from the *New Oxford Dictionary of English* (Pearsall, 1998). This consisted entries labeled as prepositions and from phrases based on **definitions** that have (1) a single-word or phrasal preposition definition, (2) a prepositional phrase + a preposition, (3) (an optional leading string) + a transitive present participle, or (4) a leading string + an infinitive of a transitive verb. The definitions of these TPP entries constituted the prepositional sense inventory. In tagging sentences for a SemEval task of preposition disambiguation, Litkowski and Hargraves (2007) expand the inventory by 10 percent. In the development of the Pattern Dictionary of English Prepositions (PDEP, (Litkowski, 2014)), the sense inventory increased by another 10 percent. Many of these new senses arose from corpus data showing frequent preposition complements.

Schneider et al. (2015, 2016) proposed in PrepWiki[2] (henceforth "v1") a supersense inventory of 75 categories proposed for English for adpositions for use in a corpus. Hwang et al. (2017) have developed a new "v2" inventory of 50 categories and have been using it to annotate the same corpus. During this annotation, discussions arose how to annotate some phrases that appeared to be idiomatic (e.g., *by far* and *for free*), We suggested that using the definitions of such phrases might help the annotation.

When we looked at such prepositional phrase (PP) idioms and their definitions, we wondered whether TPP and PDEP include them. Such PP idioms are multiword expressions (MWEs) whose first word is a preposition. While it would seem that such phrases would appear in the entries for the first word (preposition), they do not. Instead, they mostly appear in the dictionary as **phrases** under the entries for the preposition object (e.g., *by far* is defined as "by a great amount" in the entry for *far* and *for free* is defined as "without cost or payment" in the entry for *free*). These definitions suggest, respectively, that *by far* has the supersense Extent and that *for free* has the supersense Cost.

A systematic examination of whether PP idioms are properly handled in TPP and PDEP requires several additional questions:

- What are PP idioms? (See Section 2.)
- How can PP idioms be identified from dictionaries? (See Section 3.)

(see to-do list in section C). This paper and its supporting data are available at `https://github.com/kenclr/ppidioms`.

[2] `http://tiny.cc/prepwiki`

- What PP idioms are present in the PDEP corpora and have been tagged? (See Section 4.)
  - Have some PP idioms been incorrectly annotated in the PDEP corpora?
- What PP idioms are missing from PDEP sense inventories? (See Section 5.)
- How can the PP idioms be integrated into the senses for the appropriate preposition entry? (See Section 6.)
  - What are the procedures to analyze the definitions of the PP idioms?
- What corpora data can be used to support the PP idiom justification? (Can some be obtained from OUP sentence senses?) (See Section 7)
- How do tokens tagged as PP idioms in the STREUSLE corpus correspond to the dictionary PP idioms? (See Section 8.)

2 Characterizing PP Idioms

PP idioms[3] are multiword expressions (MWEs), consisting of two or more words and usually beginning with a preposition. Intuitively, it might seem that the location of the meaning for a PP idiom might be found in the entry under the preposition. However, Lew (2012) indicates that prepositions are unlikely to be found there, but rather from some other place. In most cases, the location is likely to be found in the entry for the object/complement of the preposition.

Cieślicka (2015) provides several models by which native speakers store idioms and suggests several varieties of compositional constituents in PP idioms. Her analysis of these models predicts that PP idioms have several types, but does not provide criteria that will predict new idioms. Baldwin et al. (2006) provided an analysis of "determinerless PPs" (PP-Ds), characterizing them into four types of classes.

- Lexical listing: P + N combinations in the lexicon

[3] "A group of words established by usage as having a meaning not deducible from those of the individual words", `https://en.oxforddictionaries.com/definition/idiom`

- Prepositions with defective NPs: Noun phrases occur as subjects and objects also without determiners (hence making them acceptable in PP-Ds)
- Prepositions with idiosyncratic NPs: Nouns can take only a restricted set of modifiers
- Prepositions that select $\bar{N}$: the complement is constrained to be an unsaturated $\bar{N}$

Many of these instances are incorporated in a dictionary into senses for a preposition (e.g., *by* with a sense defined as "(followed by a noun without a determiner) in various phrases indicating how something happens" such as *by chance*) (Pearsall (NODE; 1998) and Stevenson and Soanes (ODE; 2003)). In particular, the definitions are provided words having clear meanings, and thus, not idiomatic. As a result, just being a "determinerless PP" does not suffice to identify something as a PP idiom, as will be indicated below.

These points will be discussed in more detail below when the full sets of idioms are described.

3 Identifying the Inventory of PP Idioms

As implied above, there do not appear to be criteria for identifying what makes an MWE idiomatic. In the absence of such criteria, an initial list of PP idioms may be identifiable by using dictionaries, i.e., where lexicographers have provided MWEs as treated them as idiomatic. In NODE and ODE, as with other dictionaries, phrases and phrasal verbs are append to main entries. Such phrases may be initially viewed as possibly idiomatic. However, as discussed in Lew (2012), it is difficult to find these phrases, since there are not likely to have a simple alphabetical arrangement, as suggested above.

More than just finding prepositions, Litkowski (2001, 2002) described the procedures for using NODE data in creating an alphabetic dictionary that included all phrases. Subsequent editions of ODE were also processed in similar procedures. As mentioned, these MWEs were the basis for the construction of The Preposition Project (TPP;Litkowski (2005)), but they were used only to identify phrasal prepositions. They did not cover other PP idioms described above (*by far* and *for free*).

Working with the latest ODE (in data from 2008), we worked with entries containing all phrases with all senses. We created one file for all senses,

where each line consists of the entry name, a sense number in parentheses, a dash, a three-letter identifier for the part of speech, and the definition

- *across country (1) - (phr) not keeping to roads.*

This dictionary contains 121,006 entries and 201,389 senses. We then applied a regular expression (regex) to each line, between the first character of the line and the first left parenthesis of the line (which begins a sense number). The purpose was to identify entries that putatively correspond to PP idioms, using the two regexes:

- `^preposition [a-z ]+ \(` "recognizing a preposition beginning"
- `^[a-z ]+ preposition \(` "recognizing a preposition ending".

Each preposition in PDEP was substituted for each regex. (This could have been performed automatically, but was done manually in about one hour for each.)

An editor software (EditPad Lite[4]) was used to obtain the matches. Each of the 304 prepositions in the regex was executed to find and highlight all the lines of the total senses that matched the regex. (The software would also indicate when there were no matches for a preposition.) The software could then copy all the matches for the preposition to a file. The result was 2484 lines for the preposition beginning regex and 2785 lines for the preposition ending regex. The two files constitute the initial inventories for further analysis. Many lines are not PP idioms, but rather other kinds of phrases; these will not be considered further.

3.1 Preposition Initial PP Idioms

The 2484 instances with an initial regex preposition consisted of 63 distinct prepositions, 50 using single-word prepositions and 13 with phrasal prepositions (e.g., *out of* and *short of*). The five most frequent initial prepositions are: *in* (673), *on* (391), *at* (218), *to* (115), and *by* (97). Most of the items (76 percent) were identified as valid PP idioms (see Table 1), with 16 percent could be dismissed from further examination. Five percent of the items (126 instances) were phrases already included in TPP and PDEP (i.e., not requiring any further examination). Two categories suggested further problematic. An additional 54 instances identified phrasal prepositions (e.g., *as opposed to*, *in justice to*, and *with a view to*); these should probably be added to PDEP.

[4] `https://www.editpadlite.com/`

Table 1: Characterization of Preposition Initial Instances

Characterization	**Count**
Good PP Idioms	1891
Phrases Already in PDEP	126
Potential into PDEP	54
Double Preposition	7
Invalid Idiom	406
Total	2484

Since these phrasal instances end with a preposition, they will also arise in preposition ending instances and discussed in Section 3.2.

There are 7 instances consisting as double prepositions.[5]

Need to determine in which preposition should these be placed. These will also appear in preposition ending instances in Section 3.2. They appear to belong to the final preposition.

(Appendix A describes how this file was processed.)

For the most case, these instances corresponding the PP idioms that raised the question about supersense annotations (i.e., *by far* and *for free*). Most of the instances are entries that constitute ordinary PPs with an initial preposition following by a noun phrase (frequently determinerless). For these cases, their definitions can be used for determining where the idiom belongs in the existing PDEP inventory (if possible), as well as for identifying the supersense analysis and annotation. However, the situation is often more complicated, with several types of characteristics that required further discussion in the following list.

1. **Phrasal Prepositions**: Multiword prepositions that begin with a single-word preposition will include the phrasal prepositions in PDEP (e.g., after in *after the fashion of*). In constructing TPP and PDEP, the objective was to obtain all such phrasal prepositions. As indicated in Table 1, the regex process identified 126 PPs already included in PDEP, but also found 54 additional phrasal prepositions that should be added to PDEP.

   Phrasal prepositions are also included in regex processing and thus are allowed to generate PP idioms which begin with phrases. This has

[5] *across from*, *from above*, *in on*, *in with*, and *on to* (3 instances)

identified PP idioms for *a la*, *ahead of*, *all over*, *apropos of*, ***in spite of***, *more like*, *near to*, ***on top of***, *out of*, *short of*, *this side of*, ***up for***, and ***up to*** (the bold prepositions were triggered by the first word as well as the phrase). That is, these phrases also have further PP idioms with complements.

2. **Multiple Senses**: Many PP idioms have more than one sense. There are 1561 unique idioms in the list of 1891 idioms, i.e., 330 instances correspond to ambiguities. As with any single word entries (e.g., nouns or verbs), word sense disambiguation (WSD) issues will affect supersense annotation. Thus, an instance of a PP idiom would have to examine the context to determine which sense is intended. In addition, the multiple senses may be tied to different senses of the preposition's sense inventory.

3. **Adjective and Adverb Senses**: Some PP idioms may be used in both an adjective and an adverb sense. In the file, there is a sense in each part of speech. Identifying the appropriate sense is somewhat easier than WSD. For example, *a la mode* has 6 instances, three identified as adverbs and three as adjectives; the definitions are the same for the adverb and the adjective; as a result, these senses are not duplicated in the counts. The duplicate senses were removed.

4. **PPs with Preposition Sense**: Many PPs have a sense with a PP definition, i.e., the first word in the definition is a preposition and the remainder of the definition is a noun phrase. In many of these definitions, the initial preposition is the same as the preposition in the PP idiom (e.g., at home is defined as *at a team's own ground*). When this is the case, it is possible that the definition may identify an existing sense in the inventory.

   When the definition of the MWE begins with the prep, how should it be described in PDEP? In likely, this can be specified using the **lexset** field.

   It is likely that the majority of the definitions use a different preposition, so that it will be necessary to examine the sense inventory for a different preposition.

5. **Multiple PP Idioms**: Some PP idioms have a definition that indicate there are actually two idioms. For example, against the stream is defined as *against (or with) the prevailing view or tendency*, indicates

that with the stream is also a PP idiom (which is not identified in our analysis). In such cases, the parenthetical preposition is the opposite of the PP idiom.

When the definition of the MWE includes a parenthesis indicating that there is an opposite MWE, i.e., antonymous, it begins with "or". There are 58 instances, although some of them are not. Some of the antonyms may be useful to see whether the opposites can be identified.

Some PP idioms are not fixed, but indicate some variability. Approximately 200 include a possessive variable (e.g., *in one's bones*, where usually requires a word such *my* or *their*); in these cases, the example sentences will show the variety. About 15 cases identify a variable word, e.g., *within ?? distance*, where the questions indicate that any word such as *walking* or *hailing* can fill the space.

6. **PPs with Adverb Sense**: Many PP idioms are defined with an adverb (e.g., at bottom is defined as *fundamentally*). Many of these may have a meaning *in a manner . . .*, suggesting that is a Manner. In some cases, the idiom may have an overt manner sense, e.g., at leisure is defined as *in an unhurried manner*.

   For example, there are many definitions (150) ending with "-ly" (having the qualities of, recurring at intervals of, denoting manner or degree).

7. **Prepositional Sense Circularities**: Many PP idioms have definitions that are themselves PP idioms (e.g., *at stake* is defined as *at risk*, which is also a PP idiom). It is possible that some of these definitions are circular. When there are such PP idioms as definitions, a search should be found for a non-idiomatic definition as the basis for the meaning.

   There is no simple method for identifying cases with circularity definitions. It's easy to determine, but this would require examining the 1891 definitions as shown in Table 1.

8. **PPs with Multiple Senses**: Some PP idioms have more than one definition, usually two, raising the question whether there are distinct senses that should not be included under the same sense. For example, at need is defined as *when needed; in an emergency*; while it is likely

that the two senses are very close in meaning, such cases need to be examined properly.

There are 270 instances containing a semicolon, ";", indicating the presence of multiple definitions.

9. **PPs with Various Parts of Speech**: Many PP idioms have definitions that would be parsed with verb forms, adjectives, or (as mentioned) adverbs. These definitions are likely to provide a clue to which sense of the base preposition should be linked. To characterize such senses, it would be useful to examine sentence examples, where the definitions would best indicate how the idiom can be interpreted in context.

10. **Missing Definitions**: Many PP idioms (228) did not included definitions (identified as "**(phr) ()**") in the full set of 2484 instances. Some of these have definitions as phrases under other words (typical nouns) that need to be identified. Some of these require a double step to find the definition (e.g., *at one fell swoop* is not under the entry for *swoop*, but rather has to go to the adjective sense of *fell* to find the idiom). Some of these are possibly just a placeholder that refers to the phrase group under the main entry (e.g., *at a loss* has three senses, but the first sense only indicating that the main entry as the source for the phrases is *loss*.

    Cases in our data that do not have a definition will be examined from the online dictionary to add the definition in our file.

11. **Other Phrase Types**: Several phrases that begin with a word that has a preposition form (e.g., ***bar*** *mitzvah* (a noun phrase) or ***as*** *the crow flies* (a clause)) are not prepositional phrases. There are 395 phrases of the 2484 preposition initial instances. There would be no need for further investigation of these instances.

These items above describe the patterns in the preposition initial PP idioms. The to-do items above identify the tasks that will be followed for each item. As items are dealt with, the to-do items will be deleted and described the steps that have been performed.

Table 2: Characterization of Preposition Final Instances

Characterization	**Count**
Phrases Already in PDEP	182
Potential into PDEP	141
Verb Phrases	2188
Miscellaneous Phrases	274
Total	2785

3.2 Preposition Final PP Idioms

The 2785 instances with an ending regex preposition consisted of 66 distinct prepositions, 51 using single-word prepositions and 15 with phrasal prepositions (e.g., *out of* and *short of*[6]) The five most frequent ending prepositions are: *up* (417), *of* (353), *on* (317), *off* (246), and *down* (172). Most of these instances are not PP idioms, but rather verb particle/preposition constructions (VPCs). (Appendix B describes how this file was processed.)

To begin the process of examining these instances, the first step was to create a sheet in Excel containing the full set. A new column was added to identify a line number for each instance. Then, in an text editor, the regex "`[a-z]+ \(\d+`" was used to identify the last word in the dictionary entries followed by a space, an opening parenthesis, and one or more digits (corresponding to a sense number); the parenthetical sense number was then stripped off to provide the list of prepositions. Another column was added after the instance number and the list of prepositions was added, thus identifying the final preposition that triggered the instance. However, when the instance had been triggered by a phrasal preposition, such as *along with*, *with* was the last word and thus appeared out of alphabetical order, i.e., *with* appeared after the instances for *along* and before the instances for *around*. This non-alphabetic order enabled the identification of the 15 phrasal prepositions. After these prepositions were identified, we were able to create a table in Excel showing the number of instances for each preposition.

Next, we added a column to code the type of each instance, as follows:

1. **Existing PDEP Phrases**: Phrases already included in PDEP, e.g., *with reference to* (182 instances)

[6]Curiously, these two phrasal prepositions occur in the preposition initial PP idioms as well.

2. **Missing PDEP Phrases**: Phrases appearing to be phrasal prepositions that perhaps should be added to PDEP, e.g., *in the wake of* (141 instances)

3. **Double Prepositions**: Two-word phrases beginning and ending a preposition (the same 7 instances as in Table 1, and already identified in type 2 above)

4. **Verb Phrases**: Phrases beginning with a verb and judged as a verb particle or preposition (e.g., *set by*)

5. **Noun Phrases**: A noun phrase, containing a noun definition (e.g., *snack bar*)

6. **Usage Phrases**: Phrases with definitions beginning with such words as *used* or *said* (e.g., *how about* defined as "used to make a suggestion or offer")

7. **Adverb Phrases**: Phrases with an adverb definition (e.g., *by the by* defined as "incidentally")

8. **Adjective Phrases**: Phrases with an adjective definition (e.g., *all in* defined as "exhausted")

9. **Assorted Phrases**: Subordinating conjunctions (*as long as*), complete sentences (*when the chips are down*), and various idiomatic phrase (*and so on*).

Only a small proportion (11.6 percent) of the items were identified as valid PP idioms (items 1 and 2 in Table 2).

Most of the items (78.6 percent) were VPC phrases (item 4 above). The first word of these phrases is a verb and the final word is a preposition (e.g., *get at*) or a particle (i.e., adverb, e.g., *fall behind*). In the Oxford dictionary, these items are "phrasal verbs" that appear under the entry for the verb. When the final word is a preposition, the definition indicates that the phrase would follow by a direct object (e.g., *get at* is defined as "bribe or unfairly influence (someone)"). When the final word is a particle, the definition is a complete phrase (e.g., *fall behind* is defined as "fail to keep up with one's competitors"). While these instances are not needed for examining PP idioms, these VPCs could be very rich for study of verb MWEs (VMWEs).

The remaining preposition final instances (10 percent) include almost all other part-of-speech types (in items 5 through 9 above), although not

having much significance for the study of PP idioms. In Table 2, these are grouped together as miscellaneous phrases invalid idioms.

4 PP Idioms in Existing Corpora

When a PP idiom appears in the PDEP corpus, the tagging has already been identified as corresponding to one of the senses for the preposition. In these cases, the idiomaticity appears to have been somewhat weaker, making it possible to infer (or slightly deducible) a sense. For example, for *ahead of*, all the five PP idioms were present in the CPA corpus, tagged in two of the four senses. Two senses (*schedule* and *time* appeared in 17 of 92 instances; the other three (*the game* and two senses of *one's time*) appeared in 24 of 103 instances. Thus, it would seem that the lexicographers constructing the sense inventory this preposition had already incorporated the PP idioms.

To determine the extent to which the idioms occur in the TPP corpora, we use Sketch Engine (SE) implementation of the preposition corpora.[7]. While it is easy to examine each of the 1891 PP idioms, this quite time-expensive.

Need to develop an automatic procedure for determine whether each idioms is present in the current corpora. It would have to take into account some of the non-frozen idioms, such as *under one's arm* or *out of one's mind*. It might be easier to use the "vertical" file used to create the SE implementation.

Some considerations:

- When a PP idiom is present in any of the corpora, particularly the CPA corpus, have all the occurrences been tagged consistently?
- Should I look at the FN and OEC corpora as well as the CPA corpus? (Probably yes)
- What file should be used to find the results? (SE has a mechanism for identifying MWEs; this might be useful.)
- Should I look only at those in which the target is the preposition and the complement (or in other instances as well)? (Probably not)

[7] `https://www.sketchengine.co.uk/english-preposition-corpus-1/`

As an example, note ***within ?? distance***, which has it's own OEC senses. There are also two preposition initial instances, *within spitting distance* and *within striking distance*. All three have their own set of example sentences, in addition to the OEC sense for *within* (*5(2)*) in PDEP.

5 Missing PP Idioms

There are two types of missing PP idioms from PDEP: (1) preposition entries that had not been previously recognized (primarily phrasal prepositions) and (2) senses not present or characterized in the sense inventory for an existing preposition entry. Each type requires different steps.

5.1 Identifying New PDEP Entries

The items labeled "Potential into PDEP" in Tables 1 and 2 provide the initial potential entries. All of these potential (phrasal) entries, i.e., MWEs, end with a simple (one-word) preposition. Almost all entries currently in PDEP follow this pattern, with the exception of some entries that are gerunds (ending in *-ing*) and that also have been characterized as prepositions in the dictionary.[8] For the most part, the instances in Table 1 are also in Table 2. As a result, the entries will be focused in order to identify which entries get added.[9] These will be reviewed in detail to determine if these instances should be added to PDEP.

The first step for examining a potential entry is whether it already appears in the PDEP corpus. For example, for *at the mercy of*, we need to determine if the phrase occurs in either the *at* or *of* corpus sentences. (No such case has yet been seen.) In addition, we need to examine whether an instance has already been incorporated in the sense inventory for a preposition in the phrase. For example, *such as* was previously added as a sense under **as** because of its frequency in the PDEP TPP corpus.

Addition of entries to PDEP that meet this first stage will then follow the procedures used for current entries, as follows:

- Adding an entry to the set of preposition senses, with the preposition name, a sense number, and a definition (creating the basic entry)

[8] The only phrasal entry not following this is *as regards*.

[9] The following have further need of characterization: *in course of ??*, *in the event of ??*, *in the interests of something*, *up to one's ears in*, *up to one's elbows in*, and *up to one's neck in*.

- Start adding data into the fields of the pattern
- Obtaining the OEC sentences for the entry and sense (for tokenizing, parsing, and feature generation) (See Section 7 for more details.)
- Obtaining a sample of sentences from the BNC (for tokenizing, parsing, and feature generation) (from a sample using Sketch Engine)
- Tagging the BNC sentences, particularly when a new entry has multiple senses.

5.2 Identifying New Entry Senses

As indicated above, the first step of assessing each idiom is to determine whether it has already been tagged in the existing PDEP corpora. This will be done using the Sketch Engine vertical file (Litkowski, 2017). Each corpus uses a <**doc**> XML element, containing **corpus** and **preposition** attributes. Each sentence uses an <**s**> element, containing attributes for PDEP data, including the tagged sense label, the instance number for the corpus, the sense class, and the sense subclass.

When we encountering sentences for one of the 63 distinct prepositions (see Section 3.1), we will gather together the complements that will be examined for presence in each sentence. This will involve identifying multiple senses and non-fixed idioms. For example, *ahead of* has five idioms (*one's time* twice, *schedule*, *the game*, and *time*). The last three are frozen idioms and would need to be matched exactly. For *one's time*, *one's* would involve any possessive pronouns ("his", "their", or "its") or any possessive phrase (e.g., "your body clock 's time").

For each sentence, we find the target <**prep**> XML element, as well as the element's end (</**prep**>), in case we have a phrasal preposition (such as *ahead of*). When we have identified the element end, we see if one of idiom's complement occurs. When we have a match, we tally for the idiom and the tagged sense. After all sentences have been processed for the preposition, we will summarize the results, showing the counts (including those not occurring) and the sense labels (to assess the consistency of the tagging).

We will have tables for 63 distinct (50 single-word, 13 phrasal) prepositions with 1561 unique idioms (with 1891 senses) (see Section 3.1). While we have observed many cases when the idioms are present in the PDEP corpora (mostly, in the TPP corpus, but also in the OEC and FN corpora), we expect that most of the idioms are not present. As mentioned, for those

that are present, the concerns are whether the current tags appear to be correct and whether the tags are consistent when they occur in more than once in the corpora.

Implement algorithm for determining the occurrence of all idioms for each preposition to assess if tagged, using the vertical files for the PDEP corpora.

For idioms not occurring in the corpora, we will need to closely examine the complements of the idiom and the definitions for each sense of the idiom. While the complements are not obvious, i.e., requiring to treat the phrases as idiomatic, there may be a hint or trace of meaning that may link the idiom to an existing sense of the preposition.[10] The definition of the idiom will also be valuable in attempting to link to an existing preposition sense. Also, very frequently, the definition of the idiom may be another idiom. For example, *at large* has 5 senses, one having a definition "at liberty" and another having a definition "at length", both of which are also in PP idiom set. That is, the full set of idioms have some circularity that requires consideration.

As suggesting from this discussion, considerable examination of information about each idiom is needing before concluding that an idiom requires addition to the sense inventory for a preposition. Only then can we infer that the PP idiom is indeed missing from the PDEP sense inventory and that a new sense is required.

Develop a procedure for determining whether each PP idiom not in existing can be tied to the preposition's sense inventory, or whether a new sense needs to be added.

6 Incorporating PP Idioms into PDEP

The two types of PP idioms (preposition initial, Section 3.1 and preposition final, Section 3.2) will involve different processes for entering into PDEP. As indicated above, some of the preposition initial instances are also in the preposition final instances; these will be handled in the latter set. The preposition initial instances will be incorporated into the sense inventory for

[10] Orin Hargraves reminded this suggestion. In addition, some online Merriam-Webster idioms (MW, `https://www.merriam-webster.com/`) first give the meaning of the idiom and then give "word by word definitions" of the component words. For example, "tip the scales at" (= "have a (specified weight)" lists the senses for "tip" and "scales", only hinting at the meaning of the idiom. This is not done for all idioms, such as "buy the farm" = "die".

initial preposition. The preposition final instances will require a new entry in PDEP.

6.1 Processing Preposition Initial Lines

PDEP has a field **lexset**, intended to list lexical items that frequently occur as the complements for a sense. This field has not yet been used; the intention was to list such words using the PDEP mechanisms for examining word-finding and feature-extraction rules (particularly the lemma extraction rule). Instead, other fields (e.g., qualitatively characterizing the complement and identifying feature selectors from the finding rules) sufficed, pending more quantitative methods. Based on the discussion of this paper, the **lexset** field seems appropriate as the place for the complements of preposition initial PP idioms. However, this seems to require elaboration of the procedure for using this mechanism.

With this procedure, the complement will be placed into the **lexset** field. For example, in the idiom *ahead of schedule*, "schedule" would be entered in this field for the pattern "3(1b)" of *ahead of*. This is also justified since this idiom has tagged 10 times with this sense in the TPP corpus. In addition, this pattern is defined "earlier than" and the PP idiom is defined "earlier than expected or required", i.e., a more specialized definition of "3(1b)". Another idiom is *ahead of time* which also has the same definition, so that "time" should also be entered in the **lexset** field. Thus, this field might appropriately have the value "{schedule|time}". This solution, however, does not incorporate all the desirable information for the two idioms.

To characterize the use of the lexical items "schedule" and "time", it is necessary to identify the specialization for "3(1b)", i.e., the addition of "expected or required". This requires perhaps two additional senses, each identified as subsenses to "3(1b)" with the specialization. Since both of the idioms occur in the TPP instances for *ahead of*, the instances with the idioms would need retagging to indicate the subsenses. With the two new senses, it is likely that most of the fields could be copied from the parent sense. More importantly, addition of two new subsenses is similar to the process described above (Section 6.2) for adding new entries into PDEP. That is, such a process would involve the addition of OEC sentences and a sample for the BNC.

The process above (in the previous two paragraphs) would be followed for the 1891 "good" PP idioms (Table 1). Many of the idioms will follow the easier cases as just described. Others will involve more difficulty in placing the appropriate places of the existing PDEP inventory for a particular

preposition. This will be particularly difficult for those with high frequency (e.g., 673 instances beginning with *in* and 391 instances beginning with *on*). We expect that most of these will fall within the existing sense inventory for a preposition. We expect a few will require a distinct addition to the current senses. In any event, the number of subsenses, particularly for the frequent instances, will result in complicated insertions into PDEP.

Characterize the processes of adding subsenses into PDEP, particularly to determine if there are qualitatively different procedures.

6.2 Processing Preposition Final Lines

As indicated in Table 2, there are 141 instances in the preposition final lines that might require addition to PDEP. Section 1 described the procedure originally used to identify the prepositions in TPP and PDEP; we may have missed some prepositions. In the course of the current search, we identified several PP idioms ending in a possible preposition. We then continue using the original criteria: looking for phrases having a **definition** that has (1) a single-word or phrasal preposition definition, (2) a prepositional phrase + a preposition, (3) (an optional leading string) + a transitive present participle, or (4) a leading string + an infinitive of a transitive verb.

This involves examining the Oxford Dictionary for the putative prepositional phrase, looking at the definition(s) as well as any example sentences.[11] In particular, we determine whether the phrase has preposition complements. If so, the phrase would be entered as described in Section 5.1, i.e., expanding PDEP.

As preposition final lines are examined, articulate the criteria used to confirm that an instance should be added to the PDEP inventory.

6.3 Preposition Hierarchy for PDEP

In generating the corpus for TPP (Litkowski and Hargraves, 2007), the sense inventory was expanded by about 10 percent in tagging the FrameNet instances. In tagging the CPA corpus for PDEP, another 10 percent of senses was added. These additions were not entered within an organization structure. The original TPP followed the NODE (Pearsall, 1998) sense numbering. The senses were given an ordinal number, followed by parentheses with the NODE number and a letter for a subsense, The first expansion added "-1" under the sense where a new sense was placed; the second expansion

[11] `https://en.oxforddictionaries.com/`

added the next ordinal in the sense numbering to the TPP numbering, with the letter "n" (for "new") in parentheses.

The original NODE sense ordering (see Pearsall (1998)) contained a two-level hierarchy. Major senses were characterized as core senses, with the first core meaning as "the one that represents the most literal sense that the word has in ordinary modern usage". Subsequent core senses followed this same organizing principle. Subsenses had a logical relationship under the closest core sense. The relationship was a figurative extension, a specialized case, or some other extension or shift retaining one or more elements of the core.

As shown above (Section 6.1) for *ahead of*, the meanings of PP idioms are likely to involve further depth of the hierarchies. When PDEP is expanded to incorporate the effect of PP idioms, from preposition initial and preposition final instances, a further review of the interactions of the sense inventories, within each preposition and in connection with other prepositions (such as substitutable prepositions) seems warranted (similar to digraph analysis as in Litkowski (2002)).

Analyze the subsensing that appears necessary when attempting to handle as many as several hundred PP idioms within a sense hierarchy, perhaps adding a field to describe hierarchical relations.

7 ODE Corpora Data for PP Idioms

As indicated in Litkowski (2013), one constituent corpus of TPP is a set of sentences used for ODE from the Oxford English Corpus (OEC). These are viewed as a sentence dictionary characterizing each sense in ODE, with up to 20 sentences for each sense. ODE is the primary source for online ODE[12]. Each sense usually has a couple of examples and an additional set of example sentences. These example sentences (i.e., the sentence dictionary) are what were combined into the OEC corpus for PDEP.

For the PP idioms (section 3), each item also appears in the online ODE and also has up to 20 sentences for each sense of the idiom. As indicated above, the idioms, their meanings, and their example sentences are not found with the beginning words (prepositions), but can be found under the entries associated with the preposition complements. As shown in Table 1, we have identified 1891 good PP idioms. If they have an average of 10 sentences for each idiom, a new corpus for OEC sentences could have around 19,000 sentences (compared to 7485 in the OEC corpus of PDEP). Such a

[12] `https://en.oxforddictionaries.com/`

large corpus, i.e., already disambiguated, could be valuable for investigating properties of idiom recognition.[13]

Ask OUP whether we can obtain the sentences for these idioms.

As indicated in section 4, many of the idioms occur in the TPP corpus of PDEP. While they constitute as example sentences for the idioms, they were not identified with the Oxford English Corpus, but rather from the British National Corpus. Such cases may be useful in comparing the two different corpora.

8 PP Idioms in the Supersense Reviews Corpus

As indicated above, Schneider et al. (2016) described an annotation of a corpus that also involving tagging PP idioms, triggering this effort to examine the idioms in more detail. Schneider et al. (2017) provides a link to the STREUSLE 4.0 corpus containing English annotations according to supersenses.[14] The parses for the reviews use the CoNLL-U Format (containing 10 columns and adding 9 columns). The 12th column (LEXCAT) is a syntactic category that applies to *strong lexical expressions*. One such category is **PP**, i.e., a prepositional phrase MWE. In these cases, the 13th column (LEXLEMMA) contains the lemmas of the MWE. The 14th (SS) and the 15th (SS2) columns identify adposition supersenses, respectively, the role and the function (often the same). [KC L What criteria were used to identify PP idiomaticity?]

The corpus comprises 3812 sentences. The parses identify 170 PP idioms, of which 95 are unique (with 23 idioms occurring more than once). This suggests that the STREUSLE corpus contains a relatively small portion of PP idioms, compared to the 1891 idioms in Table 1. Of 95 items, only one (*at least*) had more than one supersense (Approximator and Extent), i.e., implying that the idiom is ambiguous. Of the idioms, 45 have a construal characterization, i.e., a difference between the first supersense (the role) and the second supersense (the function). (Appendix C describes how this file was processed.) [KC L Are all PP idioms in the corpus intended to be included in the Adposition Guideline? (Some have not been found.)]

As indicating in the introduction, the motivation for this paper was the possibility that using dictionary definitions might be helpful in identifying

[13] In addition, although most of the PP ending idioms (Table 2) are verb phrases, they could also provide a basis for investigating verb MWEs (VMWEs), with a corpus of more than 20,000 sentences.

[14] `https://github.com/nert-gu/streusle/`

supersenses. We searched each idiom in the dictionary and for its context in the corpus. The immediate observation is that most of the information about each idiom is to be found under the dictionary entry for the complement of the preposition of the idiom. We first examined the 95 idioms in our list of 1891 idioms and found 39 occurring exactly. These idioms will thus be included in our further studies.

We then delved further into the 56 idioms not in our list. We looked more deeply into the online Oxford[12] dictionary and also searched the online Merriam-Webster (MW) dictionary.[10] We were able to characterize 37 as being idiomatic, although not as used in the LEXLEMMA field of the parses. That these were not immediately identified as PP idioms in the dictionary provides an interesting set corresponding to the general kinds of issues pertinent to MWE behavior. In characterizing the items in STEUSLE, the typology below provides only a small sample of the ways in which MWEs may occur, as described in Cieślicka (2015).

- **Non-Idiomatic PPs (9)** Several putative idioms do not appear to be idiomatic. Some have nouns with a meaning that easily fit under the definition of a preposition sense. These include: *among other*, *from standpoint*, *in experience*, *in time of need*, *in traffic*, *on all level*, *to my surprise*, *under warranty*, and *with good reason*. While these MWEs occur frequently (possibly as collocation candidates), their meanings are immediately understood by combining the meanings of the constituent words. [KC L Is there some significance characterizing these phrases as being idiomatic?]

- **Occurring in Other Dictionary (22)** Some MWEs are not in the Oxford dictionary, but occur as idioms in the MW dictionary. With respect to the Oxford, it appears that the meaning of putative idiom can be determined the component words. For example, *in limbo* is not in the dictionary, but **in** has a sense "expressing a state or condition" and **limbo** has a definition "an intermediate state or condition".[15] These include: *as usual*, *behind the wheel*, *from Hell*, *in every way*, *in large part*, *in limbo*, *in love*, *in luv* (by inference), *in pain*, *in town*, *of all time*, *off the bat*, *on a whim*, *on schedule*, *on staff*, *on the phone*, *out front*, *out of business*, *out of town*, *over the phone*, *over the year*, and *up front*.

[15] [KC L Perhaps there is some idiomaticity of identifying the preposition that is linked to the complement, i.e., how do we know that ***in*** should be used for ***limbo***?]

- **Requiring a Determiner (7)** When searching in the dictionary, the form overtly specifies a determiner or possessives modifying the noun. These include: *around corner*⇝*around* ***the*** *corner*, *in hurry*⇝*in* ***a*** *hurry*, *on own*⇝*on* ***one's*** *own*, *on way*⇝*on* ***the***|***one's*** *way*, *out of way*⇝*out of* ***the***|***one's*** *way*, *to face*⇝*to* ***one's*** *face*, *under circumstance*⇝*under* ***the*** *circumstance****s***

- **Variable Forms (7)** The phraseology included in the STREUSLE corpus varies from what is expected in the dictionary. These include: *around the clock*⇝***round*** *the clock*, *at all cost*⇝*at all cost****s***, *by any stretch of the imagination*⇝***not*** *by stretch of the imagination*, *in the best interest*⇝*in the best interest****s*** ***of (something)***, *on a basis* (there are many adjectives for **basis** without a temporal adjective), *on the lookout*⇝***be*** *on the lookout* ***for***, *with fly color*⇝*with* ***flying*** *color*. [KC L Since the STEUSLE corpus has indicated variable MWE forms, is there some manner for describing PP idioms?]

- **Phrasal Preposition, but not PP (3)** Some MWEs are phrasal prepositions rather than prepositional phrases, i.e., not including a complement. These include: *in hope to*, *just about*, *nothing but*

- **Non-Prepositional (4)** Some idioms are not prepositional phrases, but seem to be adverbial MWEs. These include: *back and forth*, *out there*, *up and run*, and *with the way it be*.

- **Inaccurate Tagging (1)** One prepositional phrase can be idiomatic, but the instances in the corpus are not. This includes: *on the side*.

- **Misspellings (3)** These include: *a least*⇝*at least*, *in the mean time*⇝*in the meantime*, and *on line*⇝*online*.

While our detailed examination of possible PP idioms has found problems with some instances in STREUSLE, this indicates some of the difficulties in dealing with MWEs. This suggests the kind of difficulty that may arise in dealing with the much larger set of 1891 PP idioms in Table 1. In addition, in examining the annotations in the corpus, some instances were not tagged as PP idioms, such as *for free* and *in line*, and some, such as *on the side*, did not correspond to the idiomatic PP.

9 Future Work

The possibility that looking up PP idioms in a dictionary has led to a considerable amount of information describing such idioms. Several variations in the types of idioms have emerged, providing a comprehensive overview of the idioms. The overall picture has made it easier to identify further steps and their tasks, with the following priorities:

- Determine the occurrence of idioms in the PDEP corpora
  - For PP idioms in the current corpora, have they been tagged consistently?
- Obtain further corpora instances for PP idioms
  - Obtain Oxford sentence dictionary (from the Oxford English Corpus)
  - Obtain instances from the British National Corpus (to be consistent with the current PDEP CPA corpus)
- Develop procedures for determining how to link PP idioms with the current sense inventory (or determining that a new sense is required)
- Examine idioms that have definitions that are also in the list of idioms and how their circularity should be characterized

It is quite possible that other questions and tasks emerge as these other steps are performed.

Acknowledgments

I am grateful for discussions with Anna Cieślicka, Orin Hargraves, Robert Lew, and Nathan Schneider. They have provided small comments that led me to a considerable amount of effort that has been fascinating. I hope that these efforts have been worthwhile.

References

Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. *In Search of a Systematic Treatment of Determinerless PPs*, pages 163–179. Springer Netherlands, Dordrecht, 2006. ISBN 978-1-4020-3873-0. doi: 10.1007/1-4020-3873-9_11. URL https://doi.org/10.1007/1-4020-3873-9_11.

Anna B. Cieślicka. Idiom acquisition and processing by second/foreign language learners. In Roberto R. Heredia and Anna B.Editors Cieślicka, editors, *Bilingual Figurative Language Processing*, pages 208–244. Cambridge University Press, 2015. doi: 10.1017/CBO9781139342100.012.

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. Double trouble: The problem of construal in semantic annotation of adpositions. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 178–188, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S17-1022`.

Robert Lew. The role of syntactic class, frequency, and word order in looking up english multi-word expressions. *Lexikos*, 22:243–260, 2012.

Ken Litkowski. The preposition project corpora. Technical Report 13-01, CL Research, Damascus, MD 20872 USA, 2013. URL `http://www.clres.com/online-papers/TPPCorpora.pdf`.

Ken Litkowski. Pattern Dictionary of English Prepositions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1274–1283, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-1120`.

Ken Litkowski. The preposition corpus in the sketch engine. Technical Report 17-01, CL Research, Damascus, MD 20872 USA, 2017. URL `http://www.clres.com/online-papers/PrepsSkE.pdf`.

Kenneth C. Litkowski. Use of machine readable dictionaries for word-sense disambiguation in SENSEVAL-2. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 107–110, Toulouse, France, July 2001. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S01-1026`.

Kenneth C. Litkowski. Digraph analysis of dictionary preposition definition. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 9–16. Association for Computational Linguistics, July 2002. doi: 10.3115/1118675.1118677. URL `http://www.aclweb.org/anthology/W02-0802`.

Kenneth C. Litkowski. The Preposition Project. In *Proceedings of the Second ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Colchester, England: University of Essex, April 2005. Association for Computational Linguistics.

Kenneth C. Litkowski and Orin Hargraves. Semeval-2007 task 06: Word-sense disambiguation of prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24–29, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S/S07/S07-1005`.

Judy Pearsall, editor. *The New Oxford Dictionary of English (NODE)*. Clarendon Press, Oxford, 1998.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. A hierarchy with, of, and for preposition supersenses. In *Proc. of The 9th Linguistic Annotation Workshop*, pages 112–123, Denver, Colorado, USA, June 2015.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Meredith Green, Abhijit Suresh, Kathryn Conger, Tim O'Gorman, and Martha Palmer. A corpus of preposition supersenses. In *Proc. of LAW X – the 10th Linguistic Annotation Workshop*, pages 99–109, Berlin, Germany, August 2016.

Nathan Schneider, Jena D. Hwang, Archna Bhatia, Na-Rae Han, Vivek Srikumar, Tim O'Gorman, and Omri Abend. Adposition supersenses v2. *CoRR*, abs/1704.02134, 2017. URL `http://arxiv.org/abs/1704.02134`.

Angus Stevenson and Catherine Soanes, editors. *The Oxford Dictionary of English (ODE)*. Clarendon Press, Oxford, 2003.

A Processing Data of Beginning Prepositions

The 2484 instances in **pp-begin.txt** (see Section 3.1) were entered into a spreadsheet, with each instance in the first column. A new column was inserted before the instances and the instance numbers were created in this new first column. Another column was inserted as the new second column to hold the preposition name. We used the regex "`^[a-z]+`" to obtain the first words of the instances. Using our editor, we made a copy of the matches into another blank file; we copied this list and inserted this list into the second column, hence providing the first words of the idioms. These first words were generally single-word prepositions, but in several cases these were the beginning to phrasal prepositions. We examined this column and modified the cells corresponding to any phrasal prepositions (e.g., *a la* and *ahead of*).

We inserted another column (as the third) after the list of prepositions for use to make a first categorization for each instance. Table 1 identifies the five categories. By identifying the categories, we were able to count the numbers in the table and to examine other characteristics of the PP idioms. We created a tab-separated value file (TSV) (**pp-beg-cats.tsv**).[16] We tabulated statistics about the prepositions in another TSV file, **pp-beg-stats.tsv**. This table shows the number of instances for each of the 63 prepositions. Next, it shows the number of unique putative PP idioms (e.g., the 7 instances for *a la* have only two, *a la carte* and *a la mode*). Finally, the table shows the number of senses for each preposition that are viable PP idioms (i.e., not counting those that are invalid).

We used the categorized sheet in the spreadsheet and used this to copy all lines that were viewed as being good PP idioms, in the file **pp-idioms-good-a.txt**. We created another file of just the first 3 tabbed columns in another file **pp-idioms-good.txt**. We deleted the first two columns of the 1891 lines, leaving the idioms. We could then delete any duplicate lines, to identify the 1561 unique idioms, thus determining the presence of multiple senses.

B Processing Data of Ending Prepositions

Section 3.2 describes the first steps of processing the 2785 instances in **pp-final.txt**. The distinct prepositions and their counts are in the file **pp-**

[16] Using 1 for good PP idioms, 2 for phrases already present in PDEP, 3 for double prepositions, 4 for phrases to be examined for inclusion in PDEP, and 0 for phrases that are not PP idioms

fin-stats.tsv. The categorizations for the instances are shown in the file **pp-fin-cats.tsv**, with the numbered list as described in Section 3.2. This file was used to obtain the instances with each type. For our purposes, we only obtained those already in PDEP (type 1) (182 instances in **pp-final-in.txt**) and those to consider addition into PDEP (type 2) (141 instances in **pp-final-poss.txt**).

C Processing Data for PP Idioms in the Supersense Reviews Corpus

The parsed STREUSLE corpus was obtained from the file streusle.conllulex[14] and is the basis for examining its PP idioms. We searched the 75367 lines using the regex `\tPP\t`, obtaining the 170 lines containing the prepositional phrase MWEs, in the file **Streusle-4.0.PPIdioms.parse**. We entered this file into a spreadsheet and copied the 13th, 14th, and 15th columns, i.e., their LEXLEMMA, SS (role supersenses), and SS2 (function supersenses), and put them into another sheet. We copied this sheet into our editor and sorted the 170 lines alphabetically (**Streusle-4.0.PPIdioms-sorted.tsv**. We then used our editor to delete duplicate lines into the unique lines (**Streusle-4.0.PPIdioms-unique.tsv**). We next entered this file into a new sheet in Excel for use in further analysis of these putative PP idioms (in **Streusle-4.0.PPIdioms-stats.tsv**).

In the **stats.tsv** file, the first column contains the 95 LEXLEMMA unique items. The second column lists the number of instances of each lemma in the corpus (12 for *at least*, 9 for *at all*, 9 for *by far*, 9 for *in town*, and 9 for *on time*). The third column identifies the instances (with "1") that have different supersenses for the role and the function. The fourth column identifies lemmas that were identified in a dictionary (with "1") and those that were not found (with "0"). The fifth and sixth columns include comments about the putative idiom.

We made another sheet in the spreadsheet of the first, fourth, fifth, and sixth columns (in **Streusle-4.0.PPIdioms-analysis.tsv**) to facilitate the focus on the dictionary searches for each item. These comments particularly identify reasons why items were not viewed as idiomatic. Other comments describe non-fixed variations from the items that accepted them as idiomatic.

We used the unique file to identify which items in the first column are present in 1891 PP idioms (in Section A). We identified the 39 items that we found in **pp-idioms-good-a.txt**. We considered the LEXLEMMA items in Streusle and whether they corresponded to the PP idiom wording. In

Streusle 4.0, the idiom words were obtained from the lemma words in the 3rd column for the parse in successive lines after the first word in the lemma. This process resulted in idioms that were not exact. For example, the idiom *a least* consisted of the two lines with the lemmas *a* and *least*. In Streusle 4.1, the idiom was changed to *at least*. We thus infer that the objective is that LEXLEMMA should correspond to what would be found in a dictionary. Among the 56 items, we found that nine have minor variations in ODE, such as *at all costs* instead of *at all cost* and *with flying colors* instead of *with fly color*.

Todo list

- Develop a procedure for determining whether each PP idiom not in existing can be tied to the preposition's sense inventory, or whether a new sense needs to be added. 15
- Characterize the processes of adding subsenses into PDEP, particularly to determine if there are qualitatively different procedures. 17
- As preposition final lines are examined, articulate the criteria used to confirm that an instance should be added to the PDEP inventory. 17
- Analyze the subsensing that appears necessary when attempting to handle as many as several hundred PP idioms within a sense hierarchy, perhaps adding a field to describe hierarchical relations. 18
- Ask OUP whether we can obtain the sentences for these idioms. . 19