Tagging the Pattern Dictionary of English Prepositions with Preposition Supersense Examples

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

December 27, 2021

Abstract

The Pattern Dictionary of English Prepositions (PDEP, Litkowski (2014)) contains 29 fields of data for each of 1040 senses. Guidelines for English adposition and cases supersenses (Schneider et al. (2020)) provide detailed descriptions for 50 categories, particularly indicating prepositions exemplifying the categories (the Preposition Supersense Tags, PSSTs). PDEP includes fields for a **Class**, a **Subclass**, and a **Supersense**. We describe a process that tags the PDEP sense for the preposition in each example in the guidelines with the object for updating the supersenses to the most recent data.

1 Introduction

The several fields in PDEP are intended to provide an overarching characterization of prepositions. A major object is to express spatial or temporal relations or mark various semantic roles. In PDEP (Litkowski (2014)) and the guidelines (Schneider et al. (2020)), their characterization tends to focus on individual pieces of meaning. Several efforts, individually, attempt to provide a comprehensive combination of the various meanings. These two compilations have been developed somewhat differently. In our effort to synchronize the two, we have found another method, simply tagging the examples in the guidelines.

In section 2, we provide the general approach to synchronizing the two compilations. In section 3, we describe the steps using the tagging and the

initial output of this process. Section 4 then describes the output of this process, indicating the linkages between the two compilations. Finally, in section 5, we provide several observations about the linkages, suggesting where further modifications for the two compilations.

2 The Synchronization Problem

(Litkowski, 2014) provides details for the initial fields in PDEP, particularly indicating data generated during The Preposition Project (TPP, Litkowski and Hargraves (2005) and Litkowski and Hargraves (2006)). The TPP lexicographer provided an initial set of the **Class** and **Subclass** fields for each sense in PDEP. Some refinements to these fields were made, showing their current lists of classes[1] and subclasses[2].

The initial PDEP fields of **Cluster** and **Relation** were subsequently added into a **Supersense** field. This field came from before the first guideline version (Schneider et al. (2020)). As stated there, characterizing it as **Adposition Supersenses v2**, referred to version 1 based on a PrepWiki. As a result, 581 have a non-empty supersense, 113 containing multiple supersenses (i.e., containing a comma), and 458 senses with an empty field. This paper focuses on the current values for the PDEP supersense. That is, to fill senses having an empty field, to use the current guideline categories, and to consider what to do about fields having multiple values for the field.

3 Tagging Steps

In working on the synchronization problem, several pieces of information have some possible value. (Litkowski, 2021) is a working paper containing these pieces for each category of the guideline. This paper describes eleven items to characterize each PSST (definition, history, comments, instances, direct tags, senses with the same supersenses, substitutable prepositions, the PDEP hierarchy, previous PSSTs, preposition definitions, and functions in the STREUSLE data). These items are tentatively defined. A table containing these eleven items has been initiated for each of the 50 categories in the guideline. The instances and the PDEP hierarchy have been completed and constitute the discussions of the remainder of this paper.

The instances in the table are the examples of each category in the guideline. These examples were taken from the source text of the guideline for

[1] `https://www.clres.com/db/classes/ClassAnalysis.php`
[2] `https://www.clres.com/db/cllist.php?cl=all`

each category. Examples were included only if the bold item in the example is a preposition in PDEP (i.e., that required disambiguation). Several categories include examples that refer to other categories, intended to clarify the specific category; there were not included in the set of instances for the specific category. Some categories have examples that focus on adverbs or multiword expressions that are not included in PDEP; these examples were not used.

When including an example in the set of instances for a category, the specific preposition was underlined (i.e., changing it from bold). In the guideline formulation, an example frequently viewed as containing a *function* in addition to a *role*. When a category used both the *role* and a *function*, this was included in the set of instances, in case this distinction might be of some further value.

In the working paper, the first link and then as well as the characterization of the eleven items for each category, there is a table describing the examples for each category. The first column contains a 4-digit number showing the guideline hierarchy. The second column contains the category name; it is linked to the table of the eleven items for the category; this is intended to facilitate moving back and forth between the category list and the actual data. The third column shows the number of examples that have been tagged for the category; this runs from 0 to 47 examples, with a total of 594 examples that have been tagged. The fourth column counts the number of distinct prepositions have been used in the examples for the category; overall, there are 237 distinct prepositions, indicating that several examples in a category have used the same preposition.

The fifth column of the table counts the number of senses for the category. This indicates that several examples use more than one sense of a given preposition. There are 345 senses over the 50 categories. This is the essence of the basis for synchronizing PDEP and the guideline. The disambiguation of the preposition for each example involved considerable amount of effort. There were a few instances of monosemous prepositions, but most of the them involved the several polysemous prepositions. Some involved as much as 30 minutes in tagging a preposition. This involved examining the Oxford definitions and examples from the dictionary. The next source were the tags from the TPP lexicographer with the FrameNet corpus. Then, finally, the tags that had been developed in the PDEP CPA corpus. Despite this amount of effort, it is quite possible that others would differ in the tagging. It is important to indicate that making changes might not make much of a difference. Having established the mechanisms for determining the results makes it relatively easy to assess the changes.

4 Linking the Categories to PDEP

Once an example has been tagged, i.e., identifying a preposition and a sense, we can localize the (**Class,Subclass**) it in the PDEP taxonomy. To determine the location, we use the hierarchy script[3] (e.g., in the example below) specifying the guideline category (PATH), the preposition (**over**), and the sense (**11(4)**). This script gets the **Class** (here **Spatial**) and the **Subclass** (here **Path**) in the PDEP data for this sense and then prints a table of all the senses in PDEP with the same class and subclass. In this example, the taxon lists and shows the 39 senses in PDEP for this combination.

- The bird flew **over** the building. (over (11(4)))

As indicated in the table described in the previous section, PATH has 10 examples, with 9 different prepositions and 10 different senses. It would be tedious to enter each such combination. In this example, we first see that many of the examples for PATH already appear in the table, i.e., several of the examples have used prepositions and senses having similar definitions.

This table shows the **Spatial** class and the **Path** subclass. This combination is shown in the class list (footnote 2). As we complete processing with tagged examples in PATH, we see that the category also activates three other **Spatial** subclasses: **SimplePosition**, **SimplePosition:Origin**, and **SimplePosition:Destination**. In the table described above (in section 3), the sixth column counts the number of taxons activated by the examples in each guideline category. The total number of taxons used by the 50 categories is 210 (ranging from 0 to 17), indicating that many taxons were used in more than one category.[4]

When a taxon was activated, the table was converted into a PDF file. This facilitates marking a line by highlighting the preposition and sense and then marking a sticky note to record the guideline category that activated it. There are 66 taxons, but 12 were not used in any guideline category, i.e., none of the senses in these taxons were used in any guideline category; these taxa have no highlights or sticky notes.[5]

[3] `https://www.clres.com/db/hier.php?cat=Path&prep=over&sense=11(4)`

[4] A seventh column in this table counts the total number of senses in all of the taxons for a category. These range from 0 to 343, with the THEME category using 17 taxa. This column was completed, although it is not clear that it has any considerable significance. The primary interest is simply that these numbers are so high, i.e., corresponding one-quarter of the senses in PDEP.

[5] These are **Activity (Proposed)**, **Backdrop (Contrasting)**, **Exception (Exclusion)**, **MeansMedium (Agentive:Negative, Means:Negative)**, **Scalar (Less, Mathematical)**, and **Spatial (Before, Behind, Below, Equal, Presence)**

As indicated above (section 3), each example in the guideline was tagged (disambiguated) without regard to any other example. When a particular taxon was first activated by one example, many other senses in the taxon were very similar to the first one. This suggested that other senses in the taxon, not included in the guideline category, could have just as likely been suitable for examples. For example, in PATH, senses of *along* or *down* could also have been used to exemplify this category.

Two other situations arose as the tagging proceeded. First, several senses in a taxon are activated by the examples in different guideline categories, belying the seemingly synonymy of the taxon. It appears that this is the case for almost all of the taxons.[6] Second, many of the senses have been activated by more than one guideline category.[7] These situations will be discussed in the next section.

5 Observations of Synchronization Linkages

The initial reaction to the previous situations might be that there are many problems with the PDEP taxonomy and the guideline SNACS categories. Instead, further introspection suggests that these situations contain richer aspects of the preposition ambit. This suggests that further investigation is necessary to characterize what is going on with these issues.

5.1 Senses Tagged in Examples

As indicated above (footnote 7) describing use of multiple categories in tagging 594 examples, 208 different senses were used (out of 1040 senses in PDEP).[8] The used senses came from 56 distinct prepositions. Many of these (24 prepositions) used more than one sense of the preposition.[9] It is noticeable that these are the most polysemous and that almost all of the senses in these prepositions were tagged in the examples.

[6] Only 8 of the 54 activated taxa had only one guideline category. These taxons have only a small number of senses.

[7] The 594 examples used 208 senses; 73 of the senses have more than one category; the remaining 135 senses were tagged by only one category.

[8] The used senses are available in a CSV file (`http://www.clres.com/online-papers/usedsenses.csv`). This file contains six columns: the PDEP class, the PDEP subclass, the tagged preposition, the preposition sense number, the PDEP definition of the preposition and sense, and the guideline categories that were used in the examples.

[9] *on* (18), *of* (17), *with* (16), *from* (14), *to* (14), *for* (13), *by* (12), *in* (12), *at* (10), *over* (7), *out of* (6), *through* (5), *among* (4), *between* (4), *after* (3), *into* (3), *like* (3), *off* (3), *via* (3), *about* (2), *across* (2), *onto* (2), *within* (2), *without* (2)

Putatively, the totality of the senses that were tagged constitute the full coverage of preposition meanings. These senses can be examined in more detail. Many of the activated senses are defined with either single prepositions or phrases that end with another preposition. This suggests that such senses are not basic, i.e., they are derived from more basic prepositions. For example, *on top of* (sense 2(1a)) is defined as "so as to cover; over", suggesting that *over* is a hypernym to *on top of*. Complete digraphs were developed during TPP (Litkowski (2009)); these can be refined with updates based on PDEP data.

5.2 The PDEP Hierarchy

The initial set of classes developed in TPP were refined during the PDEP tagging.[10] In particular, several initial classes were combined as appeared necessary. The reasons for changes from the original are described in the class analysis, refinements as TPP evolved to PDEP. Notwithstanding, further investigation of these classes and subclasses is warranted based on activation of several supersenses in a taxon (see footnote 2). As indicated above, the senses in a taxon were expected to be semantically similar.

There are several questions that can be addressed. A first question is to understand the multiplicity of supersenses in a taxon. How are such activated senses related or interrelated to each other? How do the non-activated senses relate to those that were activated by the examples? A second question is to understand how the senses in a taxon are hierarchical to one another. As suggested in the previous subsection, how do the digraphs for the senses relate to other taxa in the PDEP hierarchy?

5.3 Taxons not Activated

There are 105 senses in the 12 taxons that were not activated (footnote 5). The largest group (**Scalar (Less)**) has 27 senses; these are similar to the other taxon (**Scalar (Greater)**) whose senses were frequently activated by the ComparisonRef category. Five of these taxa are spatial senses; these seem to be similar to the Locus category and not sufficiently requiring distinctive subgroups. The taxon (**Backdrop (Contrasting)**) has 22 senses that do not seem to be covered in the guideline. These senses correspond to prepositions such as *despite* and *in spite of* and others with similar definitions. Most of the other non-activated taxa have only a few senses that might be viewed as close to other guideline categories.

[10] `https://www.clres.com/db/classes/ClassAnalysis.php`

Despite this discussion, the number of senses that were not activated constitute 10 percent of the PDEP senses. Perhaps, further investigation is necessary about these taxa, to determine where they should be located in the guideline hierarchy.

5.4 Multiple Supersenses for a Preposition Sense

As indicated above (footnote 7), 73 of the 208 senses activated in the example have more than one supersense.[11] For example, sense (6(4a)) of *in* ("indicating the quality or aspect with respect to which a judgement is made") is tagged in 7 guideline categories (LOCUS, MANNER, CHARACTERISTIC, CIRCUMSTANCE, TOPIC, THEME, STIMULUS). Looking only with the supersense category names, the definition has some tinge with these supersenses. That is, arguably, all of the supersenses are correct.

Clearly, examining the multiple supersenses seems desirable to be sure. But, assuming that many or most of the multiples are valid, the questions is what they imply. One possibility is that the definitions for the prepositions contain multiple components that have been used to construct them.

6 Future Work

Next steps are to carry out the analyses that have been suggested above and to determine how these results are incorporated in the appropriate PDEP fields.

References

Ken Litkowski. Analysis of preposition classes. Technical Report 09-01, CL Research, Damascus, MD 20872 USA, 2009. URL `http://www.clres.com/online-papers/PrepositionClasses.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`.

[11] 44 (2), 17 (3), 4 (4), 3 (5), 3 (6), 2 (7)

Ken Litkowski. Supersense V2 Instance Maps. Technical Report 21-02, CL Research, Damascus, MD 20872 USA, 2021. URL http://www.clres.com/online-papers/PSSTMap.pdf.

Ken Litkowski and Orin Hargraves. 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*, pages 171–179, Colchester, England, University of Essex, April 2005. Association for Computational Linguistics.

Kenneth C. Litkowski and Orin Hargraves. Coverage and inheritance in the preposition project. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, 2006. URL https://www.aclweb.org/anthology/W06-2106.

Nathan Schneider, Jena D. Hwang, Archna Bhatia, Vivek Srikumar, Na-Rae Han, Tim O'Gorman, Sarah R. Moeller, Omri Abend, Adi Shalev, Austin Blodgett, and Jakob Prange. Adposition and case supersenses v2.5: Guidelines for english, 2020. URL http://arxiv.org/abs/1704.02134.