Updating Supersenses in the Preposition Pattern Dictionary

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

August 8, 2018

Abstract

One field in the Pattern Dictionary of English Prepositions (PDEP) is used to characterize supersenses. Schneider et al. (2015, 2016) proposed in PrepWiki (henceforth "v1") a supersense inventory of 75 categories proposed for English for adpositions for use in a corpus. Hwang et al. (2017b) have developed a new "v2" inventory of 50 categories and have been using it to annotate the same corpus. This paper describes the procedures to map "v1" to "v2".

1 Introduction

The Pattern Dictionary of English Prepositions (PDEP, (Litkowski, 2014)) contains several fields that provide collections of senses (**Class**, **Subclass**, **Relation**, **Supersense**, **Cluster**, and **Relation**). These collections may be viewed as coarse-grained semantics of the individual senses in the dictionary. Preposition disambiguation may be useful in using different granularities.

The **Supersense** collection, described in Schneider et al. (2015, 2016), was developed in PrepWiki[1] and used as the first version of 75 supersenses included in PDEP. Hwang et al. (2017b) have developed a new "v2" inventory of 50 categories. This paper describes procedures for updating the data into the new inventory.

[1] http://tiny.cc/prepwiki. [$^{KC}_{L}$ It is very difficult to manipulate the PrepWiki data. It frequently takes an hour to display anything.]

Section 2 describes the procedures used to enter the "v1" supersenses into PDEP. Section 3 describes the data characterizing the two supersense inventories that are relevant for mapping to the new inventory. Section 4 describes the procedures for establishing the new data in the supersense field for characterizing preposition complements. Section 5 identifies difficulties in some-to-some rather than one-to-one mapping and when "v2" supersenses do not occur. Section 6 describes the use of the "v2" inventory in characterizing construals for representing preposition governors.

2 Initial Creation of Preposition Supersenses

The initial supersenses were developed at PrepWiki. Each category, a supersense tag (e.g., **SST-Direction**), was defined and situated with other categories. A table identified elements for the category, consisting of a preposition sense, its definition, one or more examples, and the supersenses. The sense number and the definition come from the PDEP data.[2] Some senses have multiple supersenses, e.g., for *off (4(3a))*, "so as to be removed or separated from" has three supersenses, **SST-Direction, InitialLocation, Source**.

The PrepWiki categories were harvested into several JSON objects.[3] The main JSON object (in **prepwiki.json**) was characterized as TPP_PSST with each member consisting of a sense-value pair such as **"into 4(4)": "Direction"**. This object was used as a variable in a PHP script designed to synchronize with the supersense PDEP field.[4] Initially, the supersense field was empty and the script would identify the sense-value pairs that were not filled (i.e., the supersense that needed to be filled in the database). As we used the PDEP editor to enter the supersense into the appropriate field, the script would identify those not yet entered. These were quickly entered into PDEP, with 143 distinct supersense values, in **psst-v1-pdep.txt**, and 581 senses with supersense values, in **psst-v1-pdep-all.tsv** (with 459 senses not having supersenses identified, in **pdep-no-psst.tsv**).

[2] Several senses are not prepositions, but rather adverbs, particles, or other forms. The discussion in this paper refers only to preposition senses.

[3] From Vivek Srikumar and Nathan Schneider from November 2014 through September 2015.

[4] `http://www.clres.com/db/syncpsst.php`. See `http://www.clres.com/db/syncpsst-Copy.php` that was used add the supersense field for each item.

3 The Second Version of Supersenses

Hwang et al. (2017b) initiates the second version of supersense, importantly presenting the notion of adposition construals. Guidelines (Schneider et al., 2017)[5] have been developed to characterize each of the 50 supersenses in considerable detail. A supersense-to-PDEP map was not included in these guidelines. This paper is designed to provide such a map.

To prepare a map, we first examined the current set of supersenses, generating several subsets to help in understanding how the second version set might relate to the original supersenses.

3.1 Describing Current PDEP Supersenses

The several papers describing the initial set of supersenses had some differences. We first had to identify the fixed set to use for any mapping. We use the PrepWiki version of the 75 supersenses (**psst-v1-wiki.txt**). But, in addition, PrepWiki also identify 85 "SST conflations", i.e., categories consisting of two or more supersense combinations, as mentioned above in section 2. Some changes had been made from the final set used in the last PDEP updating, which included four supersenses not in the PrepWiki version.[6]

PDEP does not include all of the supersenses and conflation categories. At present, there is a total of 143 different supersenses, 62 containing "pure" (i.e., single-word) supersenses (**psst-v1-pdep-pure.txt**, which includes the 4 supersenses not in the PrepWiki version) and 81 using conflated supersenses (**psst-v1-pdep-conf.txt**, i.e., having a comma in the field)[7]. An example of a conflation is ***across 2(2)*** which has Location, 1DTrajectory in the PDEP supersense field. There are 17 supersenses that do not occur as pure in PDEP, 10 occurring as part of conflations[8] and 7 not occurring at all[9].

To characterize the current PDEP supersenses, we first created the list of supersenses from the PrepWiki. Next, we created a PrepWiki table of all single-word supersenses and their "short" definitions (**psst-v1-defs.txt**); this was done to compare the definitions of second version. We created a

[5] `https://arxiv.org/pdf/1704.02134v3.pdf`

[6] Cause, Other, Partitive, and Speaker

[7] PrepWiki lists 85 conflations; perhaps some of these don't correspond to preposition senses, but rather to adverbs or particles.

[8] Contour, Co-Participant, Co-Patient, Donor/Speaker, Experiencer, Purpose, Recipient, Reciprocation, Transit, and Whole

[9] Affector, Configuration, Participant, Path, Place, Traversed, and Undergoer

text of the 143 PDEP supersenses and two subsets: the 62 "pure" supersenses (**psst-v1-pdep-pure.txt**) and the 81 conflation supersenses (**psst-v1-pdep-conf.txt**). These lists are designed to help comparison with the "v2" supersenses.

3.2 Comparing "v2" to "v1" Supersenses

We used the guideline paper for listing (Schneider et al. (2017)) (**psst-v2.txt**) since this also provide the definitions of the 50 supersenses. As for the "v1" supersenses, we also created a list of the short definitions for "v2", using the boxes from each supersense in the guideline (**psst-v2-defs.tsv**). This provides the basis for comparing the two sets.

We first observed that 36 of the "v2" supersenses occur in the "v1" set and 14[10] do not occur in the "v1" supersenses. Thus, it is necessary to examine the 39 supersenses that no longer occur in the current set. We created the list of 36 supersenses that are the same for both sets (**psst-v1-same.txt**) and the list of 39 supersenses not in "v2" (**psst-v1-merged.txt**).

For the common set of supersenses, we examine whether the definitions in two sets are the same (**psst-v1-same.txt**). The first assumption is that they are the same. Only a few supersenses have identical definitions. Many of the supersenses (perhaps 15) differ slightly that would seem to be the same. The remainder (about 17) have a more elaborate characterization, but essentially the same meaning, with a sharpened definition. Based on this discussion, it would be valid to use any PDEP supersenses with the same name.

For the other 39 supersenses, the objective is to map the "v1" supersenses into the "v2" supersenses (**psst-v1-merged.txt**). There are several pieces of information (from the guideline) that can be used to determine how to make the mapping. Sections 1.4 and 1.5 describe some of the major changes from "v1", particularly identifying several supersense names that were simply removed as being not productive for tagging. In addition, in the discussion of individual supersenses, there was a discussion in ***history*** blocks that describe maps from "v1" to "v2".

Seven supersenses were overtly characterized as having been removed[11] and five were not discussed at all (with the impression that they were re-

[10] Characteristic, ComparisonRef, Cost, Gestalt, Identity, InsteadOf, Inverval, OrgRole, Originator, PartPortion, Possession, RateUnit, SocialRel, and Stuff

[11] Comparison/Contrast, EndState, Scalar/Rank, StartState, State, Value, and Value-Comparison

Type	**Code**	**PSSTs**	**Senses**
Same Name	Same	29	146
Map to One PSST	Map1	20	208
Identical Merging	Ident	6	7
New Conflation	Conf	57	84
One Left PSST	OneLeft	15	18
Removed from “v2”	Removed	10	91
Not Discussed in “v2”	NotDisc	2	17
Not in PrepWiki	NotPW	4	10
Total		143	581

Table 1: PDEP Distinct “v1” Supersenses and Affected Senses.

moved)[12]. The other 27 supersenses are specifically mapped to one of the single-word “v2” supersenses. It is not simply the elimination or merging of some supersenses, but also has involved renaming in the “v2” set. In addition, three supersenses (Participation, Configuration, and Temporal) are intended only to organize subtrees of the “v2” hierarchy and not to be used directly.

4 Updating the PDEP Supersense Field

As indicated above, PDEP currently has 581 senses with 143 distinct supersenses. The first step is to determine what should be the “v2” value for each of the current supersenses. This requires identifying the new supersense for each current supersense. To do this, we see that several current supersenses have similar patterns (described in **pdeppsst-v1.tsv**).

To construct the map (in **psst-v2-new.tsv**), we first listed each of the 143 PDEP supersenses (in the 3rd column). Next, we counted the number of senses with each supersense (in the 2nd column). We then characterized the type of mapping that should be applied (in the 1st column) to each supersense, based on the analysis in the **same** and **merged** files. Finally, we determined how to construct a combination of the “v2” supersenses (in the 4th column). The procedures for each type of mapping are described in the paragraphs in the remainder of this section. Table 1 summarizes these results.

[12] Affector, Elements, Place, Superset, and Undergoer

4.1 Mapping Criteria for Each Supersense

We identify that 29 supersenses have the same supersense ("**Same**") in both the current inventory and the new inventory, corresponding to 146 PDEP senses. These correspond to 29 of the 36 supersenses that occur in both inventories. There are no occurrences for seven of the supersenses, i.e., not occurring in PDEP senses: Configuration, Experiencer, Participant, Path, Purpose, Recipient, and Whole. We presume that the meaning of the common PSSTs are essentially the same, with some sharpening in the new definitions.

A second pattern occurs when a "pure"[13] "v1" supersense does not occur in the "v2" inventory but is mapped into a "pure" supersense ("**Map1**"). This is the case for 20 supersenses, covering 208 senses. One sense (mapping Location to Locus) covers 86 senses. Five of these (1DTrajectory, 2DArea, 3DMedium, Course, and Via) are mapped into one supersense (Path).

Subsumed conflations consist of two or more supersenses, of which all members are identical in merging ("**Ident**"). For example, the conflation 1DTrajectory, 2DArea each have been mapped into Path. Six supersenses occurred in this type, also affecting 7 senses.

To map conflations, each component PSST is considered individually based on its mapping, generating a new conflation using the "v2" inventory ("**Conf**"). For example, the conflation DeicticTime, Location is mapped to the conflation Interval, Locus. There are 57 conflations that were subjected to this type of mapping, corresponding to 84 PDEP senses. In 45 cases, the conflation was used in only one sense in PDEP, with the others occurring a small number of senses.

Fifteen PSSTs consisted of conflations all of which but one supersense ("**OneLeft**"). For example, in the conflation InitialLocation, StartState, the PSST StartState had been removed from the "v2" inventory. As a result, we removed the eliminated PSST, leaving only one (in this case, mapped to Source), as indicating in the guideline. These cases corresponded to 18 PDEP senses.

In describing the new inventory, seven supersenses were removed as being no longer meaningful ("**Removed**"), as indicated above[11]. Also, three conflated supersenses have two supersenses that were removed. These 10 PSSTs cover 91 PDEP senses, three of which corresponded to a large number of PDEP senses: Comparison/Contrast (23), Scalar/Rank (29), and State (20). An additional two supersenses (Elements (14) and Superset (3))

[13] I.e., not conflated with another supersense.

were not discussed in the guideline, suggesting that they were effectively removed ("**NotDisc**"); these were use in 17 senses.[14]

As indicated above[6], four supersenses had been removed in PrepWiki ("**NotPW**"). These occur in 10 senses in PDEP; these are for senses having orthographic variants in which senses were copied from the base preposition. These will be updated from the senses at the base preposition.

4.2 Obtaining the Tentative Final Supersenses

From the 143 "v1" supersenses, we eliminated the 16 supersenses from those identified in the **Removed**, **NotDisc**, and **NotPW** categories. We then obtained the new 127 "v2" supersense names (the 4th column of the **new** file) in **psst-v2-refined.txt**. We then removed duplicate names to obtain a final list of the 89 tentative supersenses (**psst-v2-final.txt**). The resulting list contains 37 pure supersenses and 52 conflation supersenses (combination of 2 or more pure names).

5 Overview of V2 PSSTs in PDEP

Several questions need to be addressed in further analysis to make further progress in mapping from the "v1" inventory to the "v2" inventory. The first overview indicated that there are 89 PSSTs in the final map (down from 143), with 37 having "pure" PSSTs (down from 62) and 50 are conflations (down from 81).[15] The initial look indicates that 13 of the 50 "v2" PSSTs do not occur in the mapping. These are ComparisonRef, Configuration, Cost, Experiencer, Gestalt, InsteadOf, OrgRole, Participant, PartPortion, Possession, RateUnit, Recipient, and Stuff.

The major issue pertains to the PSSTs that were removed or not discussed in "v2" with respect to the "v1" inventory, as described Table 1. The affected PSSTs on both sides of the inventories are not one-to-one mapping, but rather involve some-to-some mapping. These will require further analysis, perhaps just a matter of grouping subsets by examining affected PDEP senses. Looking at the PSSTs in the previous paragraph suggests that this is the case.

As indicated above, several PSSTs were identified as having been removed, and when so, these removals also affected conflations. Mostly, these

[14] [$^{KC}_{L}$ Perhaps the appropriate maps from "v1" to "v2" can be determined by finding cases with the old PSSTs and look to see how these cases have been handled in the STREUSLE 4.0 files.]

[15] [$^{KC}_{L}$ Are conflations still valid in "v2"?]

just reduced the number of PSSTs in the conflation, but others eliminated all PSSTs (e.g., in EndState,State both components were removed, leaving no supersense for one PDEP sense, *to (5(2)* defined as “approaching or reaching (a particular condition)”). In any of the cases where some or all PSSTs have been removed, we can identify the senses affected and study the corpus instances to see what PSSTs might be appropriate.

[$^{\text{KC}}_{\text{L}}$ After making the discussion of how to assess the maps for supersenses had been removed, several additional questions arose. These questions pertain to all map types in Table 1. In the Streusle 3.0 files, the file **psst-tokens.txt** contains the tags for each preposition. Presumably, e.g., Attribute is mapped to Characteristic. However, when we examine the Streusle 4.0, the file **Streusle.conllulex**, we can see the PSST “v2” tags and there is considerable variation, rather than the suggested map. In other words, it is likely that the mapping would appear to have **confusion matrices**. I don’t know what this means or how to assess these situations.]

6 Incorporating Construals into PDEP

Based on the previous discussions mapping to “v2” PSSTs, the new supersense fields will use essentially the same procedures as described in Section 2. However, as described in Hwang et al. (2017a) and Schneider et al. (2017), the “v2” inventory adds a **role construal** characterizing a **function construal**. [$^{\text{KC}}_{\text{L}}$ What should be placed in the PDEP field? When both a scene and a function are applicable, how should this be described in PDEP? Should there be a new field?]

In the guideline, each PSST is described via a paragraph that includes a definition and several example sentences. The examples also identify applicable prepositions. Frequently, the discussion will add a construal. For example, in the paragraph describing Whole, there are four PSST construals as a function and four others as a role. It is not clear where the paragraph should be placed in the PDEP supersense field.

These are not inventoried directly in the guidelines and it is not clear how to identify all valid combinations of supersenses using role-function possibilities. An initial set of valid combinations can be obtained from the corpus that have tagged with approximately 4250 adpositions (see STREUSLE 4.0[16]). Of the adpositions, 702 were identified as construals (i.e., had differing role and function supersenses, using 105 distinct combinations).[17]

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

[17] STREUSLE 4.1 has 869 construals; these have not yet been further examined in detail.

References

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. Coping with construals in broad-coverage semantic annotation of adpositions. *arXiv:1703.03771 [cs.CL]*, March 2017a. URL http://arxiv.org/abs/1703.03771. arXiv: 1703.03771.

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 2017b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S17-1022.

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.

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.