Preposition Analysis Using Correspondence Analysis

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

November 23, 2020

Abstract

Several proposed characterization of preposition sense groupings have been developed over the years. In general, such groupings have involved discussions of fine- and coarse-grained senses. All of these discussions are based on qualitative judgments, beginning with lexicographers creating dictionaries and continuing with computational linguists using distributional methods. Correspondence analysis (CA) offers a different approach for examining sense similarities, using features developed in parsing preposition instances. CA methods first provide graphical visualizations of the similarities and then provide quantitative distances between senses, analyzing the variances of contingency tables in the expected values. We examine these methods in enhancing characterizations of preposition behavior patterns.

Contents

1 Introduction

Litkowski (2019) discussed future plans for the Pattern Dictionary of English Prepositions (PDEP, Litkowski (2014)), subsequent to honing the files installed into Sketch Engine[1](SE). These plans described improvements in PDEP supersenses, reviewing the corpora tagging, completing fields in the preposition patterns, analyzing substitutable prepositions, and extending preposition idioms in multiword expressions. As stated, there was no ordering for these areas. Several aspects of these tasks have been started, but now it seems that concrete steps have suggested that the tasks can be integrated, and in a way that may lead to a novel perspective for analyzing similarities of preposition senses.

Completing further fields in the preposition patterns had generally involved judgment of how each field should be filled[2]. The most basic field involves determining the general part of speech of the preposition's complement and governor. For the complement, PDEP envisioned common nouns, proper nouns, wh-forms, and gerunds, as well as the possibility of indicating that a sense complement could be a small set of lexical items. For the governor, PDEP envisions that a noun, a verb, or an adjective governed the prepositional phrase. The PDEP software included several kinds of interactive analysis that could be used to help fill the various fields, particularly using features that were generated in parsing the corpora. Instead of examining the corpora associated with the tagged senses, one at a time, it was

[1] `https://www.sketchengine.eu/`
[2] `https://www.clres.com/db/TPPEditor.html`

clear that writing some simple scripts could examine all the corpora at one time.

The first script[3] simply created tables of parts of speech for each sense for each preposition, for each of the three corpora. These tables are cross-tabulations, suitable for using correspondence analysis (CA), prompted by McGillivray et al. (2008), as further described in Greenacre (2017), summarized in Appendix A. CA provides spatial visualizations of cross-tabs showing the multidimensional relations for the preposition senses based on a singular-value decomposition of the table. In particular, the CA analysis shows how the senses of a preposition are related to each other. In this paper, we describe how each of the planned improvements can built on the correspondence analyses.

Section 2 details how tables are generated from the features that were used in support-vector machines used in modeling the preposition sense disambiguation. Section 3 provides a more detailed multiple correspondence analysis that allows examination of individual corpus instances. Section 4 shows how it is possible to compare independent tagging against the dictionary definitions for each of the senses. Section 6 enables an examination of the **substitutable prepositions** field in PDEP, to allow the null hypothesis that the features across these substitutes are essentially highly similar. Section 7 allows similar examination of the PDEP field for **supersenses**, also allowing the examples used in the guidelines for supersenses in Schneider et al. (2017). Section 8 discusses multiword expressions (MWEs) added to PDEP, not included in the original sense inventories for 70 prepositions; these as well as other MWEs need to be analyzed in conjunction with the base sense inventories for these prepositions. Section 9 describes techniques for comparing the tagging of each instance in the CPA corpus, based on the distance to the other senses.

2 Cross-Tabulation of PDEP Features

Simple correspondence analysis (as described in Appendix A.1) begins with the generation of cross-tabulations in contingency tables, with sense lists in the rows and features (such as parts of speech) in the columns.[4]

As described in Litkowski (2016), PDEP parsed 81509 sentences, using a

[3]The script featanal.py is available at `https://github.com/kenclr/ca4pdep`. The project **ca4pdep** contains details of the processes, code, and data used in this paper. The feature files can be downloaded from `https://www.clres.com/db/feats/` (not yet).

[4]See https://www.clres.com/ca/pdepca01.html for code and output in this section.

Table 1: Parts of Speech for Complements of *above* in CPA Corpus

CPA	cd	dt	jj	nn	nnp	nnps	nns	pdt	prp	vbg	wp
1(1)	0	0	0	23	1	0	2	0	2	1	0
2(1a)	0	2	0	8	0	0	2	0	2	0	0
3(1b)	0	0	0	10	1	0	0	0	2	0	0
4(2)	0	1	0	10	0	0	4	0	1	0	0
5(2a)	0	1	0	3	2	0	0	0	1	0	0
6(2b)	0	1	0	5	1	0	1	0	2	1	0
7(2c)	0	0	0	0	0	0	1	0	0	1	0
8(2d)	0	0	0	5	0	0	0	0	0	0	0
9(3)	18	3	1	33	2	1	13	0	2	1	1
10(n)	0	48	2	1	0	0	3	3	0	0	0

dependency parser (Tratz and Hovy, 2011), each focused on one preposition. On average, about 1250 features were generated for each sentence; for a typical set of 250 sentences for a preposition, about 70,000 distinct features were generated. Features are comprised of three components, (1) a word-finding rule (**wf**), (2) a feature extraction rule (**fer**), and (3) the value of the feature (**wf:fer:**). The initial cross-tabulation looks at the feature (**hr:pos:**), the part of speech of the heuristic identification of the preposition complement. The Python script above created a table of the parts-of-speech for each preposition sense for each corpus (CPA, OEC, and FN[5]). Table 1 shows the table for *above* in the CPA corpus[6] [7]. In this case, there were 250 instances in the corpus, but only 229 were prepositions and the remaining 21 were adverbs.

A cursory examination of Table 1 provides some indication of how the pattern for each sense might be marked in the editor, e.g., noting the presence of cardinal numbers for sense 9(3) and determiners for sense 10(n). Since neither of these emphasized parts of speech are currently checkmark options in the pattern manager, describing the behavior requires character-

[5]Corpus Pattern Analysis, Oxford English Corpus, and FrameNet

[6]Senses: 1(1) in extended space over and not touching; 2(1a) extending upwards over, 3(1b) higher than and to one side of; overlooking; 4(2) at a higher level or layer than; 5(2a) higher in grade or rank than; 6(2b) considered of higher status or worth than, too good for; 7(2c) in preference to; 8(2d) at a higher volume or pitch than; 9(3) higher than (a specified amount, rate, or norm); 10(n) more so than anything else

[7]Parts of Speech: cd (cardinal number), dt (determiner), jj (adjective), nn (noun (sing. or mass), nnp (proper noun, singular), nnps (proper noun, plural), nns (noun (plu.), pdt (predeterminer), prp (personal pronoun), vbg (verb, gerund or present participle), wp (wh-pronoun)

ization in the Selectors box.

The first question about this table is whether there is any difference between any of the senses. It is difficult to discern the differences among the other senses by inspection. A chi-square test determines if the distributions of the categorical variables differ from each another, i.e., testing the null hypothesis that there is no difference. This is the beginning step in correspondence analysis. In examining a cross-tabulation, if the null hypothesis is true, the observed and expected frequencies will be close in value. In Table 1, the question is whether the several senses have similar behavior. In this case, the chi-square statistic χ^2 is 286.95, indicating that the patterns are different. The chi-square divided by the sum of the table (229) is known as the (total) inertia (ϕ^2), characterized the variance in the table, in this case equal to 1.253063.

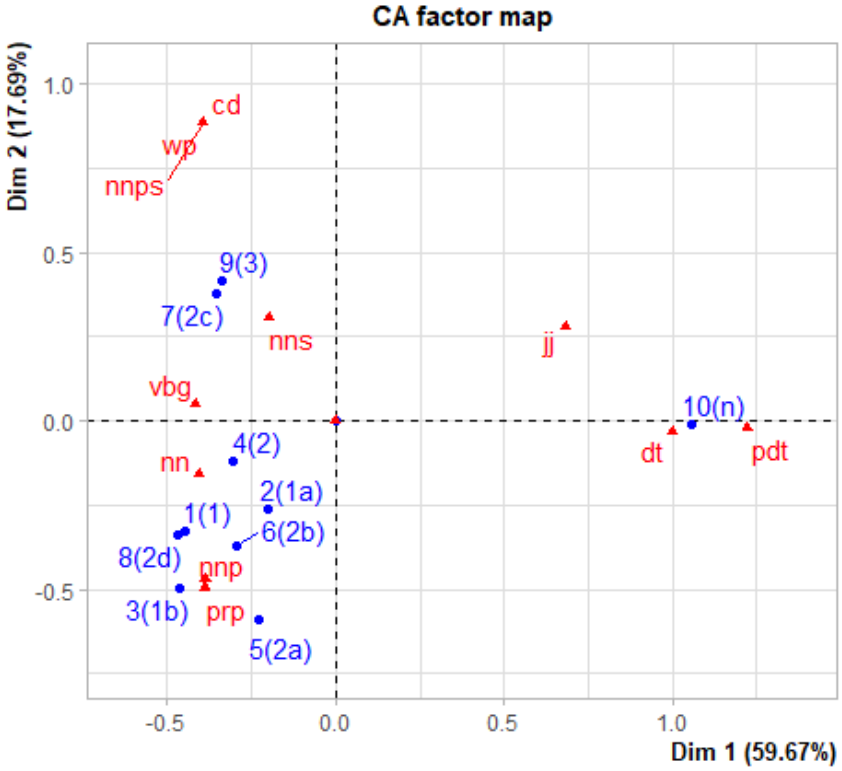

Figure 1: Sense Similarities Based on CPA Corpus

The objective of CA is to determine where the variance lies in the table.

The first step creates an independence or correspondence matrix (CM), dividing each cell by the sum of all the cells, so that the sum of the cells of the CM is equal to 1. This is a matrix of standardized residuals, which is used to perform a singular value decomposition (SVD) into its factorizations. With the eigenvalues from the SVD, the CA can be used to identify the contribution for each of the rows, columns, and cells into their contributions to the inertia. This permits a plot of the rows and columns of the original table, in this case shown in Figure 1. The figure visualizes how the senses and the parts of speech relate to one another.[8]

Usually, the result summary first identifies the chi-square value. Next, the summary identifies the eigenvalues resulting from the singular value decomposition of the table, particularly showing the variance for each dimension. The total inertia (the sum of the variances) is 1.253063. A summary next provides the details for each row and each column, showing an analysis of these details. First, a column identifies the inertia for each row or column; the sum of these individual amounts is equal to the total inertia. Thus, it is possible to see, in this case, which senses and parts of speech have the largest portion of the variance. For Table 1, this indicates that senses 10(n), 9(3), and 7(2c) and the parts of speech "dt", "nn", "cd", and "vbg" account for the greatest variances. The summary results next identify where each sense and part of speech should be placed in factor map. These locations correspond to the first two dimensions for each

(here)From the visualization, several further observations can be made and analyzed further. As indicated above, sense 10(n) seems to be somewhat different from the others. The figure shows that sense 10(n) is very extreme, very different from the others, probably based on the likelihood that the sense is idiomatic ("above all"), with its complement either a determiner or predeterminer. Having made this observation, it is possible to drop this sense, and thus provide a better idea of how the blob can be distinguished. In dropping that sense, it is necessary to drop the "pdt" (predeterminer) column, since dropping 10(n) leaves only 0 values and would result in a degenerate matrix for the SVD. With the smaller table, χ^2 is 96.46, still rejecting the null hypothesis.

The plot (not shown) that removes sense 10(n) still shows a blob of most of the senses, with outliers for sense 7(2c) ("in preference to") and the part of speech "vbg" (gerundial). Examining Table 1, we see that this sense

[8]There are several packages for correspondence analysis, particularly in R, one in Python, and others in various statistical software. They can also perform the components in CA or can be implemented by developing the computations.

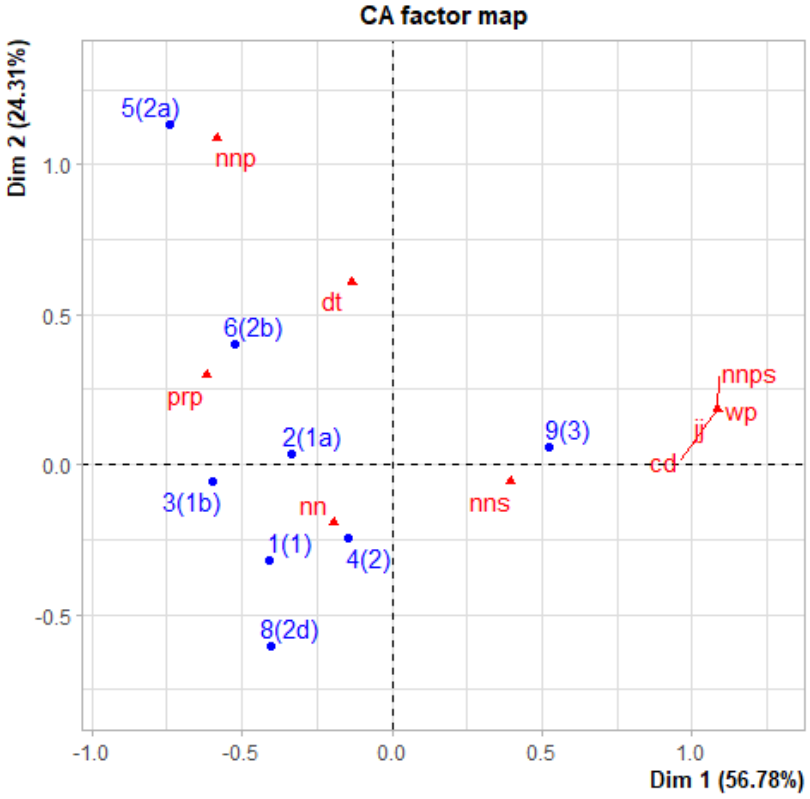

Figure 2: CA Factor Map with Some Removed Instances on CPA Corpus

has only 2 instances and the part of speech has only 4 instances. With these small counts, it seems that their significance in the plot appear too dominant, suggesting that these outliers can be dropped. With the smaller table, χ^2 is 68.49, still rejecting the null hypothesis. Figure 2 shows the resulting plot. This figure shows more spread of the senses. The first set of senses (1(1), 2(1a), 3(1b)) are close to one another; the second group (4(2), 5(2a), 6(2b), 8(2d)) are spread out, but also in the negative hemisphere, suggesting similarity; and the third group (9(3)) is alone in the positive hemisphere.

Table 2: Instance Data for *above* in CPA Corpus

Inst	S	C	G
c1	3(1b)	prp	vbd
c2	1(1)	nn	nn
c3	9(3)	nns	vbd
c4	9(3)	prp	vbz
c5	9(3)	nns	vb
c6	1(1)	nn	vb
c7	10(n)	dt	cc
c8	9(3)	nn	vbp
c9	9(3)	cd	nn
c10	10(n)	dt	vbz

3 Instance Analysis

The contingency table (Table 1) is based on the features for each of the instances in the source corpus. These instances can be entered in another table (Table 2), with the rows corresponding to the corpus instances number (here listed only 10 of the 225 rows). Each row has three columns. The first column is the corpus instance identifier (Inst); the second column is the sense tag (S); the third column is the complement's part of speech (C); the fourth column is the governor's part of speech (G). This table is used as the

Need to refer to multiple CA and show what it can do here.

4 Dictionary Analysis

The PDEP senses (footnote 6) were provided from the *Oxford Dictionary of English* (ODE), Stevenson and Soanes (2003)). In addition, as described in Litkowski (2013), 7650 example sentences were also made available, from the Oxford English Corpus (OEC). These sentences were also parsed, generated with the same sets of features, characterizing the prepositions behavior. One notable aspect of this corpus is that the sentences are essentially simple, not compound sentences, thus likely having more accurate parses and features.

It is possible to compare independent tagging against the dictionary definitions for each of the senses. The PDEP data also include features for the sentences used to exemplify each of the senses. Table 3 shows the parts of speech for the complements of *above* in the OEC corpus. This table shows several variations that may occur. First, there is no sense for "10(n)"

Table 3: Parts of Speech for Complements of *above* in OEC Corpus

OEC	cd	dt	nn	nnp	nns	prp	vbg
1(1)	0	0	9	3	3	5	0
2(1a)	0	0	16	0	2	2	0
3(1b)	0	0	8	8	3	1	0
4(2)	0	0	12	0	2	2	0
5(2a)	0	1	8	2	1	8	0
6(2b)	0	0	7	0	4	0	0
7(2c)	0	1	12	0	5	0	1
8(2d)	0	0	20	0	0	0	0
9(3)	3	0	11	0	6	0	1

in the OEC corpus. This was added to the PDEP senses since the last sense occurred frequently in the CPA corpus. This sense is also an occurrence of a multiword expression (MWE) that should be considered along with the main senses of *above*, where MWEs are further discussed in section 8 below. Second, the parts of speech occurring in the OEC corpus are slightly different from those in the CPA corpus. Table 1 includes "jj" (adjective), "nnps" (plural capital nouns), "pdt" (predeterminers), and "wp" (*wh*-pronouns) that do not occur in Table 3. It is worth commenting that (1) these parts of speech are very infrequent, (2) they may not have been correctly parsed, and (3) these instances do not arise to a frequency that suggests they should have been incorporated into the set of dictionary senses.

Stevenson and Soanes (2003)

Here generate the simple CA graph and compare with Figures 1 and 2

5 Feature Selection

Reference to Glynn paper to discuss the issues of how to select which features to use in a CA.

Glynn (2014) and Krawczak and Glynn (2019)

6 Substitutable Prepositions

an examination of the **substitutable prepositions** field in PDEP, to allow the null hypothesis that the features across these substitutes are essentially highly similar

7 Supersenses

Note that this is also similar to the previous discussion on substitutable prepositions

allows similar examination of the PDEP field for **supersenses**, also allowing the examples used in the guidelines for supersenses in Schneider et al. (2017)

8 Multiword Expressions

About 70 senses were added to PDEP that were added to the sense inventories of the prepositions, based on their occurrence the CPA corpus. Many of these senses corresponded to multiword expressions (MWEs), with their own entries in the ODE dictionary (Stevenson and Soanes (2003)).

9 Reviewing the Corpora Tagging

10 References

References

Herve Abdi and Lynne J. Williams. Correspondence analysis. In Neil Selkind, editor, *Encyclopedia of Research Design.* Sage, Thousand Oaks, CA, 2010.

Dylan Glynn. Correspondence analysis: Exploring data and identifying patterns. In Dylan Glynn and Justyna Robinson, editors, *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*, pages 443–486. John Benjamins Publishing Company, 2014.

Michael Greenacre. *Correspondence analysis in practice, Third Edition.* CRC press, Boca Raton, FL, 2017.

Karolina Krawczak and Dylan Glynn. Operationalising construal: A corpus-based study in cognition and communication constructions. *Jezikoslovlje*, 20(1):1–30, 2019. URL https://hrcak.srce.hr/219568.

Ken Litkowski. The preposition project corpora. Technical Report 13-01, CL Research, Damascus, MD 20872 USA, april 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. Pattern dictionary of english prepositions. In Mona Diab, Aline Villavicencio, Marianna Apidianaki, Valia Kordoni, Anna Korhonen, Preslav Nakov, and Mark Stevenson, editors, *Essays in Lexical Semantics and Computational Lexicography - In Honor of Adam Kilgarriff*. Springer, 2016. URL http://www.clres.com/online-papers/LitkowskiPDEP.pdf.

Ken Litkowski. Honing the sketch engine prepositions. Technical Report 19-01, CL Research, Damascus, MD 20872 USA, 2019. URL http://www.clres.com/online-papers/HoneSkE.pdf.

Barbara McGillivray, Christer Johansson, and Daniel Apollon. Semantic structure from correspondence analysis. In *Coling 2008: Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing*, pages 49–52, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://www.aclweb.org/anthology/W08-2007.

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.

Stephen Tratz and Eduard Hovy. A fast, accurate, non-projective, semantically-enriched parser. In *Proceedings of the 2011 Conference on*

Empirical Methods in Natural Language Processing, pages 1257–1268, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D11-1116.

A Correspondence Analysis Techniques

Correspondence analysis has several varieties, grouped into two major types, simple and multiple. Simple CA examines contingency tables; multiple CA works with (See Greenacre (2017), Abdi and Williams (2010), and Glynn (2014).)

A.1 Simple Correspondence Analysis

The singular value decomposition (SVD) factors the standardized residual matrix into a diagonal matrix known as the singular values, explained the variance of the contingency table . The table has the original size of m x n. The diagonal matrix has n-1 singular values, in decreasing magnitude. The sum of the singular values is known as the total inertia, constituting 100 percent over the n-1 dimensions. A figure, such as in Figure 1, shows the percentage of each dimension that is covered. A biplot shows the results for two of those dimensions, identifying how much of the inertia (the total variance) is covered for each. In general, the hope is that the total of the first two dimensions will cover at least 80 percent of the variance.

After this basic statistic, CA allows considerable analysis. The first step creates a correspondence matrix (CM), dividing each cell by the sum of all the cells, so that the sum of the cells of the CM is equal to 1. This matrix is used to compute a matrix of standardized residuals, which is used to perform a singular value decomposition (SVD) into its factorizations. This permits a plot of the rows and columns of the original table, in this case shown in Figure 1. The figure visualizes how the senses and the parts of speech relate to one another. There are several packages for correspondence analysis, particularly in R, one in Python, and others in various statistical software. They can also perform the components in CA or can be implemented by developing the computations.

A.2 Supplementary Points

In a contingency table, as described in a A.1, the rows and columns establish the principal axes and the basis for the plots. These axes are viewed as

active. Each active has a different force of attraction - profiles farther from the average have more "leverage" in orienting the map. Sometimes, we wish to examine points that have no mass at all (i.e., their contribution to the inertia is zero). Such points are called *supplementary points* or *passive*. There are three common situations: an additional column, an additional row, or another row which is the sum of two rows. In these situations, the procedure is to add the supplementary rows or columns, as if they were to be analyzed as part of the contingency table, but then label them as supplementary. In the analysis, it is still possible to compute what inertia the supplementary points would have and we can show how these points relate to the original table, i.e., to determine the closest points of the original table. (See Greenacre (2017), pp. 89-96 and 263-264.)

A.3 Multiple Correspondence Analysis

Todo list