Principal Component Analysis of Verbs Introduced by Prepositions in Italian Adjunct Clauses

Martina Rizzello

January 16ᵗʰ 2023

1 Introduction

Although case is generally understood as applying to nouns, it has been documented to appear on verbs and at clause edges in languages outside of Europe, expressing relations between clauses (Aikhenvald, 2008; Franco, 2012; van Gijn, 2019). Most Romance languages are not traditionally analysed in terms of case; however, prepositions seem to perform the same role as affixal case marking observed in other languages, signalling the relation between arguments (Blake, 2001).

My research has been focusing on finding the relation between prepositions in Romance languages and the clause types they introduce (e.g. causal, purpose, result etc.). To complement this work, this report aims to examine whether the verbs introduced by different prepositions in adjunct clauses share some relevant characteristics. To do so, a Principal Component Analysis (PCA) of Italian adjunct clauses has been carried out. PCA is a powerful method used to reduce the dimensionality of large data sets while at the same time preserving as much information as possible. The aim is to determine if certain verbs are employed preferentially in conjunction with specific prepositions and if so, what semantic characteristics they share. The PCA has shown a possible correlation between prepositions introducing *before* clauses and verbs of motion and process, as well as between prepositions introducing causal relations and experiential verbs.

2 Methodology and Results

The data was sourced from the CORpus di Italiano Scritto (CORIS) (Rossini Favretti, 1998). Relevant examples were searched with the help of regular expressions and POS tags: these include prepositional phrases introducing adjunct clauses, followed either by a general verb form (i.e. all forms, inflected or not) or by an auxiliary and a past participle. Example 1 below shows one of the results of the query, with the addition of a gloss and a translation.

(1) possono essere bruschi tanto \<da   aver   causato\> alcuni disastri aerei
      can.3PL be   abrupt so.much PREP have.INF caused   some   disasters aerian
      'They can be so abrupt that they have caused plane crashes'

The examples were saved in different files, then cleaned and lemmatised. The result is a collection of 14 text files, one for each preposition, and each including a list of roughly 800 verbs. The distinction between the prepositions followed by an auxiliary plus a past participle and a general verb form was maintained, with the first being labelled as *_aux_pp* and the second as *_v*. The file names are composed of the prepositional phrase and the above mentioned labels.

The data was tokenised and turned into a frequency table. After calculating the relative frequency of each data point, a PCA was performed on the 70 most frequent verbs. To make the graph more legible and avoid overcrowding, the variables with the lowest contribution were filtered out. The resulting biplot of individuals and variables is shown in Fig. 1.
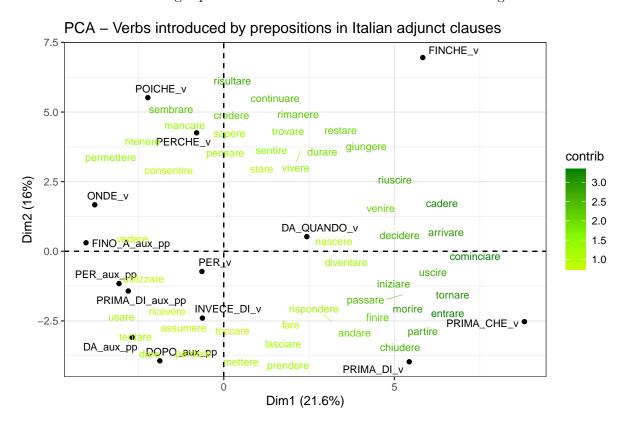

Figure 1: PCA of verbs introduced by prepositions in Italian adjunct clauses.

The first two principal components appear to explain respectively only 21.6% and 16% of the variance, suggesting that the analysis is not particularly reliable. This might be due to the limited amount of data, or it could simply indicate there is no strong correlation between the prepositions and the verbs that follow them. Some patterns are still noticeable, although they should be interpreted carefully. I will rely on Cook (1979) for the semantic classification of verbs.

All the prepositions followed by an auxiliary and a past participle seem to cluster together, suggesting that the verb form itself might be an important factor. The verbs that gravitate around these prepositions – 'use', 'give', 'receive, 'touch', 'lose' – mostly appear to belong to the category of action and action-process verbs. It is possible that these types of verbs are more likely to be used in passive infinitive constructions or in their past infinitive form, as a completed action. Note that the prepositions in the bottom-left quadrant are used in *before* and *after* clauses, in purpose and in result clauses.

Moving to the bottom-right quadrant, the verbs close to 'before' (*prima che* and *prima di*) followed by general verb forms belong to the category of motion ('go out', 'go through', 'come back', 'enter', 'go') and process ('leave', 'start', 'end', 'die'). In the upper-right quadrant the verbs close to 'since' (*da quando*) – namely 'arrive', 'come', 'fall', 'decide', 'accomplish' – also belong to the category of motion and process verbs.

Finché ('until') appears as an outlier at the upper-right edge of the graph, suggesting that it is distinct from the other prepositions while also not being particularly related to any of the

verbs shown. It is possible that the verbs more closely related to it were filtered out due to their low percentage of contribution to the PCA.

In the upper-left quadrant, *poiché* and *perché* – prepositions followed by a complementiser introducing causal (i.e. *because*) relations – appear to be mostly surrounded by verbs that Cook classifies as experiential, that is, related to human cognition and feelings. Some of these verbs are 'seem', 'think', 'believe', 'know', 'miss' and 'consider'. In this case the predictor for the type of verb employed is clearly the clause type and not the preposition itself, as other instances of *per* appear elsewhere in the graph.

3 Conclusion

Overall, despite the low percentage of variance, some patterns have surfaced. Regardless of preposition and clause type, all the prepositions introducing an auxiliary followed by a past participle appear in the same area and seem to correlate with action and action-process verbs. Moreover, *before* clauses including general verb forms appear close to verbs of motion and process. Finally, prepositional phrases encoding causal relations seem to correlate with experiential verbs. Further research is needed to explain and confirm the above mentioned patterns, as well as to possibly reproduce the experiment with bigger data sets.

4 Supplementary Materials

The data presented, the code employed and all other resources associated with this report have been archived and are available at the following link:

- `https://doi.org/10.5281/zenodo.7541464`.

References

Aikhenvald, A. Y. (2008). Versatile cases. *Journal of Linguistics*, 44(3):565–603. Publisher: Cambridge University Press.

Blake, B. J. (2001). *Case.* Cambridge University Press.

Cook, W. A. (1979). *Case grammar: development of the matrix model (1970-1978).* Georgetown University Press.

Franco, L. (2012). On Case and clauses: Subordination and the Spell-Out of non-terminals. In *Proceedings of the First Central European Conference in Linguistics for postgraduate students*, volume 82, page 103. Pázmány Péter Catholic University Budapest.

Rossini Favretti, R. (1998). Corpus di italiano scritto. `https://corpora.ficlit.unibo.it/TCORIS/`. Accessed: 2023-01-06.

van Gijn, R. (2019). Case markers as subordinators in South American indigenous languages. *Nominalization in Languages of the Americas. Amsterdam: John Benjamins*, pages 197–247.