The Meaning of *forma* in Thomas Aquinas. Hierarchical Clustering from the *Index Thomisticus* Treebank

Marco Passarotti[1], Gabriele Cantaluppi[1], Francesco Mambrini[2]
[1]Università Cattolica del Sacro Cuore, Milan (Italy); [2]University of Cologne (Germany)
marco.passarotti@unicatt.it; gabriele.cantaluppi@unicatt.it; f.mambrini@gmail.com

Background and Motivation

Started in 1949 by father Roberto Busa (1913-2011), the *Index Thomisticus* (IT; Busa, 1974-1980) represents the first digital corpus of Latin and has been a groundbreaking project in computational linguistics and literary computing. The IT contains the opera omnia of Thomas Aquinas (118 texts) as well as 61 texts by other authors related to Thomas, for a total of around 11 million tokens. The corpus is morphologically tagged and lemmatized and it is available on paper, CD-ROM and on-line (www.corpusthomisticum.org).

The *Index Thomisticus* Treebank (IT-TB: http://itreebank.marginalia.it) is an ongoing project that aims at performing the syntactic annotation of the whole IT corpus. The IT-TB is a dependency-based treebank consisting of around 150,000 annotated tokens for a total of approximately 8,000 sentences from three works of Thomas: *Scriptum super Sententiis Magistri Petri Lombardi, Summa contra Gentiles* and *Summa Theologiae*. The IT-TB shares the same annotation guidelines with the Latin Dependency Treebank (LDT), developed by the Perseus Digital Library (Boston, MA) on texts of the Classical era. These guidelines resemble those of the Prague Dependency Treebank of Czech (PDT) and are similar to those of the PROIEL corpus of the New Testament in several translations in Indo-european languages (Oslo, Norway).

The IT-TB is part of a bigger project named "Lessico Tomistico Biculturale" (LTB). LTB aims at building a new lexicon of Thomas Aquinas by empirical confrontation with the evidence provided by the IT. Indeed, the entries of the available lexica of Thomas are systematically biased by the criteria for the selection of the examples adopted to describe the different meanings of lemmas. This limitation can now be overcome by exploiting the data of the IT and of the IT-TB.

The first lemma we want to analyse for the purposes of LTB is *forma*. This lemma has 18,357 occurrences in the IT corpus, 16,525 of which in Thomas' works and 1,832 in the texts of other authors. Thus, we devoted the first years of the project to annotate those sentences that feature at least one occurrence of *forma*. Presently, 5,191 occurrences of *forma* have been annotated in the IT-TB, corresponding to around one third of all the occurrences of *forma* in Thomas' works.

Forma is a 'technical' word in Thomas' writings, showing high polysemy. In the lexicon of Thomas Aquinas by Deferrari & Barry (1948-1949: 433-438), *forma* has 5 meanings: (a) "form, shape", synonym of *figura*; (b) "form", the configuration of an artificial thing as distinct from "figure" (which is the configuration of natural things); (c) "form", the actualizing principle that makes a thing to be what it is; (d) "mode, manner"; (e) "formula". In Multiwordnet, *forma* has 21 senses, which do not include all those present in Thomas.

Contribution

We apply word hierarchical clustering techniques to cluster the occurrences of *forma* in both IT and IT-TB into groups, so that occurrences showing similar behaviour fall in the same cluster(s). Different word clustering techniques usually follow a two-step procedure: (1) classification: each occurrence is represented as an observation in a matrix (a feature vector) and the similarity of two observations is computed; (2) clustering: some clustering algorithm is applied, such that similar occurrences are grouped together.

Our theoretical starting point is the notion of "context of situation" by Firth, who points up the context-dependent nature of meaning, as reported in his famous quotation: "You shall know a word by the company it keeps" (Firth, 1957: 11).

We carried out a DIvisive hierarchical clustering ANAlysis (Kaufman & Rousseeuw, 1990) by using the function DIANA (Maechler *et al.*, 2011), available in the library cluster of the R free statistical software (http://www.r-project.org) starting from a dissimilarity matrix generated by considering a modification of the *simple matching* distance (Sokal & Michener, 1958). The *simple matching* distance between two observations, r and s, with categories (x_{r1}, x_{r2}, …, x_{rk}) and (x_{s1}, x_{s2}, …, x_{sk}) over k variables is:

$$dist(r,s) = \frac{k - \sum_{j=1}^{k} sim(x_{rj}, x_{sj})}{k}, \qquad sim(x_{rj}, x_{sj}) = \begin{cases} 0 \ \ if \ x_{rj} \neq x_{sj} \\ 1 \ \ if \ x_{rj} = x_{sj} \end{cases}.$$

Since, for every pair of observations, we considered the similarity in groups of variables, e.g. for the 2 words preceding and following the occurrence of *forma* (and not for each variable as specified by the *simple matching* distance, e.g. the first word following the occurrence of *forma*) we defined:

$$dist(r,s) = \frac{sim_{\max} - \min\left(\sum_{g=1}^{w} sim(x_{rg}, x_{sg}), \sum_{g=1}^{w} sim(x_{sg}, x_{rg})\right)}{sim_{\max}}$$

where w is the number of groupings of variables; $sim(x_{rg}, x_{sg})$ is an asymmetric measure for the number of elements in the s observation matching with the elements in the r observation for group g (namely multiple occurrences of the same term may occur in a group); and

$$s_{\max} = \max_{r,s}\left(\min\left(\sum_{g=1}^{w} sim(x_{rg}, x_{sg}), \sum_{g=1}^{w} sim(x_{sg}, x_{rg})\right)\right)$$

is the overall observed maximum number of matches.

We produced two matrices of data, one from the IT-TB and one from the IT:

- a matrix consisting of 5,191 observations. For each occurrence of *forma* in the IT-TB, the observations report the lemmas of: (a) its parent and grandparent in the dependency tree, (b) all its attributives (dependent nodes with syntactic label "Atr" in the tree), (c) all its coordinated nodes in the tree, (d) up to 2 words preceding and 2 words following the occurrence of *forma* concerned in the observation.

  While (a), (b) and (c) report information extracted from the IT-TB (i.e. syntactic information), (d) features information concerning the linear word order of the text (taken from the IT);
- a matrix consisting of 18,357 observations. For each occurrence of *forma* in the IT, the observations report the lemmas of up to 3 words preceding and 3 words following the occurrence of *forma* concerned.

We performed several experiments on these matrices, by choosing different settings for grouping the variables. For instance, in a number of experiments we excluded specific kinds of words (like function words, pronouns and some verbs) when computing the similarity/dissimilarity of the observations. In other experiments, we chose different settings for grouping the values of the columns for each observation.

The same experiments were always performed on both all the 18,357 observations provided by the IT-based matrix and on a subset consisting of the 16,525 observations of *forma* in Thomas' works only.

Evaluation and Results

In order to evaluate the results, we built two different gold standards.

- Gold standard A (GsA): we manually annotated the meaning of 672 randomly chosen occurrences of *forma* (approx. 13% of the total in the IT-TB). We used a tagset of 10 different values that were defined according to Deferrari & Barry, Latin Wordnet and lexico-syntactic information from the IT-TB;
- Gold standard B (GsB): among the observations of GsA, we selected a subset of 356 featuring a clear (i.e. not ambiguous) meaning of *forma*.

When plotting the results of the hierarchical clustering, we highlight in specific colours the observations respectively recorded in GsA and GsB. This allows us to visually check if those observations annotated with the same meaning in the gold standards do indeed appear or not into similar clusters in the plot.

For each experiment, we evaluate the results by using several evaluation metrics, among which are precision, recall and f-score. The best performing setting on 5,191 observations is the following:

- function words, pronouns and verb *sum* excluded from computing similarity/dissimilarity;
- grouping setting: 2 separate groups, namely syntactic information and textual information.

With this setting, we reached the best f-score of 0.93 (precision 0.95; recall 0.90), for the GsB observations tagged with label 6 (*forma materialis*; *forma* connected with *materia*). The GsB observations tagged with other labels show lower scores (ranging from 0.86 to 0.5). If GsA observations are concerned, the f-score ranges from 0.8 (precision: 0.93; recall: 0.7) for label 6, to 0.28 for label 2 (*forma* as *anima corporis*).

References

BUSA, R. (1974-1980). *Index Thomisticus*. Stuttgart-Bad Cannstatt: Frommann-Holzboog.

DEFERRARI, RJ. & BARRY, M.I. (1948-1949). *A Lexicon of St. Thomas Aquinas: based on the* Summa Theologica *and selected passages of his other works*. Washington: Catholic University of America Press.

FIRTH, J.R. (1957). *Papers in Linguistics 1934-1951*. London: London University Press.

KAUFMAN, L. & ROUSSEEUW, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

MAECHLER, M., ROUSSEEUW, P.J., STRUYF, A., HUBERT, M. & HORNIK, K. (2011). *cluster: Cluster Analysis Basics and Extensions*. R package version 1.14.1. http://CRAN.R-project.org/package=cluster.

SOKAL, R.R. & MICHENER, C.D. (1958). "A statistical method for evaluating systematic relationships". *Univ. Kansas Sci. Bull.*, 38, 1409-1438.