Abstract

“QuoVadis” is a corpus of entities and semantic relations between them, built on the “Quo Vadis” book authored by Henryk Sienkiewicz. The process of producing the corpus is briefly presented in the paper. Then, levels of annotation, partly automatic, partly manual, are summarised and basic statistics over the corpus are reported. The corpus is freely offered to the scientific public interested to study anaphora resolution and entity linking on the Romanian language.

Key words — anaphora, corpus, entity linking, manual annotation, Romanian resources.

[1] Introduction

A book is a world of its own, a creation of its author. Nonetheless, it is re-created each time a reader reads it, discovering its action, characters, plot and facts, while also imposing on them a personal interpretation. As language users, readers make easily and subconsciously connections between characters, recognize the referential correspondences of each character that build up linguistic chains, have intuitive predictions about the evolution of the plot and of the acting characters. Thus, above a basic level of primary information conveyed by the text, an inner world of understanding and linking with the general experience of the reader is elaborated.

In general, there should be a common, basic level of understanding of a text, which should reside in the decoding of the primary messages it conveys, by means of deciphering the mentions of characters and the relations linking them. These relations can be of a referential (anaphoric) and a non-referential nature. Referential relations involve identification of multiple mentions that refer the same character or related properties. For example, reading a Romanian version of “Quo Vadis”, the book of the Nobel prized laureate Henryk Sienkiewicz, a reader would be able to discover that Vinicius, el (he), Marcus, tânărul patrician (the young patrician), ruda lui Petronius (Petronius’ relative), un tribun militar (a military tribune) etc. all refer only one character of the book, whose name is Marcus Vinicius.

Of a non-referential nature are other types of relations holding between the characters of the novel. For instance, the affective relationships between Lygia and Vinicius, that are contradictory at the beginning, develop onto a beautiful love-story. Other examples are the dominance relations between the emperor Nero and his courtiers, or the kinship relation between Lygia and his adoptive parents, Aulus and Pomponia Graecina.
In this paper we describe a corpus encoding mentions of persons, gods, groups of persons and gods, and body parts of persons and gods, as well as semantic relations linking these entity types, classified in 4 categories: referential, affective, kinship and social. The corpus uses as a hub document a Romanian version of the already mentioned “Quo Vadis” text. The selection of this particular book should be attributed to reasons that include: the density of characters and relations, freeness of copyright and, not the least, the fact that the novel being translated in so many languages, we envisaged the possibility to exploit the semantic annotations in the Romanian version for other languages, by applying exporting techniques.

The motivations for creating this corpus were two-fold. First, among the Romanian textual resources, a gold corpus that could be used for training programs to reproduce human expertise in the recognition of entities and their correlations, including anaphoric and semantic links is still lacking. Second, the activity itself to build the corpus was organised as a complex annotation exercise in the benefit of our master students in Computational Linguistics. The experience gained during the process of annotation and the organization of this extremely elaborate and time-consuming task helped us to acquire a level of know-how that, we believe, can be exploited in future large-groups textual annotations tasks.

[2] Studies in anaphora and semantic links on Romanian

Huang (2000) defines anaphora as “a relation between two linguistics elements, wherein the interpretation of one (called an anaphor) is in some way determined by the interpretation of the other (called an antecedent)”. A referential expression needs a source for its saturation. The notion of anaphora should be considered a contextual one, because many referential expressions are, only by themselves, therefore without a context, ambiguous with respect to what can be considered an antecedent. This is why a corpus, and not a collection of anaphor-antecedent pairs, is the proper resource to train a recognition program.

We have found relatively few studies about anaphora concerning the Romanian language. In the following we list a number of studies concentrating on Romanian anaphora from a linguistic formal level: some contrastive approaches between Romanian and French are discussed by Tasmowsky (1990); types of replays anaphora are investigated by Iliescu (1988); approaches to discursive anaphora are inventoried by Manoliu-Manea (1993); Pană Dindelegan (1994) brings into discussion the neutral value of the pronoun o (functioning as pro-form, when an antecedent is referred, vs. non-substitute, when it is not bound with the nominal substitute) and the functions of the clitics in Romanian; a thorough analysis of syntactic anaphora is elaborated by Dobrovie-Sorin (1994/2000); means of anaphoric realisations and expression and a typology of anaphora is reported by Zafiu (2004); in Gramatica limbii române (the Grammar of the Romanian Language), the compendium realised by the Romanian Academy, anaphoric and cataphoric phenomena in Romanian are inventoried in the section dedicated to the Sentence (2005/2008, 2nd volume); finally, a discursive approach to anaphora and cataphora can be found in (Oroian, 2006). However, none of these studies use a corpus as an organised repository of examples.

All computational approaches on Romanian anaphora, instead, place corpora, even if of small dimensions, at the base of the tools built and the reported evaluations. We mention here some of these

\[\text{Translation by Remus Luca and Elena Lință and published at Tenzi Publishing House in 1991.}\]
There are extremely many studies presenting corpora containing annotations of entities and semantic links, in other languages then Romanian, many of them for English: the MUC (The Message Understanding Conference) and the ACE (Automatic Content Extraction) corpora (Doddington et al., 2004), or ARRAU (Poesio & Artstein, 2008). For other languages: AnCora corpus – for Catalan and Spanish (Recasens, 2011), DAD – for Italian (Navarretta & Olsen, 2008), COREA – for Dutch (Hendrickx et al., 2008), etc.

Our work, we believe, could be of real support to both theoreticians and experimentalists. As limited as it is now, (one genre, one author, and an exact temporal placement), the QuoVadis corpus could provide evidences for their hypothesis, could be a source of positive and counter examples, or could be used to trains programs designed to recognise entities and relationships among them.

[3] Building the corpus

The process of developing the corpus was an elaborate and time consuming activity. A research theme to annotate the “Quo Vadis” novel with semantic relations was discussed with the first year master students in Computational Linguistics at the Faculty of Computer Science of the “Alexandru Ioan Cuza” University of Iași in the autumn-winter university term of 2012. Since then, all along 4 terms, annotation conventions were presented and refined in lab sessions, special annotation cases were analysed, students having to work by themselves or grouped by two. More versions were produced and improved iteratively, while, to the end, the process being left only in the hands of the most talented annotators. The whole activity is thoroughly described in (Cristea et al., 2014). The experience gained within the group of students was used by the first author to produce an entirely new version of the corpus, described in (Bibiri, 2014). The statistics presented in this paper reflect this final version of the QuoVadis corpus.

Before manual annotation, the book, in cleaned text format, was submitted to a chain of pre-processing steps, by accessing the web services of the NLP-Group@UAIC-FII², see also (Simionescu, 2012): the text was first segmented at sentence boundaries, then tokenised, then POS-tagged and lemmatised, and finally a chunker marked noun phrases (NPs) and their heads. Marking noun phrases was necessary, as referential expressions (the textual realisations of entities) have been selected by the annotators among the NP elements already marked. In the identification of entities, heads are important clues, because annotated recursive NPs should have distinct heads and when more recursive NPs have the same head, only the longest is considered to realise an entity. For instance, in the following recursive referential expressions, heads are underlined: [oameni din [toate stările sociale]] ([people of [every position]]). In cases when the entity does not have a lexical realisation (zero anaphora), as in contexts where the subject is not expressed in the stretch of the text, part of the morphological and syntactic features of the included entity are recognisable in the person and number of the verb and the annotation reflects this situation. An example of a null entity notation follows: [te], [iubesc; REALISATION=INCLUDED].

² http://nlptools.infoiasi.ro/
Marcus; here the subject of *iubesc* (*love – v.*) is included in the predicative form of the verb, unlike in the English version, where the subject is always expressed: *[I]₂ *love* *[thee]*₁, Marcus.

As for relations, they always hold between two arguments and, with the exception of referential relations, they are signalled by a word or an expression (the trigger). An important concern in the creation of the corpus was to mark as span of a non-referential relation the minimal stretch of text in which the relation is expressed. This concern is applied with an eye open for the future, thinking at systems presumed to generalise patterns from relations instances, drilling thus a recognition process. It is clear that the shorter the stretches of text that contain relations are, the higher the probability to infer correct patterns out of the annotated examples. The length of a relation is the minimal linear span that includes the two arguments and the trigger. Excepting for referential relations, where arguments can be sometimes quite distant in text one to the other, usually relations are expressed locally, within a sentence, a clause, or even a noun phrase. Directionality of referential relations is always marked from the right argument to the left one, even if not necessarily the closest linearly in the case of coreferentiality.

Example:

*Dacă* [Nero]₁, ar fi poruncit să fii răpită pentru [el]₁, nu te-ar fi adus la Palatin.

*If* [Nero]₁ *had given command to take thee away for* [himself]₁, *he would not have brought thee to the* Palatine.

where the identical indexes show a coref relation.

For the other types of relations, if arguments are non-intersecting, the normal reading of the trigger gives the direction, as in this example of a social type of relation:

<Eliberând>-[o]₁, [Nero]₂…

[Nero]₂, when he had <freed> [her]₁…,

where the relation holds as: [2] superior-of [1], on the ground that only a superior can free someone, and the trigger (marked between angle parenthesis) is *Eliberând* (*freed*) In case of nested arguments the direction is, by convention, from the external argument towards the inner one, as in this example of a kinship type of relation:

[<sora> [lui]]₂₁

[[his]₂ <sister>].₁

Here the relation is marked as [1] sibling-of [2] and the trigger is the word *sora* (*sister*).

For the manual annotation, the PALinkA tool³ was used (Orăsan, 2003). Six XML types of elements record all manual annotations: ENTITY – making entities, TRIGGER – marking relations’ triggers, and REFERENTIAL, AFFECTIVE, KINSHIP and SOCIAL – marking the eponymous relations. In all, we annotated 9 subtypes of referential relations, 7 subtypes of kinship, 11 subtypes of affective and 6 subtypes of social relations, described in (Cristea et al., 2014).

Following is a more complex example:

³ PALinkA (accessible at http://elg.wlv.ac.uk/projects/PALinkA/) was created by Constantin Orăsan in the Research Group in Computational Linguistics, at the School of Law, Social Sciences and Communications, Wolverhampton.
... cui i-ar fi putut trece prin minte că [un patrician], [nepot şi [fiu de [consul]]], ar putea să se găsească printre groperi.

Besides, into whose head could it enter that [a patrician], [the grandson [of one consul]], [the son [of another]], could be found among servants, corpse-bearers.


[4] Statistics and discussions over the corpus

The figures in Table 1 present the corpus at a glance.

<table>
<thead>
<tr>
<th>Table 1: General statistics over the corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>#sentences</td>
</tr>
<tr>
<td>#tokens, punctuation included</td>
</tr>
<tr>
<td>#tokens summed up under all relations</td>
</tr>
<tr>
<td>#entity mentions</td>
</tr>
<tr>
<td>#referential relations</td>
</tr>
<tr>
<td>#AKS relations (Affective + Kinship + Social)</td>
</tr>
<tr>
<td>#triggers</td>
</tr>
</tbody>
</table>

Figure 1 shows the histogram of lengths of non-referential relations. The good news is that most of the relations have very short length spans. We might even risk to conjecture that the accuracy of a relations recognition program, after being trained on this corpus, will very much fall over this curve.

In the annotation of entities and referential links we were interested to re-compute referential chains. A referential (anaphoric) chain is made up of the complete set of referential expressions coreferring the same entity, ordered in the linear unfolding of the text. To overcome the lack of inter-annotator agreement tests, a set of software filters were designed and run on the XML file, each error triggering new correction phases. The final goal was to obtain a one-to-one mapping between the set of referential chains and that of entities of the book, being them singular or collective. Thus, a correct chain should include all and only the mentions of one character of the book.

Figure 1: Relations occurrences (y-axis) in correlation with the relations’ span length, in number of words (x-axis)
The corpus allows for quantitative and qualitative insights over the characters that are part of the novel. For instance, Figure 2 shows how many times the top 8 characters are referred in the text and in how many relations they occur (this diagram reflects the importance of the characters in the novel, the main ones being Vinicius and Lygia, followed by Nero, Petronius, Chilon, Ursus, Christos and Apostle Peter).

Semantic graphs coding affective and social interactions of characters as well as kinship relations can also be drawn. A semantic graph grouping two affective relations are displayed in Figure 3: the love relation between Vinicius and Lygia and the other characters of the novel, and the worship relationship versus Christos from his adepts. Figure 4 shows two affective relations: this graph displays the feeling of fear that Nero’ loyalists and obedients have against him, and the feeling of hate arising between characters as conveyed by the plot.

Finally, we show in Figures 5 and 6 the complex interactions of the main character of the novel: the Roman patrician Marcus Vinicius. Figure 5 reflects the AKS relations between him and other characters of the book. The thickness of arrows suggests the frequency of mentions of relations. As seen in Figure 6, subordination and love are the dominant relations in the book. Simplifying a lot, they show at a glance what the novel is all about: a love story in a time of predominant social subordination.

Figure 2: Occurrences of the top 8 characters in the novel

---

4 Graphs were realised with the NodeXL open-source template for Microsoft® Excel® that automatically generates graphical representations for network edge lists stored in worksheets.
Figure 3: Affective relations love and worship in the corpus
Figure 4: Affective relations fear and hate in the corpus
We presented in this paper “QuoVadis”, a corpus of entities and semantic links. Lack of financial resources did not allow redundant annotation and calculation of inter-annotator agreement. This aspect, of a tremendous importance if we aim at achieving a high enough accuracy to qualify it for a ‘gold’ corpus, is planned to be solved in the near future, by organising sample-based evaluations. For the time being, to raise the quality of annotation we have made use of software filters, described in (Cristea et al., 2014). For instance, it is highly improbable that a coreferential link contains only pronouns, and it is highly probable that all common nouns in a chain have the same number and gender morphological values. Errors found in the listings generated by these filters were manually corrected during repeated correction phases. Moreover, a special interface was designed to visualise the coreference links.

In books, relations are either stable (for instance, if not contested, those describing family links) or may evolve or even completely change, as the story unfolds. To give some examples, the sexually motivated interest that Vinicus shows to Lygia and the lack of interest or even disgust that she has for him evolve both into love; the friendship of Petronius versus Nero depreciates in hate; and Vinicus’ lack of understanding versus Christ develops into worship. However, we do not record in this variant of the corpus time frames, so these dynamics are impossible to be caught now.

The following work will concentrate on more directions: first to certify the accuracy of the corpus by organising inter-annotator agreements tests, then to train programs to recognise entities’ mentions, to test an anaphora resolution platform for the Romanian language (Cristea et al., 2002a) and to improve it, then to recognise semantic relations belonging to the classes referential, affective, kinship and social, and, finally, even to try experiments of semantic inferences that would exploit combinations of relations (for instance, the difference between the paternal love of Lygia versus her adoptive parents and the one she develops versus Vinicius).

http://nlptools.infoiasi.ro/QuoVadisVisualization/
The corpus is freely available for research at http://nlptools.infoiasi.ro/Resources.jsp and will also be included in COROLA, the Computational Representational Corpus of Contemporary Romanian Language.

Acknowledgements

Part of the work described in this paper was done in relation with COROLA—The Computational Representational Corpus of Contemporary Romanian Language, a joint project of the Institute for Computer Science in Iași and the Research Institute for Artificial Intelligence in Bucharest, under the auspices of the Romanian Academy.

The annotation conventions used in the corpus represent largely work done in preparation of the project “MappingBooks – Let me jump in the book!”, financed within the PARTENERSHIP programme of the 2013 Competition (PCCA 2013), in a joint consortium made up of the “Alexandru Ioan Cuza” University of Iași, SC SIVECO Romania SA and „Ștefan cel Mare” University of Suceava.

We thank to our master students in Computational Linguistics from the “Alexandru Ioan Cuza” University of Iași, Faculty of Computer Science, who along the university years 2012-2014 have annotated and then corrected the first version of the “Quo Vadis” corpus.

References


