Quo Vadis: A Corpus of Entities and Relations

Dan Cristea[1,2], Daniela Gîfu[1], Mihaela Colhon[3], Paul Diac[1], Anca-Diana Bibiri[4], Cătălina Mărănduc[5], and Liviu-Andrei Scutelnicu[1,2]

[1] Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi
[2] Institute for Computer Science Romanian Academy - the Iaşi branch
[3] Department of Computer Science, University of Craiova
[4] Department of Interdisciplinary Research in Social-Human Sciences, "Alexandru Ioan Cuza" University of Iaşi
[5] "Iorgu Iordan-Al. Rosetti" Institute of Linguistics of the Romanian Academy
{dcristea,daniela.gifu}@info.uaic.ro,
mcolhon@inf.ucv.ro,paul.diac@info.uaic.ro,
anca.bibiri@info.uaic.ro,catalina_maranduc@yahoo.com,
liviu.scutelnicu@info.uaic.ro

Abstract. This chapter describes a collective work aimed to build a corpus including annotations of semantic relations on a text belonging to the belletristic genre. The paper presents conventions of annotations for 4 categories of semantic relations and the process of building the corpus as a collaborative work. Part of the annotation is done automatically, such as the token/part of speech/lemma layer, and is performed during a preprocessing phase. Then, an entity layer (where entities of type person are marked) and a relation layer (evidencing binary relations between entities) are added manually by a team of trained annotators, the result being a heavily annotated file. A number of methods to obtain accuracy are detailed. Finally, some statistics over the corpus are drawn. The language under investigation is Romanian, but the proposed annotation conventions and methodological hints are applicable to any language and text genre.

Key words: semantic relations, annotated corpus, anaphora, XML, annotation conventions

1 Introduction

When we read books we are able to discover in the sequence of words, with apparently zero effort, the mentions of entities, the relationships that connect entities, as well as the events and situations the characters are involved in. Entities, relations, events and situations can be of many types. For instance, entities could be: persons, animals, places, organisations, crafts, objects, ideas, etc. as well as any grouping of the above. Relations linking characters could be anaphoric (when the interpretation of one mention is dependent on the interpretation of a previous one), affectional (when a certain feeling or emotion, is expressed in

the interaction of characters), kinship (when family relationships are mentioned, sometimes composing very complex genealogical trees), social (when job hierarchies or social mutual ranks are explicitly remarked), etc. Moreover, the text could include mentions about relations holding between characters and other types of entities: persons are in places, persons belong to organisations, mutual positioning of locations in space, etc. Deciphering different types of links is a major step in understanding a book content.

We address in this paper the issue of building a corpus that makes explicit a strictly delimited set of binary relations between entities that belong to the following types: persons, gods, groups of persons and gods, parts of bodies of persons and gods. The relations marked in the corpus belong to four binary types: anaphoric, affectional, kinship and social.

Very often the interpretation of semantic relations is subjective being the result of a personal interpretation of the text, more precisely, of inferences developed by the reader or obtained by putting on stage some extra textual general knowledge. In building this corpus we avoided anything that is not explicitly uttered in the text, trying thus to keep the subjective interpretation to a minimum. Also, we were not interested to decode time moments, nor the different ways in which time could be connected to entities or to the events they participate in.

The motivation for this endeavour is to base on this "gold" corpus the construction of a technology able to recognise these types of entities and these types of semantic relations in free texts. The experience acquired while building such a technology could then be applied in extrapolating it to other types of entities and semantic relations, finally arriving to a technology able to decipher the semantic content of texts. When a human reads a text, she/he first deciphers the semantic links as they are mentioned in the text, and only then these isolated pieces of knowledge are connected by inferences, that often engage particular knowledge of the reader, thus obtaining a coherent, albeit personal, interpretation of the text, as a whole. A novel could be the source of as many interpretations as readers it has. The freedom to build a proper interpretation makes the relish of most readers. The issue of building a machine interpretation of a novel should therefore be considered also from this perspective (which could be a rich source of philosophical questions, too): *how much do we want our technology be able to add above the strict level of information communicated by the text?* We believe that the answer to this question shall be rooted in the actual type of application that is desired, but the primary challenge is this one: *are we capable to interpreted basic assertions expressed in a book?*

We are tempted to conclude this introduction by making a parallel with the way human beings do grow up, the accumulation of their memories and readings assembling in time their personalities, characters, predisposition for certain types of behaviour. Similarly, one may imagine the formation of intelligent behaviours in agents, which could be rooted in the artificial interpretation of books, this way short-circuiting a real life-time formation. Fictional worlds, if proper selected and fed into agents, are extremely rich in life examples, and a plurality of semantic

aspects belonging to real or fictional lives could be studied if recognized in other contexts.

The paper is organised as follows. In the following section we give reasons for selecting the text of a novel as the basis for this corpus. Then we make a brief tour in the literature to configure the state-of-the-art in work related to semantic relations. Section 4 presents the annotations conventions for entities and the four categories of relations annotated in the corpus. Then Section 5 details the activities for the creation of the corpus. The notations in the corpus allowed to make different counts and comparisons. They are presented in Section 6. Finally, Section 7 makes concluding remarks and presents ways of exploitations of the corpus.

2 Why a corpus displaying semantic relations in free texts?

To understand a language one needs not only means of expression, but also vehicles of thought, necessary to discover new ideas or clarify existing ones by refining, expanding, illustrating more or less well specified thoughts [50]. Semantic relations describe interactions, connections. Connections are indispensable for the interpretation of texts. Without them, we would not be able to express any continuous thought, and we could only list a succession of images and ideas isolated from each other. Every non-trivial text describes a group of entities and the ways in which they interact or interrelate. Identifying these entities and the relations between them is a fundamental step in text understanding [32]. Moreover, relations form the basis for lexical organization. In order to create lexical resources, NLP researchers proposed a variety of methods, including lexico-syntactic patterns and knowledge-based methods exploiting complex lexical resources. Somehow similar situations occur when we reverse the task and go from text interpretation to text production. As Michael Zock and collaborators have shown [51], [52] lexical resources are only truly useful if the words they contain are easily accessible. To allow for this, association-based indexes can be used to support interactive lexical access for language producers. As such, co-occurrences can encode word associations and, by this, links holding between them.

Developing a lexical-semantic knowledge base to be used in Natural Language Processing (NLP) applications is the main goal of the research described in this chapter. Such resources are not available for many languages, mainly because of the high cost of their construction. The knowledge base is built in such a way as to facilitate the training of programs aiming to automatically recognise in text entities and semantic relations. As we will see in the next sections, it includes annotations for the spans of text that display mentions of entities and relations, for the arguments (poles) of the relations, as well as for the relevant words or expressions that signal relations.

We are not concerned in the research described in this paper neither with the automatic recognition of entities, nor with the recognition of the relationships

between entities. Instead we present how a significant corpus marking entities and relations has been built. The recognition problem will be our following objective. Năstase et al. [32] show many examples of NLP tasks that are based on the ability to identify semantic relations in text, such as: information extraction, information retrieval, text summarization, machine translation, question answering, paraphrasing, recognition of textual entailment, construction of thesauri and of semantic networks, word-sense disambiguation, language modelling. Although the issue of automatic semantic relations recognition has received good attention in the literature, we believe that the problem is far from being exhausted. Moreover, most of the research in this area is focused on corpora composed of press articles or on Wikipedia texts, while the corpus we describe is built on the skeleton of a fictional text.

As a support for our annotations we have chosen a novel, i.e. a species of the epic genre, which is particularly rich in semantic relations. The text used was the Romanian version of the novel "Quo Vadis", authored by the Nobel laureate Henryk Sienkiewicz[6]. In this masterpiece the author narrates an extremely complex society, the pre-Christian Rome from the time of the emperor Nero. The text displays inter-human relations of a considerable extent: love and hate, friendship and enemy, socio-hierarchical relationships involving slaves and their masters, or curtains and the emperor, etc. The story is dynamic, it involves many characters and the relations are either stable (for instance, those describing family links) or change in time (sexual interest, in one case, and disgust, in another, both evolve into love, friendship depreciates in hate, lack of understanding develops into worship). Appreciative affective relations may differ depending on whether they hold between human characters, when it takes the form of love, or between humans and gods, when it shapes as worship. Another aspect of interpretation relates the contrast between the social and affective relationships, dictatorial, based on fear and abeyance, at the court of Nero, and those found in the evolving but poor Christian society, based on solidarity and forgiveness.

Annotating a novel to entities and semantic relations between them represents a significant challenge, given the differences between journalistic and literary style, the first being devoid of figures of speech that create suggestive images and emotional effects into the mind of the reader.

Not the least, the novel "Quo Vadis" is translated in extremely many languages. We have thought at the possibility to exploit the semantic annotations of the Romanian version for other languages, by applying exporting techniques. Rather many results in the last years have shown that, in certain conditions, annotations can be exported on parallel, word-aligned, corpora, and this usually shortens the annotation time and reduces the costs [36], [15].

Of course, the annotation conventions are expressed in English to facilitate their use for similar tasks in other languages. As is usual the case with semantic labels, they are expected to be applicable without adaptation effort to any language. Most of the examples are extracted from the Romanian version of

[6] The version is the one translated by Remus Luca and Elena Linţă and published at Tenzi Publishing House in 1991.

the book and then aligned passages are searched in an English version[7]. The examples are meant to show also the specificities of different language versions: syntactic features, but mainly idiosyncrasies of translation make that in certain cases the notations pertinent to the two versions differ. We make special notes when these cases occur.

3 Similar work

Murphy [31] reviews several properties of semantic relations, among them unaccountability, which basically marks relations as an open class. If considered between words and not entities, then relations complement syntactic theories, like functional dependency grammars [49] or HPSG [42]. Taken to its extreme, we might say that the very meanings of words in contexts is constituted by the contextual relations they are in [11]. Lyons [23], another important schooler of the British structural semantic tradition, considers that (p.443) "as far as the empirical investigation of the structure of language is concerned, the sense of a lexical item may be defined to be, not only dependent upon, but identical with, the set of relations which hold between the item in question and the other items in the same lexical system".

In NLP applications, as in this work, the use of semantic relations is, generally, understood in a more narrower sense. We are concerned about relations between concepts or instances of concepts, i.e. conceptualisations, representations on a semantic layer of persons, things, ideas, which should not be confused with relations between terms, i.e. words, expressions or signs, that are used to express these conceptualisations. For instance, novel, as a concept, should be distinguished from *novel*, as an English word, or *roman*, its Romanian equivalent. The concept novel may be expressed by the terms or expressions *novel*, *fiction* or *fantasy writing*. The relation between *novel* and *fiction* is a synonymy relation between two words. These words are synonyms because there exist contexts of use where they mean the same thing. Examples of lexical databases are WordNet [28], where lexical items are organized on synonymy sets (synsets), themselves representing concepts placed in a hierarchy, or Polymots [16], that reveals and capitalizes on the bidirectional links between the semantic characterization of lexical items and morphological analogies.

We place our work in the semantic register, there where unambiguous meanings about words or expressions in contexts have been formed and what is looked for are the relations that the text expresses between these conceptualisations, or entities. The domain is known as *entity linking*.

One kind of semantic relation is the hyperonym relation, also called *is a* (and usually noted ISA), linking a hyponym to a hyperonym, or an instance to a class, in a hierarchical representation, for instance in a taxonomy or an ontology (A is a kind of B, A is subordinate to B, A is narrower than B, B is broader than A). Dictionaries usually use the Aristotelian pattern to define a *definiendum* by

[7] Translation in English by Jeremiah Curtin, published by Little Brown and Company in 1897.

identifying a *genus proximus* and showing the *diferentiae specificae* with respect to other instances of the same class [14]. Long time ago, Quillian [37] imagined a model of human memory, thus introducing the concept of semantic network, a knowledge graph representation in which meaning is defined by labelled relations (connections) between concepts and their instances. The most common relation in these representations are the ISA relations (for example, "Petronius *is a* proconsul in Bithynia"), but other types include *part of*, *has as part*, etc.

Most of the work in semantic relations and entity linking addresses the recognition problem and, on a secondary scale, the issue of building significant corpora to support this activity. If NLP systems are to reach the goal of producing meaningful representations of text, they must attain the ability to detect entities and extract the relations which hold between them.

The term *entity linking* usually expresses the task of linking mentions of entities that occur in an unstructured text with records of a knowledge base (KB). As such, the most important challenges in entity linking address: name variations (different text strings in the source text refer the same KB entity), ambiguities (there are more than one entity in the KB a string can refer to) and absence where there is no entity description in the KB to which a string in the source text representing an entity could possibly match with [7], [12]. Let's note also that in the interpretation of fictional texts, each time a process starts, the KB is empty and it will be populated synchronously with the unfolding of the text and the encountering of first mentions.

Another point of interest is the detection of relations holding between entities and events. Mulkar-Mehta et al. [30] for instance, focused on recognising the part-whole relations between entities and events as well as causal relations of coarse and fine granularities. Bejan and Harabagiu [4] and Chen et al. [9] showed that coreferences between events can be detected using a combination of lexical, part-of-speech (POS), semantic and syntactic features. Their corpus, the *EventCorefBank*, restricted by the Automatic Content Extraction exercise to the domains of *Life*, *Business*, *Conflict* and *Justice*, contained articles on 43 different topics from the Google News archive. Also, Cybulska and Vossen [13] annotated the *Intelligence Community Corpus*, at coreference mentions between violent events as bombings, killings, wars, etc.

Corpora detailing semantic relations in fictional texts are rare if not inexistent. Usually supporting the activities of semantic analysis are texts belonging to the print press.

Some have been annotated with predicate-argument structures and anaphoric relations, as the *Kyoto University Text Corpus* [21] and the *Naist Text Corpus* [20]. The anaphoric relations are categorized into three types [26]: coreference, annotated with the "=" tag, bridging reference, that can be expressed in the form B of A, annotated by "NO:A" to B, and non-coreference anaphoric relations, annotated with "$\simeq$". The *Balanced Corpus of Contemporary Written Japanese* (BCCWJ)[8] includes publications (books, magazines) and social media texts (blogs and forums) annotated with predicate-argument structures as de-

[8] `http://www.tokuteicorpus.jp/`

fined in FrameNet [33]. They do not annotate inter-sentence semantic relations. Although the predicate-argument structures of FrameNet include the existence of zero pronoun, referents are not annotated if not existent in the same sentence. Since anaphoric relations are not annotated, they do not annotate the inter-sentence semantic relations.

Another type of annotation regarding the semantic level, focusses specifically on anaphoric relations, including zero anaphora, as is the *Live Memories Corpus* [40], originating in the Italian Wikipedia and blogs. Since in Italian pronoun-dropping only occurs in the subject position (same as in Romanian), they transfer to the corresponding predicates the role of anaphors. To the same category belongs also the *Z-corpus* [39], incorporating Spanish law books, textbooks and encyclopedia articles, treating zero anaphora. There too, pronoun-dropping is marked in the subject position.

In all cases, corpora annotated to semantic links are intended to be used to train recognition algorithms. In principle, the annotation layers, the constraints used in the annotation, and the annotation conventions should be related to the set of features used by the recogniser and not to the learning methods used in training. For that reason, in the following we will revise briefly algorithms and sets of features used in the training of semantic links detection.

As regards the methods used in recognition, some approaches use supervised machine learning to match entity mentions onto their correspondent KB records. Rao et al. [38] score entities contained in the KB for a possible match to the query entity. Many studies make use of syntactic features extracted by deep or shallow parsing, POS-tagging and named entity annotation [34]. The 2012 Text Analysis Conference launched the *Knowledge Base Population Cold Start* task, requiring systems to take a set of documents and produce a comprehensive set of <Subject, Predicate, Object> triples that encode relationships involving named-entities. Starting with [18], many studies have used patterns incorporating information at the lexical and syntactic level for identification of instances of semantic relationships ([3], [17], [47]). The system KELVIN, for instance, integrates a pipeline of processing tools, among which a basic tool is the BBN's SERIF (Statistical Entity & Relation Information Finding). SERIF does named-entities identification and classification by type and subtype, intra-document co-reference analysis, including named, nominal and pronominal mentions, sentence parsing, in order to extract intra-sentential relations between entities, and detection of certain types of events [6].

An accurate entity linking technique is dependent on a diversity of NLP mechanisms, which should work well in workflows. Ad-hoc linking techniques are on the class of one-document and cross-document anaphora resolution ([2], [43], [45]). RARE is a system of anaphora resolution relying on a mixed approach that combines symbolic rules with learning techniques ([10], [36]). A recently improved version of it, developed for the ATLAS project[9] has given good results for a number of European languages ([1]).

[9] `http://www.atlasproject.eu/`

4 Annotation conventions

Vivi Năstase, citing Levi [22] and Ó Séaghdha and Copestake [44] in her book [32] enumerates a set of principles for relation inventories:

- the inventory of relations should have good coverage;
- relations should be disjunct, and should describe a coherent concept;
- the class distribution should not be overly skewed or sparse;
- the concepts underlying the relations should generalize to other linguistic phenomena;
- the guidelines should make the annotation process as simple as possible;
- the categories should provide useful semantic information.

In this section we present a set of annotation conventions that observe the above principles and were put at the bases of the "Quo Vadis" corpus.

4.1 Layers of annotation

The Romanian automatic pre-processing chain applied on the raw texts of the book consists of the following tasks, executed in sequence:

- segmentation at sentence level (marks the sentence boundaries in the raw book text);
- tokenization (demarcates words or word compounds, but also numbers, punctuation marks, abbreviations, etc.);
- part-of-speech tagging (identifies POS categories and morpho-syntactic information of tokens);
- lemmatization (determines lemmas of words);
- noun phrase chunking (explores the previous generated data and adds information regarding noun phrase boundaries and their head words) [46].

Let's note that we have not find a standard for annotating entities and relations. Năstase [32] says on this issue: "A review of the literature has shown that almost every new attempt to analyze relations between nominals leads to a new list of relations. We observe that a necessary and sufficient list of relations to describe the connections between nominals does not exist". As such, we went on with our own suggestions, knowing well that, at any moment in the future, if a need to adopt a standard will arise, an automatic conversion will be possible.

4.2 Annotating entities

Let's note that our intention is to put in evidence entities such as are they mentioned in a piece of literature. These are characters or groups that play different roles in the development of the story. A human reader usually builds a mental representation for each of them the very moment those characters (or groups) are mentioned first, and these representations are recalled from memory any time they are evoked subsequently. The mental representation associated

with a character may change while the story unfolds, although a certain mapping remains constant. It is just like we associate a box or a container with each character and afterwards we fill it with details (name, sex, kinship connections, composition, beliefs, religion, etc.). Some of these details may change as the story goes on, only the container remains the same. Any mention of that character is a mapping from a text expression to the corresponding container. In text, we annotate mentions, not containers, but recreate them after processing the coreference mappings, as will become clear in the next sections. So, what we call entities are these containers, or empty representation structures, as holders to associate text mentions on. However, as will be shown later in section 4.3, the notation we use for entities' mentions is an XML element also called `ENTITY`.

We concentrate only on entities of type `PERSON` and `GOD`, but group of persons are also annotated as `PERSON-GROUP`, occupations or typologies - coded as `PERSON-CLASS`, human anatomical parts, coded `PERSON-PART` and names of persons, coded `PERSON-NAME`. Similarly, there will be: `GOD-GROUP`, `GOD-CLASS`, `GOD-PART` and `GOD-NAME`,although very few of these last types, if any, really occurred. It is well known that an isomorphism exists between the world of humans and that of gods. In the Greek and then in the Roman antiquity, co-exist the same types of relations as those holding among humans. The man Christ became a god in the Christian religion. As such, to talk about men and women and to neglect gods was not an option, because would have created an artificial barrier.

Syntactically, the text realisation of entities are nominal phrases (NPs). Using a term common in works dedicated to anaphora, we will also call them referential expressions (REs), because the role of these expressions is to recall from memory (or refer to) the respective containers (where the features add on). We will use the term NP when we discuss syntactic properties of the expression and RE, when we discuss text coverage and semantic properties. It will be normal, therefore, to say that a NP has a syntactic head and a RE mentions (or represents, or refers) an entity. A noun phrase normally has a nominal or pronominal head and can include modifiers: adjectives, numerals, determiners, genitival particles, and even prepositional phrases. Some examples are[10]: [*Ligia*], [*Marcus Vinicius*], [*împaratul*] ([*the emperor*]), [*al lui Petronius*] ([*of Petronius*]), [*el*] ([*he*]), [*imperiul Roman*] ([*the Roman empire*]), [*un grup mare de credincioşi*] ([*a big group of believers*]). There is one exception to this rule: nominal phrases realised by pronominal adjectives, as [*nostru*] (*our*) or [*ale noastre*] (*ours*). We do not include relative clauses (relative pronouns prefixing a verb and, possibly, other syntactic constituents) in the notation of entities. A relative pronoun is marked as an individual entity. Example: [*Petronius*], [*care*] *era...* ([*Petronius*], [*who*] *was...*).

Not marked are also the reflexive pronouns in the reflexive forms of verbs, like in: *ei se spală* (*they* REFL-PRON *wash*); but other reflexive pronouns not appearing in a verbal compound are marked: *sieşi*, *sine* (*herself*, *himself*), etc.

[10] In all examples of this chapter we will notate occurrences of entities between square brackets, and we will prefix them with numbers to distinguish among them, there where their identities are important.

A NP may textually include another NP. We will say they are "imbricated", and, by abuse of language, sometimes we will say the corresponding entities are also "imbricated". It should be noted that imbricated NPs have always separate heads and they represent always distinct entities. NPs heads would be, therefore, sufficient to represent entities. Still, because we want our corpus to be useful inclusively for training NP chunkers, as REs we notate always the whole NP constructions, not only their heads. When more imbricated NPs have the same head, only the longest is annotated as an entity. Example: [*alte femei din* [*societatea înaltă*]] ([*other women of* [*the high society*]]), and not [*alte* [*femei*] *din* [*societatea înaltă*]] or [*alte* [*femei din* [*societatea înaltă*]], because [*femei*] as well as [*femei din societatea înaltă*] and [*alte femei din societatea înaltă*] all have the same head: "*femei*". Another syntactic constraint imposes that there are not NPs that intersect and are non-imbricated.

We have instructed our annotators to try to distinguish identification descriptions from characterisation descriptions. For instance, in *acest bărbat, stricat până-n măduva oaselor* (*this man, rotted to the core of his bones*), *stricat până-n maduva oaselor* (*rotted to the core of his bones*) is a caracterisation description. It does not help in the identification of a certain men among many and should be neglected in the notation of a RE. Alternatively, in *convoi de fecioare* (*a band of maidens*) - the sequence *de fecioare* (*of maidens*) is an identification description, because it uniquely identifies "the band" among many others and it should be included in the RE aimed to refer that band as an entity group. Only identification descriptions will be marked as REs.

4.3 Annotating relations

One class of anaphoric relations and three classes of non-anaphoric relations are scrutinised, each with sub-types. We present annotation conventions and methodological prerequisites based on which a corpus that puts in evidence characters and relations mentioned as holding between them has been manually built.

As will be seen in the following sub-sections, each relation holds between two arguments, that we will call *poles*, and, with one exception, is signalled by a word or an expression, that we will call *trigger*. In general, when marking relations we want to evidence the minimal span of text in which a reader deciphers a relation. Excepting for coreferential relations, in which poles can be sometimes quite distant in text and there is nothing to be used as a trigger, usually relations are expressed locally in text, within a sentence, within a clause, or even within a noun phrase. As such, excepting for coreferentiality, each relation span should cover the two poles and the trigger.

Our notations are expressed in XML. Basic layers of annotation include: borders of each sentence (marked as `<S></S>` elements, and identified by unique IDs) and words (marked as `<W></W>` and including unique IDs, lemma and morpho-syntactic information). Above these basic layers the annotators marked three types of XML elements:

- `ENTITY` - delimiting REs, including the attributes: `ID`, `TYPE` and, optionally, `HEAD`; as will be explained below, for included subjects (pronoun-dropping) the verb is annotated instead as an `ENTITY`;
- `TRIGGER` - marking relations' triggers; it delimits a word (`<W>`) or an expression (sequence of `<W>`);
- `REFERENTIAL`, `AFFECT`, `KINSHIP` and `SOCIAL` - mark relations. With the exception of coreferential relations, these markings delimit the minimal spans of text that put in evidence these types of relations. Their attributes are: a unique ID, the sub-type of the relation (listed below), the two poles and the direction of the relation (the attributes `FROM` and `TO`), and the ID of a trigger (the attribute `TRIGGER`).

The two poles of a relation could be intersectable or not. It they are intersectable, then they are necessarily nested and the direction convention is to consider the `FROM` entity the larger one and the `TO` entity the nested one.

1:[*celui de-al doilea <soţ>* 2:[*al Popeii*]] (1:[*to the second husband* 2:[*of Popeea*]]) ⇒ [1] spouse-of [2][11]

As already mentioned, the coreferential relation could never be expressed between nested REs. If the RE poles are not nested, we adopted a right-to-left direction in annotating coreferential relations. This decision will be defended in the next sub-section. For the rest of relations the convention is to consider the direction as indicated naturally by reading the trigger and its context. For instance, in the text "X loves Y", the relation `love`, announced by the trigger *loves*, is naturally read as [X] `love` [Y], therefore with `FROM`=X and `TO`=Y, but in "X is loved by Y", the relation will be [X] `loved-by` [Y].

It could happen that a pole of a relation is not explicitly mentioned. This happens in cases of included subjects, when the subjects are expressed by null (or dropped) pronouns. In Romanian, the morphological properties of the subject are included in the predicate, such that the missing pole will be identified with the predicate.

Example. *dar* (1:[*îl*] *şi* 2:[<*iubeau*>, `REALISATION="INCLUDED"`]) *din tot sufletul* (2:[<*loved*> `REALISATION="INCLUDED"`] 1:[*him*] *with the whole soul*)) ⇒ [2] `loves` [1]
It should be noted that a word could be simultaneously marked as a token (`<W>`), trigger (`<TRIGGER>`) and entity (`<ENTITY>`). For instance, *iubeau* (love-PAST-TENSE) in the notation below has all three markings. The value of the `FROM` attribute of the `AFFECT` element will filled in by the `ID` of the verb *iubeau*, marked as an `ENTITY`, while the value of the `TRIGGER` attribute in the same relation will be the `ID` of the `TRIGGER` element covering the same word.

[11] To save space, in the notations showing relations on our examples, we will mark in labeled square brackets, as before, the entities and in pointed brackets - the triggers; the relations themselves are indicated by their sub-types; sometimes, when there is a need to associate triggers to their corresponding relations, these are also labeled.

```
dar
 <AFFECT ID="..." TYPE="LOVE" FROM="E47" TO="E46" TRIGGER="T17">
   <ENTITY ID="E46">
     <W ID="..." POS="..." LEMMA="el">îl</W>
   </ENTITY>
   şi
   <ENTITY ID="E47" REALISATION="INCLUDED">
     <TRIGGER ID="T17">
        <W ID="W45" POS="..." LEMMA="iubi">iubeau</W>
     </TRIGGER>
   </ENTITY>
</AFFECT>
din tot sufletul
```

When a relation is expressed through the eyes of another character, being perceived particularly as such by this one, or is still uncertain or to be realised in the future we say that the relation is "interpreted". As such, a relation `R` will be marked as `R-interpret`. All types and sub-types of relations could have `interpret`-ed correspondents.

4.4 Referential relations

When the understanding of one RE-to-entity mapping depends on the recuperation in memory of a previously mentioned entity (the container together with its accumulated content), we say that a referential relation occurs between this RE (called anaphor) and that entity (called antecedent). In the literature, this definition presents variants, some (as [29]) insisting on the textual realisation of the relation, the anaphor and the antecedent being both textual mentions, while others ([10]) putting in evidence its cognitive or semantic aspects as such, the anaphor being a textual mention and the antecedent - an entity as represented on a cognitive layer, therefore, in our terms - a container plus its content. Supplementary, some authors also make the distinction between anaphora (when a less informative mention of an entity succeeds a more informative one; for instance, a pronoun follows a proper noun) and cataphora (when the other way round is true; for instance, the pronoun mention comes before the proper noun mention). It is to notice however, as [48] and others have noticed, that cataphora could be *absolute* (when the text includes no more informative reference to the entity before the less informative one) or *relative* (when the inversion takes place at the level of a sentence only, a more informative mention being present in a sentence that precedes the one the pronoun belongs to).

In order to mark the direction of a referential relation, for non-imbricated REs, in connection with text unfolding (a more recent RE mentions an entity introduced or referred by a previously mentioned RE), the annotation of the poles of the coreferential relations are as follows: the source (`FROM`) is the more

recent one and the destination (TO) is the older one. In Romanian[12], this direction is from the right of the text to its left. For example, in *pe* 1:[*Ligia*] 2:[*o*] *iubesc* (1:[*Ligia*] *is* 2:[*the one*] *I love*), the relation is marked from [2] to [1]. Although perhaps less intuitive, the direction of cataphoric relations comply with the same right-to-left annotation convention, on the ground that a container (possibly including only scarce details) must have been introduced even by a less informative mention, while the more informative mention, coming after, refers back in memory to this conceptualisation, and injects more information in there [10]. In the text 1:[*îl*] *chemă pe* 2:[*Seneca*] (*he* 1:[*him_*clitic] *summoned* 2:[*Seneca*]) the direction is also from [2] to [1]. Deciphering anaphoric relations in the Romanian language is perhaps more complex than in other languages, mainly due to the duplication of the direct and indirect complements by unaccented forms of pronouns. But we will refrain from making anaphora resolution comments in the present study as this topic is outside our declared intent.

We established nine sub-types of referential relations, listed and exemplified below.

- `coref`: by slightly modifying the definition given above for referentiality, we say that we have a coreferential relation between a RE and an entity E when we understand the RE-to-E identity mapping based on the recuperation in memory of E, a previously mentioned entity. `Coref` is a symmetric relation, where poles could be of types `PERSON`, `PERSON-GROUP`, `GOD` and `GOD-GROUPS`, but always with both poles of the same category. It is important to notice that a `coref` relation can never occur between imbricated REs. Examples:

  1:[*Marcus Vicicius*]... 2:[*el*]... (1:[*Marcus Vicicius*]... 2:[*he*]...) ⇒ [2] `coref` [1];
  1:[*Ligia*]... 2:[*tânara libertă*]... (1:[*Ligia*]... 2:[*the young libert*]...) ⇒ [2] `coref` [1];
  Nu avea nici cea mai mică îndoială că 1:[*lucrătorul acela*] *e* 2:[*Ursus*]. (*He had not the least doubt that* 1:[*that laborer*] *was* 2:[*Ursus*].)
  ⇒ [2] `coref-interpret` [1];
  L-am prezentat pe 1:[*acest Glaucus*] *ca pe* 2:[*fiul Iudei*] *şi* 3:[*trădător al tuturor creştinilor*]. (*I described* 1:[*Glaucus*] *as* 2:[*a real son of Judas*], *and* 3:[*a traitor to all Christians*].) ⇒ [2] `coref-interpret` [1], [3] `coref-interpret` [1];

- `member-of`(a `PERSON` type RE is a `member-of` a `PERSON-GROUP` entity and, similarly, a `GOD` is a `member-of` a `GOD-GROUP`), a directed relation. Example:

  1:[*o femeie din* 2:[*societatea înaltă*]] (1:[*a woman of* 2:[*the high society*]])
  ⇒ [1] `member-of` [2];

- `has-as-member` (the inverse of `member-of`, from a `PERSON-GROUP` to a `PERSON`, or from a `GOD-GROUP` to a `GOD`), directed:

[12] contrary, for instance, to Semitic languages

1:[*Petronius*]... 2:[*amândurora*] (1:[*Petronius*]... *to* 2:[*both of them*]) ⇒ [2] `has-as-member` [1];
1:[*Ursus*]... 2:[*Ligia*]... 3:[*voi*] (1:[*Ursus*]... 2:[*Ligia*]... 3:[*you_*PL]) ⇒ [3] `has-as-member` [1]; [3] `has-as-member` [2];

– `isa` (from a `PERSON` type RE to its corresponding `PERSON-CLASS`, or from a `GOD` to its corresponding `GOD-CLASS`), directed;

1:[*naşă*] *să-mi fie* 2:[Pomponia] (*and I wish* 2:[*Pomponia*] *to be* 1:[*my godmother*]) ⇒ [2] isa [1] only in the Romanian version; in the English version the two REs are inverted, which gives here the inverse relation (see next);

– `class-of` (the inverse of `isa`, from a `PERSON-CLASS` to an instance of it of type `PERSON`, or from a `GOD-CLASS` to a `GOD` type instance), directed:

Dar nu eşti 1:[*tu*] 2:[*un zeu*]? (*But are* 1:[*thou*] *not* 2:[*a god*]?) ⇒ [2] `class-of-interpret` [1] ([1] is seen as a God by someone);
daţi-mi-1:[*o*] *de* 2:[*nevastă*] (*Give* 1:[*her*] *to me as* 2:[*wife*]) ⇒ [2] `class-of-interpret` [1][13];
Se trezise în 1:[*el*] 2:[*artistul*], 3:[*adoratorul frumuseţii*]. (2:[*The artist*] *was roused in* 1:[*him*], *and* 3:[*the worshipper of beauty*]) ⇒ for the Romanian version: [2] `class-of-interpret` [1]; [3] `class-of-interpret` [1]; for the English version, because of the inversion: [1] `isa` [2]; [3] `class-of-interpret` [1];

– `part-of` (a RE of type `PERSON-PART` is a part of the body of an entity of type `PERSON`, or a `GOD-PART` is a part of the body of an entity of type `GOD`), directed:

1:[*mâna* 2:[*lui*] *dreaptă*] (1:[2:[*his*] *right hand*]) ⇒ [1] `part-of` [2];

– `has-as-part` (the inverse of `part-of`: a `PERSON` type RE has as a component part a `PERSON-PART` entity, or a `GOD` type RE has as a component part a `GOD-PART` entity), directed;

chinurile, 1:[*sângele*] *şi moartea* 2:[*Mântuitorului*] (*the torment,* 1:[*the blood*] *and the death* 2:[*of the Saviour*]) ⇒ [2] `has-as-part`[1];

– `subgroup-of` (from a subgroup, i.e. a `PERSON-GROUP` type RE, to a larger group, i.e. also a `PERSON-GROUP` type entity which includes it, and similarly for `GOD-GROUP`'s poles), directed:

1:[*a*], 2:[*b*], 3:[*c*] *şi* 4:[*alte femei din* 5:[*societatea înaltă*]] (1:[*a*], 2:[*b*], 3:[*c*]

[13] [1] could become a wife of the speaker but is actually not.

and 4:[*other women of* 5:[*the high society*]]) ⇒ [5] `has-as-member` [1], [5] `has-as-member` [2], [5] `has-as-member` [3], [4] `subgroup-of` [5];
Christos 1:[*i*]*-a iertat şi pe* 2:[*evreii*] *care i-au dus la moarte şi pe* 3:[*soldaţii romani*] *care l-au ţintuit pe cruce.* (*Christ forgave* 2:[*the Jews*] *who delivered him to death, and* 3:[*the Roman soldiers*] *who nailed him to the cross*) ⇒ for the Romanian version only: [2] `subgroup-of` [1]; [3] `subgroup-of` [1]. The `subgroup-of` relation holds in the Romanian version because of the existence of the anticipating pronoun 1:[*i*], which signifies both groups. In the English equivalent no such mention appears and a `subgroup-of` relation cannot be formulated;

– `name-of` (inverse of `has-name`, linking a `PERSON-NAME` RE to a `PERSON` entity), directed:

  1:[*numele lui* 2:[*Aulus*]] (1:[*the name of* 2:[*Aulus*]]) ⇒ [1] `name-of` [2];
  Petronius ... care simţea ca pe statuia 1:[*acestei fete*] *s-ar putea scrie*: 2:["*Primavara*"]. (*Petronius ... who felt that beneath a statue of* 1:[*that maiden*] *one might write* 2:["*Spring.*"])⇒ [2] `name-of-interpret` [1] (`-interpret` because Petronius is the one that gives this name).

4.5 Kinship relations

Kinship (or family) relations (marked KINSHIP as XML elements) occur between `PERSON`, `PERSON-GROUP`, `GOD` and `GOD-GROUP` type of REs and entities. Seven subtypes have been identified, detailed below:

– `parent-of` (the relation between a parent or both parents and a child or more children; a RE A is in a `parent-of` relation with B if A is a parent of B, i.e. mother, father, both or unspecified), directed:

  1:[<*tatăl*> 2:[*lui Vinicius*]] (1:[2:[*Viniciu's*] <*father*>]) ⇒ [1] `parent-of` [2];

– `child-of` (inverse of `parent-of`; a RE A is a `child-of` B if the text presents A as a child or as children of B), directed:

  1:[*Ligia mea*] *este* <*fiica*> 2:[*regelui*] (1:[*My Lygia*] *is the* <*daughter*> 2:[*of that leader*].) ⇒ [1] `child-of` [2];
  1:[<*copilul*> *drag al* 2:[*celebrului Aulus*]] (1:[*a dear* <*child*>] 2:[*of the famous Aulus*]) ⇒ [1] `child-of` [2];

– `sibling-of` (between brothers and sisters), symmetric:

  1:[*sora lui* 2:[*Petronius*]] (1:[2:[*Petronius's*] <*sister*>]) ⇒ [1] `sibling-of` [2];
  1:[*nişte* <*fraţi*> *ai* 2:[*tăi*]] (1:[*some of* 2:[*your*] <*brothers*>]) ⇒ [1] `sibling-of` [2];

– `nephew-of` (we have A `nephew-of` B, if A is a nephew/niece of B), directed:

  1:[*scumpii* 2:[*săi*] <*nepoţi*>] (1:[2:[*his*] *dear* <*nephews*>]) ⇒ [1] `nephew-of` [2];

– `spouse-of` (symmetric relation between husbands):

  ... *cu* 1:[*care*] *mai târziu* 2:[*Nero*], *pe jumatate nebun, avea să se* <*cunune*> (... *to* 1:[*whom*] *later* 2:[*the half-insane Nero*] *commanded the flamens to* <*marry*> *him*) ⇒ [2] `spouse-of` [1];
  1:[*Vinicius*] *ar putea să* 2:[*te*] *ia de* <*nevastă*> (1:[*Vinicius*] *might* <*marry*> 2:[*thee*]) ⇒ [2] `spouse-of-interpret` [1];

– `concubine-of` (symmetric relation between concubins)

  1:[<*concubina*> 2:[*ta*]] (1:[2:[*your*] <*concubine*>]) ⇒ [1] `concubine-of` [2];

– `unknown` (a kinship relation of an unspecified type):

  1:[*o* <*rudă*> *de-a* 2:[*lui Petronius*]] (1:[*a* <*relative*> *of* 2:[*Petronius*]]) ⇒ [1] `unknown` [2];
  1:[<*strămoşilor*> 2:[*lui Aulus*]] (1:[2:[*Aulus's*] <*ancestors*>]) ⇒ [1] `unknown` [2]

4.6 Affective relations

Affective relations (marked as AFFECT elements in our XML notations) are non-anaphoric relations that occur between REs and entities of type `PERSON`, `PERSON-GROUP`, `GOD` and `GOD-GROUP`. There are eleven subtypes, as detailed below:

– `friend-of` (A is a `friend-of` B, if the text expresses that A and B are friends), symmetric:

  1:[<*tovarăşii*> 2:[*lui*]] (1:[2:[*his*] <*comrades*>]) ⇒ [1] friend-of [2];
  1:[*Vinicius*] *e un nobil puternic, spuse el, şi* <*prieten*> *cu* 2:[*împăratul*]. (1:[*Vinicius*] *is a powerful lord, said he, and a* <*friend*> *of* 2:[*Cæsar*].) ⇒ [1] `friend-of-interpret` [2];

– `fear-of` (A is in a relation `fear-of` with B if the text expresses that A feels fear of B), directional:

  1:[*oamenii*] <*se tem*> *mai mult de* 2:[*Vesta*] (1:[*people*] <*fear*> 2:[*Vesta*] *more*) ⇒ [1] `fear-of` [2];
  1:[*Senatorii*] *se duceau la* 2:[*Palatin*], <*tremurând de frică*> (1:[*Senators*], <*trembling in their souls*>, *went to the* 2:[*Palatine*]) ⇒ [1] fear-of [2];

– `fear-to` (inverse of `fear-of`: A is in a relation `fear-to` B is the text expresses that the RE A inspires fear to the entity B), directional:

1:[*Nero*] *îi <alarma> chiar şi pe* 2:[*cei mai apropiaţi*] (1:[*Nero*] *did <roused attention>, even in* 2:[*those nearest*]) ⇒ [1] `fear-to` [2];

– `love` (A is in a relation `love` to B, if A loves B), directional:

1:[*Ligia*] *simţi că o mare greutate i s-a luat de pe inimă. <Dorul> acela fără margini după* 2:[*Pomponia*] (1:[*She*][14] *felt less alone. That measureless <yearning> for* 2:[*Pomponia*]) ⇒ [1] `love` [2];
<îndrăgostit> ca 1:[*Troilus*] *de* 2:[*Cresida*] (*<in love>, as was* 1:[*Troilus*] *with* 2:[*Cressida*]) ⇒ [1] `love` [2];

– `loved-by`(inverse of `love`: A `loved-by` B, if A is loved by B):

<iubită> este 1:[*Ligia*] *de* 2:[*familia lui Plautius*] (*<dear>* 1:[*Lygia*] *was to* 2:[*Plautius*]) ⇒ [1] `loved-by` [2];

– `rec-love` (A `rec-love` B if the text mentions a mutual love between A and B), symmetric:

<îndrăgostiţi> 1:[*unul*] *de* 2:[*altul*] (*in <love> with* 1:[*each*] 2:[*other*]) ⇒ [1] `rec-love`[2];

– `hate` (A `hate` B, if the text mentions that A hates B), directional:

Pe 1:[*Vinicus*] *îl cuprinse o <mânie> năprasnică şi împotriva* 2:[*împăratului*] *şi împotriva* 3:[*Actei*] (1:[*Vinicius*] *was carried away by sudden <anger> at* 2:[*Cæsar*] *and at* 3:[*Acte*].) ⇒ [1] `hate` [2], [1] `hate` [3];

– `hated-by` (A `hated-by` B, if A is hated by B), directional:

<ura> pe care 1:[*i*]*-o purta* 2:[*prefectul pretorienilor*] (*<hatred toward>* 1:[*him*] *of* 2:[*the all-powerful pretorian prefect*]) ⇒ [1] `hated-by` [2]

– `upset-on` (A `upset-on` B, if the text tells that A feels upset, disgust, anger, discontent, etc. on B), directional:

1:[*<Dispreţuia>* `REALISATION="INCLUDED"`] 2:[*mulţimea*] (1:[*He*] *had a two-fold <contempt for>* 2:[*the multitude*]) ⇒ [1] `upset-on`[2];

– `worship` (A `worship` B, if the text mentions that A worships B), directional:

[14] In the English equivalent, the mention of Ligia ([1]) is missing.

1:[*oamenii aceia*] *nu numai că-şi* <*slăveau*> 2:[*zeul*] (1:[*those people*] *not merely* <*honored*> 2:[*their God*]) ⇒ [1] `worship` [2];
1:[*Ligia*] *îngenunche ca să se* <*roage*> 2:[*altcuiva*]. (*But* 1:[*Lygia*] *dropped on her knees to* <*implore*> 2:[*some one else*].) ⇒ [1] `worship` [2];

– `worshiped-by` (*A* `worshiped-by` *B* if the text mentioned that *A* is worshiped by *B*), directional:

1:[*un zeu cu totul neînsemnat*] *dacă n-are decât* 2:[*două* <*adoratoare*>] (1:[*a very weak god*], *since he has had only* 2:[*two* <*adherents*>]) ⇒ [1] `worshiped-by` [2]

4.7 Social relations

The group of social relations (marked `SOCIAL` in our XML annotations) are non-anaphoric relations occurring only between `PERSON` or `PERSON-GROUP` REs and entities. They are grouped in six subtypes, as detailed below:

– `superior-of` (*A* `superior-of` *B*, if *A* is hierarchically above *B*), directional:

<*Eliberând*>-1:[*o*], 2:[*Nero*] (2:[*Nero*], *when he had* <*freed*> 1:[*her*]) ⇒ [2] `superior-of` [1];
1:[*Nero*] *a ordonat* <*predarea*> 2:[*mea*] (1:[*Nero*] *demanded* 2:[*my*] <*surrender*>) ⇒ [1] `superior-of` [2];
1:[*un centurion*] <*în fruntea*> 2:[*soldaţilor*] (1:[*a centurion*] <*at the head*> 2:[*of soldiers*]) ⇒ [1] `superior-of` [2];

– `inferior-of` (inverse of `superior-of`, *A* `inferior-of` *B* if *A* is hierarchically subordinated to *B*), directional:

1:[<*consul*> *pe vremea* 2:[*lui Tiberiu*]] (1:[*a man of* <*consular*> *dignity from the time* 2:[*of Tiberius*]) ⇒ [1] `inferior-of` [2];
1:[*Tânărul*] *luptase* <*sub comanda*> 2:[*lui Corbulon*] (1:[*The young man*] *was serving then* <*under*> 2:[*Corbulo*]) ⇒ [1] `inferior-of` [2];
1:[<*libertei*> 2:[*lui Nero*]] (1:[2:[*Nero's*] <*freedwoman*>]) ⇒ [1] `inferior-of` [2];

– `colleague-of` (*A* `colleague-of` *B* if the text explicitly places *A* on the same hierarchical level with *B*), symmetrical:

1:[<*tovarăşii*> 2:[*săi*]] (1:[2:[*his*] <*companions*>]) ⇒ [1] `colleague-of` [2];

– `opposite-to` (*A* `opposite-to` *B*, if *A* is presented in a position that makes her/him opposing to *B*), directional:

Să nici nu-ţi treacă prin gând să 1:[*te*] <*împotriveşti*> 2:[*împăratului*] (*Do not even* 1:[*think*; `REALISATION="INCLUDED"`] *of* <*opposing*> 2:[*Cæsar*]) ⇒ [1] `opposite-to-interpret` [2];
1:[*Pomponia şi Ligia*] *otrăvesc fântânile*, <*ucid*> 2:[*copiii*] (1:[*Pomponia and Lygia*] *poison wells*, <*murder*> 2:[*children*]) ⇒ [1] `opposite-to` [2];

– `in-cooperation-with` (*A* is `in-cooperation-with` *B* if the text present *A* as performing something together with *B*), directional:

1:[*Vannius*] *a chemat* T1:<*în ajutor*> *pe* 2:[*iagizi*], *iar* 3:[*scumpii săi nepoţi*] *pe* 4:[*ligieni*] (1:[*Vannius*] *summoned to his* T1:<*aid*> 2:[*the Yazygi*]; 3:[*his dear nephews*] T2:<*called in*> 4:[*the Lygians*]) ⇒ [1] `in-cooperation-with` [2], trigger: <T1>; [3] `in-cooperation-with` [4], trigger: <T2>;

– `in-competition-with` (*A* is `in-competition-with` *B*, if *A* is presented as being in a sort of competition with *B*), directional:

1:[*Petronius*] 2:[*îl*] <*întrecea*> *cu mult prin maniere, inteligenţă* (1:[*Petronius*] <*surpassed*> 2:[*him*] *infinitely in polish, intellect, wit*) ⇒ [1] `in-competition-with` [2].

4.8 Examples of combinations of relations

In the end of this section we will give a few examples showing complex combinations of relations.

1:[*Vinicius*] ... *e* 2:[*o* <*rudă*> *de-a* 3:[*lui Petronius*]] (1:[*Vinicius*] ... *is* 2:[*a* <*relative*> 3:[*of Petronius*]]) ⇒ [2] coref-of [1], [2] `KINSHIP:unknown` [3];

Se repezi la 1:[*Petru*] *şi, luându-*2:[*i*] 3:[*mâinile*], *începu să* 4:[*i*] 5:[*le*] *sărute* (*... seized* 3:[*the hand of* 1:[*the old Galilean*]], *and pressed* 5:[*it*] *in gratitude to his lips.*)[15] ⇒ [2] `coref` [1]; [3] `part-of` [1] (or [2]); [4] `coref` [1] (or [2]); [5] `coref` [3]. It is superfluous to mark [5] `as part-of` [1] because it results by transitivity from it being coreferential with [3] and [3] being `part-of` [1].

1:[*Vinicius*] *şi* 2:[<*tovaraşii*> 3:[*săi*]] (1:[*Vinicius*] *and* 2:[3:[*his*] <*comrades*>]) ⇒ [3] `coref` [1]; [2] `SOCIAL:colleague-of` [3].

5 Creating the corpus

The realisation of a manually annotated corpus incorporating semantic relations obliges to a fine-grained interpretation of the text. This triggers the danger of non-homogeneity, due to idiosyncrasies of views, over the linguistic phenomena

[15] In the English equivalent, two mentions of Peter ([2] and [4]) are missing.

under investigation, of different annotators working each on different parts of the document. A correlation activity that would nivelate divergent views is compulsory. Let's add to this that many details of the annotation conventions usually settle down in iterative sessions of discussions within the group of annotators, following the rich casuistry picked up along the first phases of the annotation process. As such, the organisation of the work should be done in such a way as to certify that the result of the annotation process contains the least errors possible, and that the conventions are coherently applied over the whole document.

5.1 Organising the activity

The annotation activity of the "Quo Vadis" corpus was performed over a period of three terms with students in Computational Linguistics[16]. An Annotation Manual, including an initial set of annotation rules, was proposed by the first author to the students at the beginning of the activity and discussed with them. Then, the students went through some practical classes in which they were taught to use an annotation tool. Approximately half of the novel was split in equal slices and distributed to them and they begun to work independently or grouped by two. During the first term, in weekly meetings, annotation details were discussed, difficult cases were presented and, based on them, the Manual was refined. At the end of the first term their activity was individually evaluated and the results showed that only about 15% of them were trustful enough as to be given a full responsibility[17]. As a by-product, we had, at the time, a consistent set of annotation rules and PALinkA[18], our annotation tool, could incorporate rather stable preferences settings (describing the XML structural constraints).

We continued the activity during the next two terms with only the best ranked members of the former class (among them - a PhD researcher with a Philology background). At the beginning of the next year (September 2013), a few new students with a Philological background went through a rapid training period and joined the team (among them - two PhD researchers in Humanities). The quality improved a lot, but in the detriment of the speed, which continued to be very slow. At that moment it became clear to us that it will be impossible to achieve this ambitious task by going through the text three times or even only twice, as the usual norms for redundant annotation require in order to organise a proper inter-annotator agreement process. As a consequence, we started to think

[16] a master organised at the "Alexandru Ioan Cuza" University of Iaşi by the Faculty of Computer Science, which accommodates graduate students with either a background in Computer Science or in Humanities

[17] It was not a surprise that for annotation activities the most dedicated and skillful students were those having a Humanity background.

[18] PALinkA was created by Constantin Orăşan in the Research Group in Computational Linguistics, at the School of Law, Social Sciences and Communications, Wolverhampton. PALinkA was used for annotating corpora in a number of projects, for purposes including: anaphoric and coreferential links in a parallel French-English corpus, summarisation, different versions of the Centering Theory, coreferences in email messages and web pages, or for Romanian name entities.

at other methods for obtaining accuracy that would involve only one manual annotation pass. We imagined different methods for clearing up the corpus from errors, which will be detailed in the following sub-sections.

As shown already, the files opened in PALinkA have been previously annotated in XML with markings signalling word (`<W>`) and paragraph (`<P>`) boundaries. Over these initial elements the annotators have marked: entities, coreferential links, triggers of relations, and relation spans, including attributes indicating the poles and the triggers.

In building the manual annotation, there where words are ambiguous, we have instructed our annotators to use their human capacity of interpretation in order to decide the true meaning of words, the types of relations or the entities that are glued by relations[19]. For instance, words and expressions, based on their local sense, could functions as triggers only in some contexts (*father*, for instance should not be taken as signaling a `parent-of` relation if its meaning is that of priest).

5.2 Acquiring completeness and correctness

Along the whole process of building the "Quo Vadis" corpus, the two main preoccupations were: to acquire completeness (therefore to leave behind as few as possible unmarked entities or relations) and to enhance its quality (therefore to clean the corpus of possible errors). As said already, in order to distribute the text to different annotators, we splitted the text of the novel into chunks of relatively equal size (phase 1, in Figure 1). It resulted a number of 58 chunks, each including on average approximately 123 sentences. The following formula was used to estimate the density of annotations (D) to each chunk:

$$D = (E + 2 \times R + 5 \times (A + K + S))/N$$

where: E = number of marked entities; R = number of marked `REFERENTIAL` relations; A, K, S = number of marked `AFFECT`, `KINSHIP` and `SOCIAL` relations, N = number of sentences.

During the annotation process, the density scores per segment varied between 0 to more than 20. Assuming an approximately uniform density all over the novel[20], these scores allowed us to detect from the blink of an eye those chunks which received too little attention from the part of the annotators and to spot also the most diligent annotators. After the first round, only the best ranked annotators were retained in the team. In the second round, all chunks scored low, therefore contributed by dismissed students, were resubmitted for a second annotation round to the selected members remained in the refreshed team (4). At this moment, all chunks are scored over 5.5, the maximum reaching 20.2 and

[19] Not rare were cases when philologists asked: *And how would the machine recognise this relation when it was difficult even for me to decipher it here?!...*

[20] Not necessarily true, because long passages of static descriptions are bare of mentions of entities and, consequently, relations.

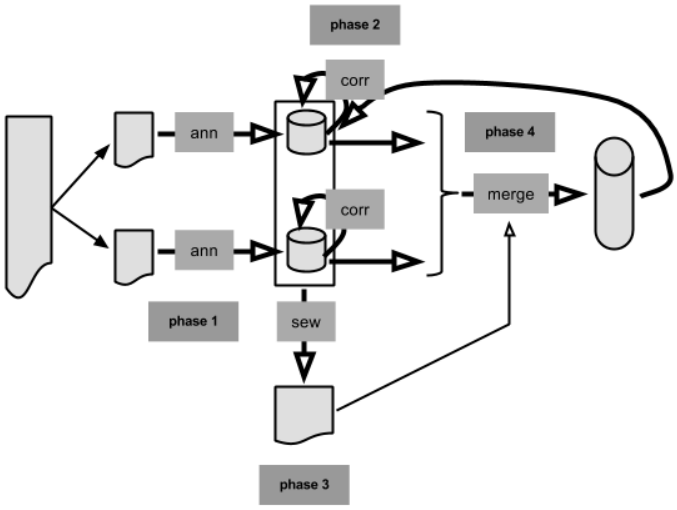

Fig. 1. Annotation-Correction-Sewing-Merging cycles in the building of the "Quo Vadis" corpus

the whole novel having an average density score of 9.4. But this score does not reflect the correctness.

The final step in the construction of the corpus was dedicated to enhancing the accuracy. As said, because of the very high complexity of the task, which makes it extremely time-consuming, and the scarcity of skilled people able to do an expert annotation task, no inter-annotator agreement has been possible to organise. However, more measures to enhance correctness were assured.

In phase 2 (Correction on Figure 1), the best trained annotators of the team received the additional task of error-proofing the annotations of their last year colleagues, updating them to the new standards and unifying them with the new ones. Then, in the 3rd phase (Sewing in Figure 1) the cross-chunks-border coreferential links were notated, as pairs of `ENTITY ID`s. These lists were then passed to the 4th phase (Merging in Figure 1), in which the chunks of annotated text were physically merged in just one file and `REFERENTIAL XML` elements with `TYPE="coref"` were added at the end of the document for all cross-border coreferential pairs. In this phase, error-detection filters were also run. These filters are described in Section 5.3. The errors signalled by the filters were passed back to annotators and they repaired the errors in the original files. Lists of coreferential entity names were also produced and these were important clues to notice errors of coreferentiality. For instance, it is impossible that an instance of Ligia appears in the same chain with Nero, and very unlikely that a plural pronoun would ever refer a character. Moreover, chains representing group characters, if containing pronouns, should include only pronouns in plural.

5.3 Error correcting filters

We list in this section a number of filterring procedures that helped to detect annotator errors.
- We call a coreference chain (CC) a list of REs whose occurrences are sequentially ordered in the text and which all represent the same `PERSON/GOD` entity or the same `PERSON-GROUP/GOD-GROUP` entity[21] ⇒ any proper noun that appears in a CC should be a variation of the name of that entity. We have extracted one occurrence for all proper names in CCs and manually verified if they are variations, inflections or nick-names of the name of the same character (Ex. *Marcus*, *Vinicius* and *Marcus Vinicius* for the character [Vinicius], or *Ligia*, *Ligiei*, *Callina*, *Callinei* for the character [Ligia], or *Nero*, *Barbă-Arămie*, *Ahenobarbus*, *Cezar*, *Cezarul*, *Cezarului*, etc. for the character [Nero]);

- All common nouns and pronouns in a CC generally have the same number+gender values[22] ⇒ For each `W` having the category common noun or pronoun in a CC we have extracted the pairs number+gender values and reported if they are not identical;

- It is improbable that an entity be referred only by pronouns ⇒ We have listed the CCs that include only pronouns and passed them to the correctors for a second look;

- In most of the cases, gods are referred to by names in capital letters ⇒ We have reported the exceptions;

- There should be a one-to-one mapping between triggers and relations ⇒ Report if a trigger is not referred in any relation or if more relations use the same trigger;

- Triggers could not appear as values of `FROM` and `TO` attributes and no element type other than `TRIGGER` could to a value of a `TRIGGER` argument of a relation ⇒ We performed an argument type checking (see [32]) by citing [35] and [41] for "matching the entity type of a relation's argument with the expected type in order to filter erroneous candidates". Combined (or coupled) constraints, as proposed by [8] in semi-supervised learning of relations in context, were not of primary interest at the moment of building the corpus.

- In the vast majority of cases, the two poles and the trigger belong to the same sentence. For instance, in the example: 1:[*e*; `REALISATION="INCLUDED"`] 2:[*-un patrician*], 3:[*prieten cu* 4:[*împăratul*]] (1:[*he*] *is* 2:[*a patrician*], 3:[*a friend* 4:[*of*

[21] Let's note that the `REFERENTIAL:coref` links should separate the whole class of `ENTITY` elements into disjoint trees. Trees and not general graphs, because considering `ENTITY`s as nodes in the graph and `REFERENTIAL:coref` relations as edges, there is just one `TO` value (parent in the graph) for each `ENTITY` node.

[22] There are exceptions to this rule: a plural may be referred by a singular noun denoting a group, or due to errors of the POS-tagger, etc.

Caesar]]), the correct annotation is as follows: [2] `class-of` [1]; [3] `class-of` [1]; [3] `friend-of` [4]. As such, the `friend-of` relation does not cross the borders of the second sentence. ⇒ We report cross-sentences non-coreferential relation spans and asked the correctors to verify them.

6 Statistics over the corpus

In this section we present a number of statistics and comment on the semantic links of the corpus from a global perspective. Table 6 presents the Corpus by numbers.

It can be seen that 20% of the tokens of the novel are covered by some manual annotation (entity, trigger, relation). The vast majority of relations are those belonging to the `REFERENTIAL` type. A comparison is shown in the diagram of Figure 2.

Counted elements	**Value**
# sentences	7,150
# tokens (`W` elements, punctuation included)	144,068
# tokens (`W` elements, excluding punctuation)	123,093
# tokens under at least one annotation (punctuation included)	28,851
# tokens under at least one relation (punctuation included)	7,520
# tokens summed up under all relations (punctuation included)	9,585
# entities	22,310
# `REF` annotations (all)	21,439
# `REF:coref` and `REF:coref-interpret` annotations	17,916
# AKS annotations	1,133
# `TRIGGER` annotations	1,097
total # annotations (`ENTITY` + `TRIGGER` + `REF` + AKS)	45,979
overall density score	10.21

Table 1. The corpus at a glance

If the 17,916 `REFERENTIAL:coref` and `REFERENTIAL:coref-interpret` relations (the most numerous) are left aside, the distribution is depicted in Figure 3. In Figure 4, the distributions of different types of `REFERENTIAL` relations (without `REFERENTIAL:coref` and `REFERENTIAL:coref-interpret`) is shown.

Figures 5, 6 and 7 show the distributions of `KINSHIP`, `SOCIAL` and `AFFECT` relations in the corpus.

Long relation spans make the discovery of relations difficult. The graphic in Figure 8 shows the ranges of lengths of `REFERENTIAL` relations spans whose lengths can be estimated, thus `REFERENTIAL:coref` and `REFERENTIAL:coref-interpret` are not considered.

As can be seen, the average span for this group of relations is placed somewhere around 20 words. In Figure 9 the same statistics is shown for the other

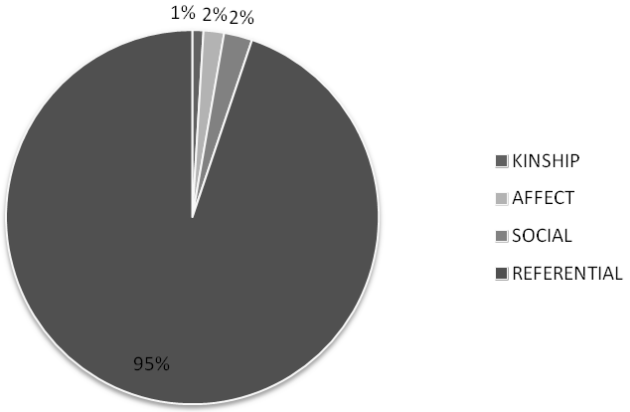

Fig. 2. Comparing families of relations

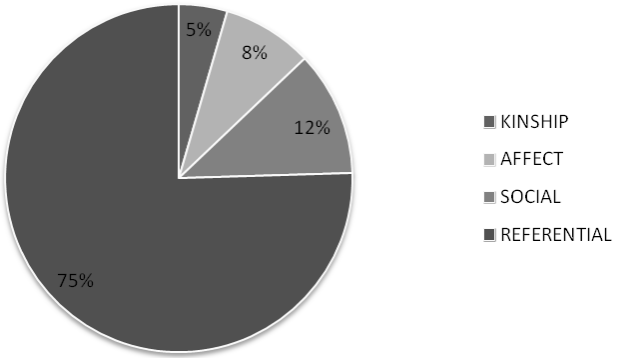

Fig. 3. Comparing families of relations (without `REFERENTIAL:coref` and `REFERENTIAL:coref-interpret`)

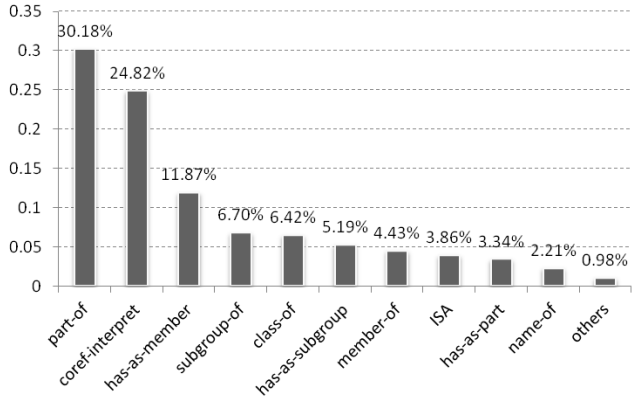

Fig. 4. Distribution of `REFERENTIAL` relations (without `REFERENTIAL:coref` and `REFERENTIAL:coref-interpret`)

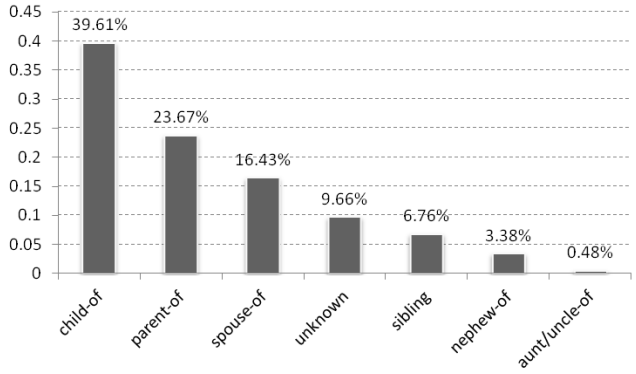

Fig. 5. Distribution of `KINSHIP` relations

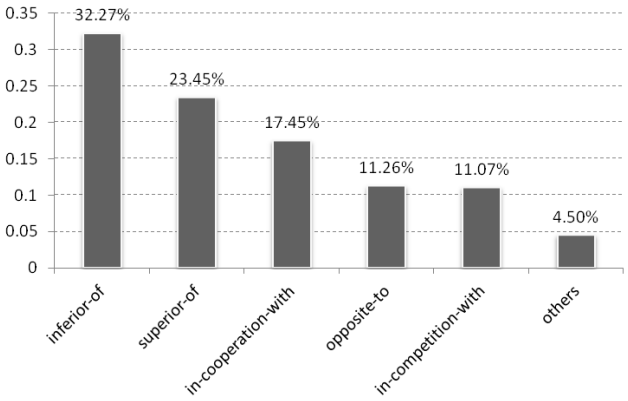

Fig. 6. Distribution of `SOCIAL` relations

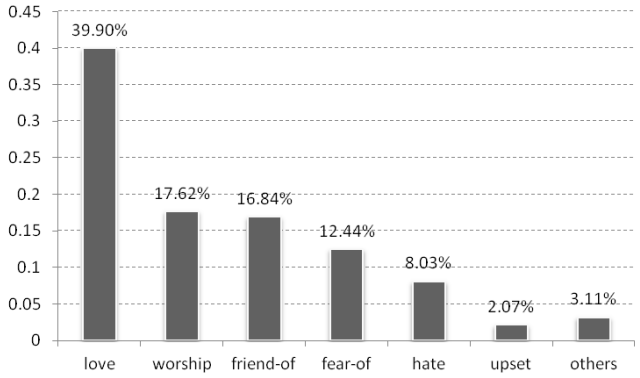

Fig. 7. Occurrences of `AFFECT` relations

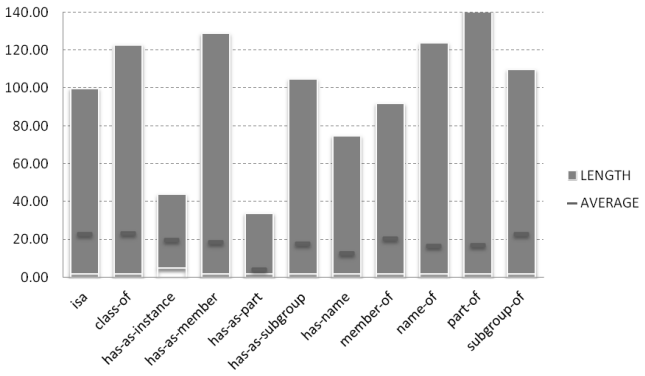

Fig. 8. Span length ranges and averages in number of words for `REFERENTIAL` relations (excepting `REFERENTIAL:coref` and `REFERENTIAL:coref-interpret`)

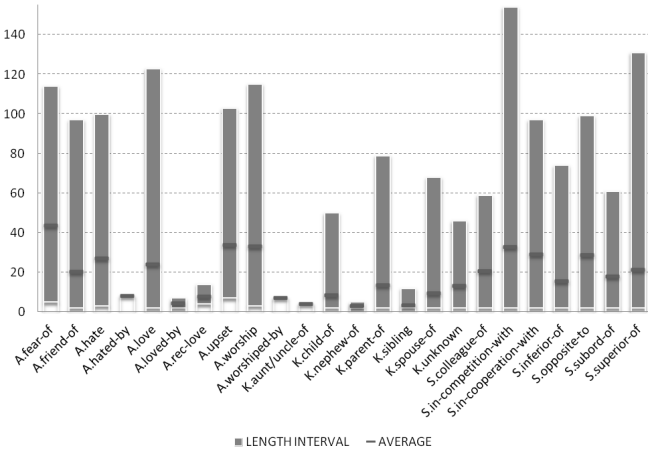

Fig. 9. Span length ranges and averages for `AKS` relations

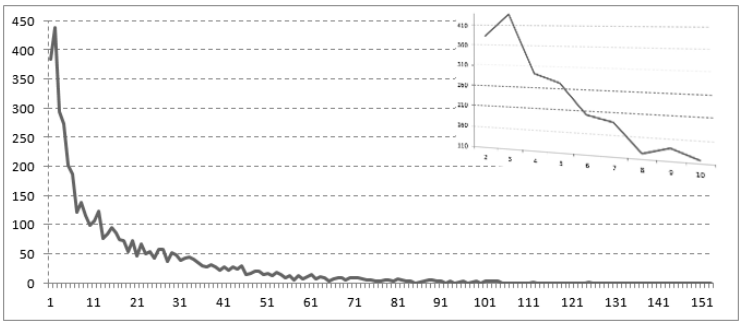

Fig. 10. Relations occurences (Oy) in correlation with the relation span length (Ox)

three families of relations. A rapid glance shows that KINSHIP relations are expressed over a shorter context than other types. This is mainly because many KINSHIPrelations are contextualised in noun-phrase expressions (*his mother*, *the son of X*, etc.).

To see how often appear in the corpus long spans with respect to short spans, Figure 10 shows the density of different lengths of relation spans (of course, excluding REFERENTIAL:coref and REFERENTIAL:coref-interpret). Its abrupt descending allure shows that short spans occur much frequently than long spans. There is a nose for length 3, indicating the most frequent span. The longest relation covers 154 words (but there is only one of this length)[23]. Supposing we would mark with $f(x)$ the function in Figure 10, the total number of words in the spans of the relations would be:

$$\sum_{x=2}^{154} xf(x) = 9,585$$

It correspondes with the total count of XML <W>...</W> markings under some relation span different than REFERENTIAL:coref and REFERENTIAL:coref-interpret. This means approximately 6.66% of the total area of the book, including punctuation (144,068 tokens).

An analysis of this kind is interesting because it reveals in approximate terms the proportion between positive and negative examples in an attempt to decipher automatically relations and could be of help when designing the sample set in a statistical approach to train from the corpus a recognition program.

Another set of statistics addresses the triggers. We are interested to know to what degree triggers are ambiguous. The sparse matrix in Table 2 allows an analysis of this type. On rows and columns all relations are placed and the number in a cell ($R1$,$R2$) indicates how many triggers, as (sequence of) lemmas, are common between the relations $R1$ and $R2$.

The last set of graphical representations are semantic graphs. Figures 11, 12 and 13 show sets of affection, family and social relations for some of the most representative characters in the novel.

Nodes in these graphs represent entities. Each one of them concentrates all coreferential links of one chain. Nodes names were formed by choosing the largest proper noun in each chain. When a character is mentioned with more different names, a concatenation of them was used (as is the case with *Ligia − Callina*). For the chains (usually small) that do not include proper nouns, one of the common nouns was used. When, doing so, more chains (nodes) got the same name, after verifying that the corresponding chains are indeed distinct, the names have been manually edited by appending digits.

[23] In this version of the corpus we did not make a thorough verification of long relations. We have noticed some errors in the annotation of poles, especially when one of the two poles are null pronouns in the position of subjects and REALISATION="INCLUDED" has not been marked on the respective verbs. In reality, a long distance coref relation would link the main verb (or its auxiliary) to an named entity, which now stands as one of the poles.

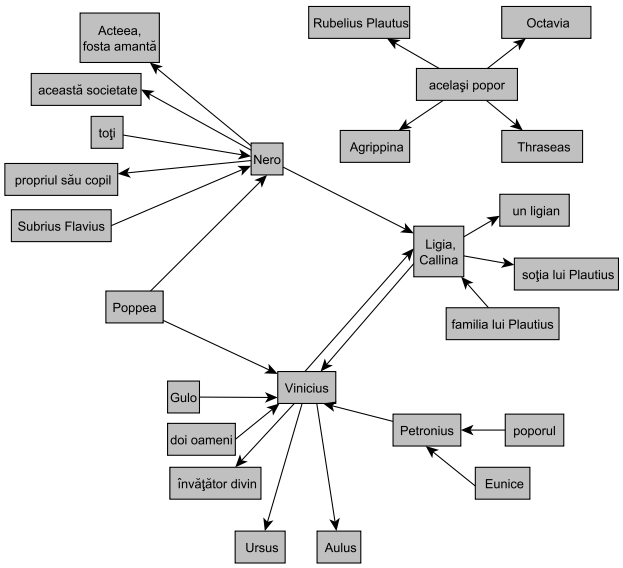

Fig. 11. A network of relations `AFFECT:love`

In Figure 11, for instance, can be read a `love` relation from Nero towards Acte, his child, his entourage (society) and, accidentally, Ligia. Also, there are reciprocal `love` relations linking Vinicius and Ligia, while Petronius is `love`d by Eunice and `love`s Vinicius.

Figure 12 concentrates both sets of relations `parent-of` and `child-of`, by reversing the sense of relations `child-of`. The family relations (not too many) expressed in the novel are now evident. Ligia-Callina has two fathers, the Phrygian king and Aulus Plautius.

Finally, Figure 12 reveals the `superior-of`− `inferior-of`pair of links (also by reversing the sense of relations `inferior-of`). Central in this graph is, as expected, the emperor Nero, socially superior to almost all characters in the novel. There is no edge pointing towards the node representing this character in the graph. Following him come: Vinicius (revealed as being superior to people of Rome, to the two slaves Demas and Croton, as well as to other servants, slaves and liberated slaves) and Petronius (linked to his servants and slaves, to pretorians, but also to his beloved Eunice). As expected, there is no superiority relation between the two close friends Petronius and Vinicius. In weaker relationships with the emperor Nero, Vinicius is not mentioned as his inferior, while Petronius, his cultural counsellor and praise-giver, he is. The only superior Vinicius seems to have over the whole novel is Corbulon, a former military chief.

Relation	#Triggers	#Common Triggers						
A.fear-of	27	S.in-coop-w:1	S.opp-to: 1					
A.friend-of	37	A.hate: 1	A.love: 2	A.loved-by: 2	A.wship: 1	S.col-of: 2	S.in-coop-w:3	S.inf-of: 1
A.hate	23	A.friend-of: 1	A.hated-by: 1	S.opp-to: 2				
A.hated-by	1	A.hate: 1						
A.love	64	A.friend-of:2	A.loved-by:3	A.rec-love:1	A.wship: 5	A.worsh-by:1	S.inf-of:1	
A.loved-by	3	A.friend-of: 2	A.love: 3	A.rec-love:1	A.wship: 1			
A.rec-love	5	A.love: 1	A.loved-by: 1	A.wship: 2				
A.upset-on	8							
A.wship	53	A.friend-of: 1	A.love: 5	A.loved-by: 1	A.rec-love:2	A.worsh-by: 2	S.inf-of: 3	S.sup-of: 2
A.worsh-by	3	A.love: 1	A.wship: 2					
K.child-of	20	K.parent-of: 7						
K.nephew-of	1							
K.parent-of	16	K.child-of: 7						
K.sibling	3	S.col-of: 1						
K.spouse-of	7							
K.unkn	8	S.col-of: 1						
S.col-of	7	A.friend-of: 2	K.sibling: 1	K.unkn: 1				
S.in-comp-w	50	A.fear-of: 1	S.in-coop-w:1	S.opp-to: 1				
S.in-coop-w	64	A.friend-of: 3	S.sup-of: 1	S.inf-of: 2	S.in-comp-w:1			
S.inf-of	69	A.friend-of: 1	A.love: 1	A.wship: 3	S.in-coop-w:1	S.sup-of: 9		
S.opp-to	41	A.fear-of: 1	A.hate: 2	S.in-comp-w:1				
S.sup-of	52	A.wship: 2	S.in-coop-w:1	S.inf-of: 12				

Table 2. The ambiguity of triggers

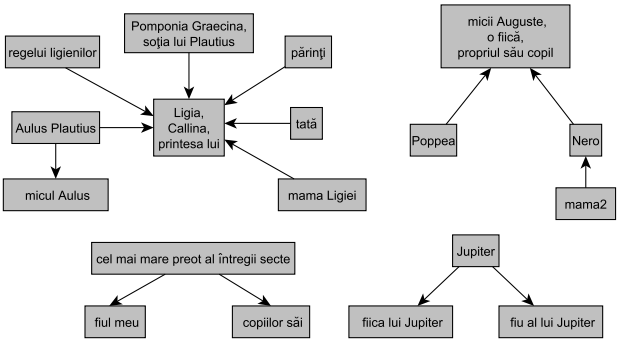

Fig. 12. A network of relations `KINSHIP:parent-of`

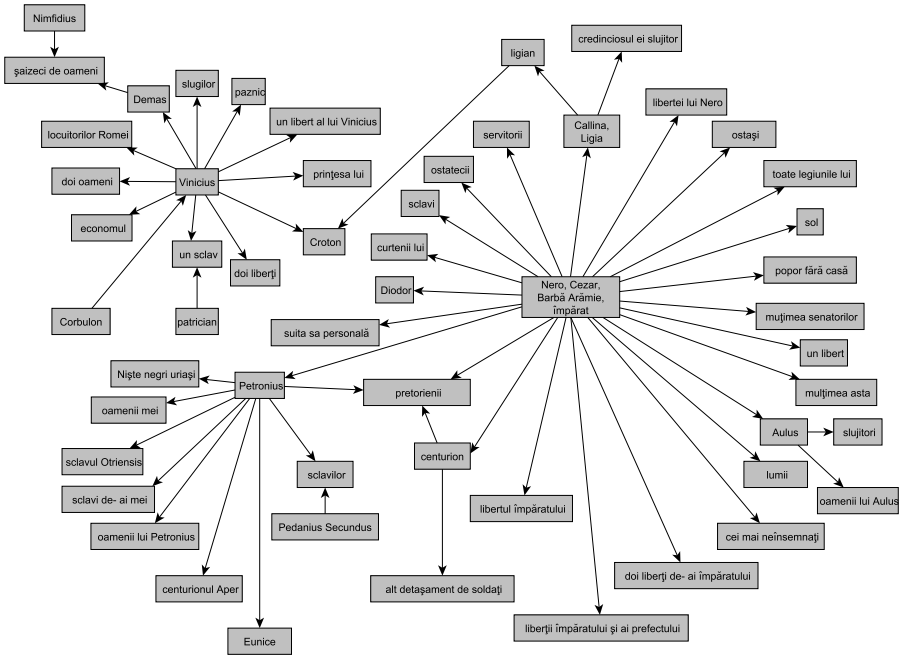

Fig. 13. A sub-network of relations `SOCIAL:superior-of`

As remarked in [32], "The importance of an entity in a semantic graph is determined not only by the number of relations the entity has, but also by the importance of the entities with which it is connected. So, for a character to be influential in the novel it is not enough to have many relations, but to be related with influential characters too. The PageRank [5] could be applied to measure the centrality/influence of an entity according to its position in the graph". Such estimations are yet to be made in a further research, but even only a simple visual inspection of our graphs puts in evidence the central characters: Vinicius, Petronius, Nero, Ligia. Let's note also that all these graphs display only once the sets of homonymous relations. More sophisticated representations, showing also the number of relations, not only their types, could put in evidence with more clarity the annotations in the corpus. Moreover, chains of social links could also evidence hierarchical positions in the society (as, for instance, the one connecting in `superior-of`relations the characters Nero, Ligia, a ligian and Croton). Combining graph relations could also evidence complex situations or plot developments, as for instance the distinction between a family type of affection (between Ligia and Plautius's wife, for instance, Plautius's wife being a parent for Ligia) and lovers (the sentiment that Vinicius develops versus Ligia and vice-versa, neither of these doubled by any kinship relation).

The examples put forth are bits of complex interpretations. They reveal that the detection of semantic relations could incumber complex reasoning steps, thus including germs for a true understanding of the semantic content of a big coherent text.

7 Conclusions and further work

The research aims to formalize relationships between characters of a novel, thus establishing precise criteria that underpin aspects of the interpretation of text. The annotations that we propose can be considered as representation bricks in a project pertaining to the interpretation of free texts. Since the world described in our corpus is a fictional one, free of any constraints, we believe that the representation of entities and their links made explicit in the annotation constitute pieces of knowledge that can be extended to other universes or world structures with minimum adaptations.

The enterprise of building the "Quo Vadis" corpus was an extremely time consuming and difficult one, necessitating iterative refinements of the annotation conventions, followed by multiple corrections, sewing and merging steps. We considered that describing this process could be interesting per se, as a model to apply in building other corpora, when the scarcity of resources do not permit passing the manual work under the critical eyes of more subjects. Instead, an iterative improvement methodology was applied, by designing syntactical and semantic filters, running them and correcting the reported errors.

The corpus is still too fresh to risk a detailed numerical report and interpretation on it. In the next few months it may still undergo further improvements[24]. The graphics and tables we presented should, therefore, be interpreted more qualitatively than quantitatively, e.g. in terms of the rates between different types of relations. They allow to sketch a perception about the density of person and god types of entities and the relations mentioned among them in a literary freestyle text. Of course, from genre to genre, style to style and document to document the densities and rates may vary dramatically, but, we believe, proportions will remain within the same orders of magnitude.

The corpus is intended to be put at the base of a number of investigations in the area of semantic links, mainly oriented towards their automatic identification. For sophisticating the features to be used in the process of training statistical relation recognition programs, other layers of annotation could be useful, the most evident one being the syntactic layer, for instance, dependency links. Then, on top of the annotations already included, other types of entities and relations could be further added. Examples of sophistications include: notation of places and relations between people and places, or between places and places. Such markings could put in evidence descriptions of journeys in travelling guides, or geographical relations in high school manual. Of a different kind, extensively studied (see [25] for a survey), are the temporal relations.

Of a certain interest could be the issue of exporting annotations between parallel texts. For instance, from the Romanian version of "Quo Vadis" to its English or Polish version. If this proves possible, then a lot of time and money could be saved.

In the process of deep understanding of texts, on top of discovering inter-human or human-god relationships could be placed superior levels of interpretation, as, for instance, deciphering groups manifesting a distinctive, stable and cohesive social behaviour (as is, in the novel, the group of Romans and that of Christians). If time is added to the interpretation, then developments of stories could be traced as well. Sometimes individuals migrate from one group to the other (see Vinicius) and the range of sentiments and social relations might change (Vinicius to Christ: from lack of interest to worship and Vinicius to Nero: from `inferior-of` to insubordination). On another hand, a society, as a whole, can be characterised by the set of inter-individual relationships and interesting contrasts could be determined. The new society of Christians, with their affective relations of `love` and `worship`, regenerate the old and decadent society of Romans, especially that cultivated at the court of the emperor. The `inferior-of`, `fear` and `hate` relations, frequent between slaves and their masters are replaced by `in-cooperation-with`, `friendship` and `love`, characteristic to the Christian model.

The corpus includes only explicitly evidenced relations (what the text says), but in many cases a human reader deduces relations on a second or deeper level

[24] One of the authors is elaborating a personal dissertation thesis (due June 2014) having as theme this corpus, being responsible for its correctness and complete statistics over it.

of inference. Moreover, some relations explicitly stated are false, insincere, as for instance the declared love or worship sentiments of some underdogs with respect to the emperor. To deduce the falsity of relations, it could mean, for instance, to recognise relations of an opposite type, stated in different contexts and towards different listeners by the same characters. All these could be subjects of further investigation, but to do such complicated things one should start by doing simple things first, as is the automatic discovery of clearly stated relations, such as those annotated in the "Quo Vadis" Corpus.

Acknowledgments. We are grateful to the master students in Computational Linguistics from the "Alexandru Ioan Cuza" University of Iaşi, Faculty of Computer Science, who, along three consecutive terms, have annotated and then corrected large segments of the "Quo Vadis" corpus. Part of the work in the construction of this corpus was done in relation with COROLA - The Computational Representational Corpus of Contemporary Romanian, a joint project of the Institute for Computer Science in Iaşi and the Research Institute for Artificial Intelligence in Bucharest, under the auspices of the Romanian Academy.

References

1. Anechitei, D., Cristea, D., Dimosthenis, I., Ignat, E., Karagiozov, D., Koeva, S., Kopeć, M., Vertan, C.: Summarizing Short Texts Through a Discourse-Centered Approach in a Multilingual Context. In: Neustein, A., Markowitz, J.A. (eds.), Where Humans Meet Machines: Innovative Solutions to Knotty Natural Language Problems. Springer Verlag, Heidelberg/New York (2013)
2. Bagga, A., Balwdin, B.:Entity-based cross-document coreferencing using the Vector Space Model. In: Proceedings of COLING '98, 1 (1998)
3. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the Web. In: Proceedings of IJCAI '07 (2007)
4. Bejan, C. A., Harabagiu, S.: Unsupervised Event Coreference Resolution with Rich Linguistic Features. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden (2010)
5. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN systems, 30(1), 107–117 (1998)
6. Boschee, E., Weischedel, R., Zamanian, A.:Automatic information extraction. In: Proceedings of the 2005 International Conference on Intelligence Analysis, McLean, VA, 2-4 (2005)
7. Bunescu, R.C., Paşca, M.: Using encyclopedic knowledge for named entity disambiguation. In: European Chapter of the Assocation for Computational Linguistics (EACL 2006).
8. Carlson, A., Betteridge, J., Wang, R. C., Hruschka Jr., E. R., Mitchell, T. M.: Coupled semi-supervised learning for information extraction. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010)
9. Chen, B., Su, J., Pan, S. J., Chew L. T.: A Unified Event Coreference Resolution by Integrating Multiple Resolvers. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, pp. 102–110, Chiang Mai, Thailand (2011)

10. Cristea, D., Dima, G.E.: An integrating framework for anaphora resolution. Information Science and Technology, Romanian Academy Publishing House, Bucharest, 4(3-4), 273–291 (2001)
11. Cruse, D.A.: Lexical semantics. Cambridge University Press, Cambridge, UK (1986)
12. Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data. In: Empirical Methods in Natural Language Processing (EMNLP) (2007)
13. Cybulska, A., Vossen, P.: Using Semantic Relations to Solve Event Coreference in Text. In: Proceedings of Semantic Relations-II. Enhancing Resources and Applications Workshop, Istanbul (2012)
14. Del Gaudio, R.: Automatic Extraction of Definitions. Ph.D. thesis. University of Lisbon (2014).
15. Drabek, R. Yarowsky, D.: Induction of Fine-Grained Part-of-Speech Taggers via Classifier Combination and Crosslingual Projection. Proceedings of the ACL Workshop on Building And Using Parallel Texts: Data-Driven Machine Translation And Beyond, June 29-30, 2005, Ann Arbor, Michigan, p. 49-56 (2005)
16. Gala, N., Rey, V, Zock, M.: A tool for linking stems and conceptual fragments to enhance word access. In Proceedings of LREC-2010, Malta (2010).
17. Girju, R., Badulescu, A., Moldovan, D.: Automatic discovery of part-whole relations. Computational Linguistics, 32(1) (2006)
18. Hearst, M.:Automatic acquisition of hyponyms from large text corpora. In: Proceedings of COLING '92 (1992)
19. Hjørland, B.: Semantics and Knowledge Organization. Annual Review of Information Science and Technology, 41, 367–405 (2007)
20. Iida, R., Komachi, M., Inui, K., Matsumoto, Y.: Annotating a Japanese text corpus with predicate-argument and coreference relations. In: Proceedings of the Linguistic Annotation Workshop, pp. 132-139 (2007)
21. Kawahara, D., Kurohashi, S., Hasida, K.: Construction of a Japanese relevance-tagged corpus. In: Proceedings of LREC '02 (2002)
22. Levi, J. N.: The Syntax and Semantics of Complex Nominals. Academic Press, New York (1978)
23. Lyons, J.: Semantics. Cambridge University Press, Cambridge, UK (1977)
24. Malmkjær, K.: Semantics. In: Malmkjær, K. (ed.) The linguistics encyclopedia, pp. 389–398. London: Routlage (1995)
25. Mani, I., Wellner, B., Verhagen, M., Lee, C.M., Pustejovsky, J.: Machine Learning of Temporal Relation. In: Proceedings of the 44th Annual meeting of the Association for Computational Linguistics, Australia (2006)
26. Masatsugu, H., Kawahara, D., Kurohashi, S. Building a Diverse Document Leads Corpus Annotated with Semantic Relations, in Proceedings of the 26th Pacific Asia Conference on Language,Information and Computation, pages 535544 (2012)
27. Mazlack, L. 2004. Granular causality speculations. IEEE Annual Meeting of the Fuzzy Information, 2004. Processing NAFIPS 04. 690695.
28. Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J.: Introduction to WordNet: An On-Line Lexical Database. International Journal of Lexicography, Vol. 3, No. 4 (winter 1990), pp. 235-244 (1990).
29. Mitkov, R.: Anaphora Resolution, in R.Mitkov (ed.) The Oxford Handbook of Computational Linguistics, Oxford University Press, 266–283 (2003)
30. Mulkar-Mehta, R., Hobbs, J. R., Hovy, E.: Granularity in Natural Language Discourse. In: Proceedings of International Conference on Computational Semantics (2011)

31. Murphy, M.L.: Semantic relations and the lexicon: antonymy, synonymy, and other paradigms. Cambridge University Press, Cambridge, UK (2003)
32. Năstase, V., Nakov, P., Ó. Séaghdha, D., Szpakowicz, S.: Semantic Relations Between Nominals. Morgan & Claypool Publishers, California (USA) (2013)
33. Ohara, K.: Full text annotation with Japanese framenet: Study to annotation semantic frame to bccwj (in japanese). In: Proceedings of the 17th Annual Meeting fo the Association for Natural Language Processing, pp. 703-704 (2011)
34. Pantel, P., Ravichandran, D., Hovy, E.: Towards terascale knowledge acquisition. In: Proceedings of COLING '04 (2004)
35. Paşca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A.: Names and similarities on the Web: Fact extraction in the fast lane. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 809-816, Sydney, Australia (2006)
36. Postolache, O., Cristea, D., Orasan, C. Transferring Coreference Chains through Word Alignment. In: Proceedings of LREC-2006, Geneva (2006)
37. Quillian, M., R.: A revised design for an understanding machine. In: Mechanical Translation, 7, 17 – 29 (1962)
38. Rao, D., McNamee, P., Dredze, M.:Entity Linking: Finding Extracted Entities in a Knowledge Base. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) Multisource, Multilingual Information Extraction and Summarization, Springer Lecture Notes in Computer Science, Berlin, Heidelberg (2012)
39. Rello, L., Ilisei, I.: A comparative study of Spanish zero pronoun distribution. In Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL), pp. 209–214 (2009)
40. Rodríguez, K. J., Delogu, F., Versley, Y., Stemle, E. W. and Poesio, M.: Anaphoric annotation of Wikipedia and blogs in the live memories corpus. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC '10) (2010)
41. Rosenfeld, B., Feldman, R.: Using corpus statistics on entities to improve semisupervised relation extraction from the Web. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 600-607, Prague, Czech Republic (2007)
42. Pollard, C., Sag, I.A.: Head-Driven Phrase Structure Grammar, University of Chicago Press (1994)
43. Saggion, H.: SHEF - semantic tagging and summarization techniques applied to cross-document coreference. In: Proceedings of SEMEVLA '07 (2007)
44. Séaghdha, D., Ó., Copestake, A.: Semantic classification with distributional kernels. In: Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08), Manchester, UK (2008)
45. Singh, S., Subramanya, A., Pereira, F., McCallum, A.: Large-scale cross-document coreference using distributed inference and hierarchical models. In: Proceedings of HLT '11, 1 (2011)
46. Simionescu, R.: Romanian deep noun phrase chunking using graphical grammar studio. In: Moruz, M. A., Cristea, D., Tufiş, D., Iftene, A., Teodorescu, H. N. (eds.) Proceedings of the 8th International Conference "Linguistic Resources And Tools For Processing Of The Romanian Language", 135–143 (2012)
47. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic taxonomy induction from heterogeneous evidence. In: Proceedings of COLING-ACL '06 (2006)
48. Tanaka, I.: The Value of an Annotated Corpus in the Investigation of Anaphoric Pronouns, with Particular Reference to Backwards Anaphora in English. Ph. thesis, University of Lancaster (1999)

49. Tesnière, L.: Éléments de syntaxe structurale. Klincksieck, Paris (1959)
50. Zock, M.: Wheels for the mind of the language producer: microscopes, macroscopes, semantic maps and a good compass. In Barbu Mititelu, V., Pekar, V., Barbu, E. (eds.) Proceedings of the Workshop Semantic Relations. Theory and Applications (2010).
51. Zock, M., Ferret, O., Schwab, D.: Deliberate word access: an intuition, a roadmap and some preliminary empirical results. Int J Speech Technol 13, pp. 201-218 (2010).
52. Zock, M. et Schwab, D.: L'index, une ressource vitale pour guider les auteurs a trouver le mot bloque sur le bout de la langue. In Gala, N. et M. Zock (eds). Ressources lexicales: construction et utilisation. Lingvisticae Investigationes, John Benjamins, Amsterdam, The Netherlands (2013).