Towards new ways of looking at texts

Andreea Macovei[1]

[1]Faculty of Computer Science "Al.I.Cuza" University Iasi, Romania
andreea.gagea@info.uaic.ro

Abstract. This paper provides an overview of my PhD thesis which in a first phase, focuses on developing an ontology which exposes the various text types, both literary and non-literary in Romanian language and in a second phase (the most complex one), on formalizing a temporal annotation scheme that can be used in order to reorder the events or more precisely, the sequences of events in novels and also, the switches that may occur in literary texts such as flashbacks, flashforwards, embedded fabulae, temporal ruptures, and transitions. The classification of texts in species and genres proposed by this ontology can be used in to create suggestions for readers, to identify a certain type of text, to extract relevant information, to analyse the format differences between two different types of texts (as those between the normative texts and the news) as a first step in automatic text writing techniques, etc., while the temporal model considers a natural order of events that the reader seems to perceive once she or he continues to read or, at the end of the book despite timelines (the representation of all the events chronologically exposed in a story) and storylines (the main story or plot of a literary text) in order to capture the actions of a character and his or her chronological evolution (time track) throughout a book.

Keywords: Classification of Texts, Ontology, Temporality, Temporal Annotation

1 Introduction

The issue of genre identification can be considered as a task of resolving the problem of text classification: as for each type of text, the genre is determined together with the other species and several specific features, the texts will be classified according to a predefined list of genres. Unfortunately, a predefined list of genres does not exist as there are numerous disagreements regarding the definition of specific genres: a genre for a theorist could be a sub-genre or even a super-genre for other theorists [1].

Observing the particularities of legal language as opposed to literary works or news articles, legal texts have a specific format, contain specialized words and are (or at least should be) completely objective. Those particularities of texts are very important in order to provide a delimitation of text types and a genre classification. Clues as impersonal character of the author, format of a text, mood of message transmission, morphological marks as *imperative mood of verbs, impersonal verbs, vocative case of*

nouns, spatial and temporal indications, etc. can be used to draw a distinction between different types of texts.

Starting from these observations, the idea of text ontology was outlined. For the representation of this genre ontology, two main text types are considered: literary genre and non-literary genre, to which a special category was added: social media texts. For each genre, we have established several sub-genres which contain species and each species is represented by a certain number of examples and features mentioned below.

Thus, we continued to work on the particularities of these types texts included in the ontology, more exactly on temporality. The variety of ways in which temporality is expressed in a literary text is extremely impressive and capturing all the events in order to outline the destiny of a character may pose challenging problems.

The interpretation of time in texts may refer to the recognition of temporal expressions, of events and their relative order, of timelines and storylines. However, literary (or belletristic) texts are quite different in perceiving time than non-literary texts such as news texts.

Although there is no rule, in novels, an ongoing story can be interrupted by a flashback and this can lead to some important changes: a possible changed perspective from which the events are told, some discontinuities in action or even, development of disparate stories without any apparent link between them, followed by surprising encounters of characters and merging of destinies; also, characters that have followed for a while common or parallel lives and at a precise moment, may split and so, their stories continue separately.

This means that despite timelines (the representation of all the events chronologically exposed in a story), there is a natural order of events that the reader seems to perceive once he continues to read or, even, at the end of the book. So, a first step could concern the identification of this order of events using our model called *the model of rail yard*: certainly, this order will lead to a better understanding of the entire story.

2 State of the art

Classifying texts is an important challenge of natural language processing field: as there are many texts published and shared every second (*literary works, newspapers, blogs, news, laws, etc.*), a possible sorting and classification of those texts is extremely necessary. Starting with Aristotle [2] and finishing with the literary theorists of these days, this process of text type classification continues to be the subject of a non-finite taxonomy.

In computer science, an ontology is a model that uses a shared vocabulary in order to define the type of concepts, their features and relations of the existing world. Ontology provides the possibility to share common understanding of the information structure among people [3]: the information that can be found on specialized websites (medical websites, technical websites, e-commerce websites) can be extracted and

aggregated and specialists (agents) can reuse it for other applications of the same domain (e.g. question answering systems).

As one of the features found in the texts included in our ontology refer to spatial and temporal indications, we continue to focus on temporal annotation of literary texts. For the temporal approach, we introduce a new computational model with an annotation scheme for temporality that may determine a different way to look at literary texts in terms of chronological reorder of events.

Recent studies consider the challenges of temporal tagging in different domains: the temporal annotation of narrative-style documents [4] such as Wikipedia articles about wars and the extraction of temporal information from a raw text [5]. Temporal annotation techniques of narratives that are rich in temporal expressions are different than the techniques used for news texts [6].

The concept of storyline [7] refers to groups of interacting timelines, or mergers of two or more timelines where the same characters or entities are taking part in the action. This happens because in the authors' view, storylines describe a kind of text structure which is not related to the flow of time; there are events ordered and anchored in timelines, but the overall relation of the storyline with the time axis seems to not exist.

Timelines get to be computed because they are sequences of event instances anchored to time expressions or relative to each other. A storyline consists of sets of events, their corresponding time points and relations that order them, while a timeline is made of those events which have the highest connectivity score.

Another computational model of storyline and timeline [8] exposes the basic elements of a storyline which are the events (defined by actors, locations and time settings), the anchoring of events to time, a basic timeline (known as fabula, a sequence of chronologically ordered and logically connected events involving one or more actors).

In this case, the study aims to detect storylines in massive streams of news, while deciphering the temporal structure of texts is less highlighted. There are also other systems that try to extract storylines or to reconstruct maps of connections in order to capture story development in an explicit manner [9] or a complex representation [10].

Ordering events may cause some problems when it comes to belletristic texts; trying to find a computational representation of temporal phenomena occurring in literary texts with the aim of reordering the cursive thread of the action seems to be a milestone quite difficult to reach. Also, this research stage aims to formalize a number of textual phenomena which seem not having received much attention before.

3 Research methodology

3.1 Ontology of Romanian texts

The proposed ontology of text types can be used to classify different types of texts as the large amount of texts belonging to different genres for each language becomes more useful if it is organized. If the types of texts are determined and their features and examples are used as indicators for automatically delimitating other similar types

of content, the ontology can become a useful tool for developing natural language processing applications.

Going beyond the classification of texts in species and genres, an ontology which exposes the various text types, both literary and non-literary, and offers a set of features to describe them could be used to automatically identify different types of texts, give suggestions in a reading recommendation system, or developing techniques of automatic writing and summarization. Our ontology can represent a first step for:

- implementing applications that propose suggestions for readers;
- collecting domain corpora in an automatic manner;
- establishing a clear and precise text typology can help specialists to determine the particularities and similarities of a type of text (these observations can be used in order to automatically identify a certain text);
- identifying a certain type of text or analysing the format differences between different types of texts;
- providing automatic text writing techniques.

Theoretical studies of Romanian literature refer especially to one text genre: literary genre, known also as fiction literary genre. There are no specifications about other genres in Romanian literature although literary critics present the idea of non-literature without providing a precise classification of species that could belong to this category. But, there is an actual and clear distinction between literary texts and non-literary texts in Romanian:

a. A literary text is a text by which the author wants to impress and to enthuse readers, expressing his or her own thoughts, ideas and feelings, using an artistic language, strongly marked by subjectivity.
b. The non-literary text is a text which aims to objectively inform the reader about certain aspects of reality in a clear and precise language, often containing scientific and technical terms.

The class of literary texts includes the literary works, divided into four main categories: epic genre, lyrical genre, drama genre and argumentative genre, each genre having its own species, while the non-literary class of texts, five categories: informative genre, instructive genre, persuasive genre, juridical genre and descriptive genre, to which a special category was added: social media texts. Table 1 covers all the genres and times of texts included in the ontology.

Table 1. Species of literary and non-literary texts included in the ontology.

Ontology of text genres			
Literary class	**Text types**	**Non-literary class**	**Text types**
Epic genre	diary, biography, autobiography, memories, epic, novel, sketch story, short story, novella, fairytale, myth, parable, ballad, fable, poem, anecdote;	**Informative genre**	weather reports, reportage, press article, news, reviews, scientific texts;

Lyric genre	elegy, ode, pastel, meditative poetry, satire, pamphlet, sonnet, rondo, ghazel, gloss, romance, hymn, doina, idyll, haiku;	**Instructive genre**	manuals, recipes, how-to guides texts;
Drama genre	drama, comedy, tragicomedy, tragedy, theatre of the absurd;	**Persuasive genre**	offers, advertisements texts;
Argumentative genre	argumentative texts;	**Descriptive genre**	travel guide, scientific and descriptive texts (brochures);
		Juridical genre	contracts and dispositions, law texts, regulations;
		Social media content	social media texts, personal data, blogs, short reviews, tweets.

A set of properties was developed to characterize the text types in our ontology. Each text type is given specific values of the considered set of properties. For instance, we have the commercial feature of several text types, such as offers and advertisements. Their commercial character is identified through the analysis of the structure of the texts, but also from morphologic and syntactic information (first person of the verbs, exclamations and indications as address or telephone number). The set of properties refer to the functional style, the presence/absence of spatial and temporal indications, the mood of the transmitted message, presence/absence of figures of style, specialized or general words, presence of a dominant entity (more precisely, if the texts are centered on a single entity or not), the length of the text, prosody, commercial style etc.

Such an ontology can also be used to create suggestions for readers, to identify a certain type of text or to analyze the format differences between two similar types of texts (as offers and advertisements, for instance), to extract information, to analyse the content of the text, to offer suggestions, to summarize a text, to identify a text according to its format etc.

3.2 The Rail Yard Model

After the study stage of ontology, we continue to work on different aspects of temporality found in the novels, a feature also mentioned in our ontology. Our research proposes an annotation approach known as the Rail Yard model that introduces a new

way to look at literary texts in terms of chronological reorder of events. This name of the model can be explained by the fact that in a book, the time can be represented as threads that may interlace and separate and this can be compared with railways in complex rail yards. In a literary text, the probability that the entire action is interrupted by flashbacks, memories or even by a new other narrated story is higher than in a non-literary text.

This model includes time tracks; these time tracks display the story development of one character or a stable group of characters over a period of story time throughout the novel. Also, the characters involved in action are the masterpiece of this scheme along with the spatial indications. So, when time and place change, when a character appears or disappears, the story follows a new direction.

Time tracks are made out of one or more time segments. Also, these time tracks have start points, end points, join points and split points that reveal where they start and finish, spilt and respectively merge.

Time tracks is be bordered by: a start point (there is always a beginning), an end point (where a time track finishes), join points (two different time tracks unify when two characters meet and continue to live together forever or for a while) and split points (where one time track splits into two separate time tracks; one character may disappear or two or more actors split).

Figure 1 shows the time track representation of a character in the novel of Tash Aw, Map of the Invisible World; on the one hand, there is the exact chaining of event sequences involving Adam, a main character and the other characters with whom he interacts throughout the text, and on the other hand, the same chaining is represented after reordering the entire discourse for better anchoring on a time axis (with a start point, an end point, two join points and one split point).

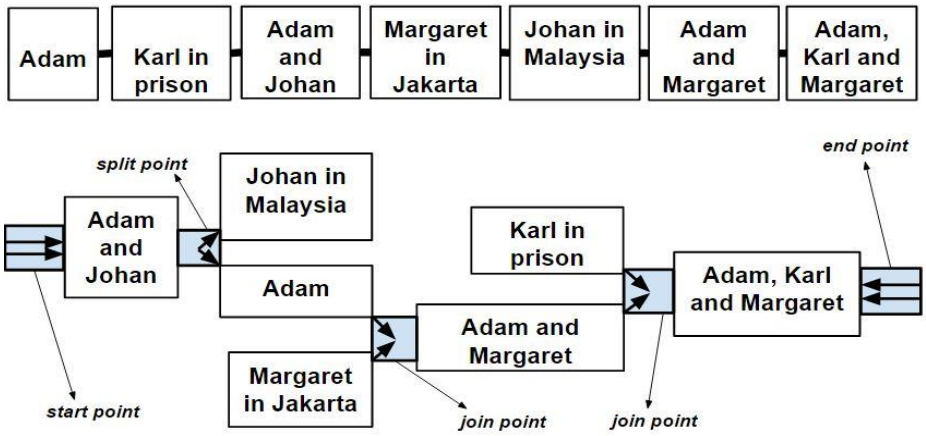

Figure 1. Example of time segments and time track representation (focussed on the main character Adam)

The first line of boxes in this figure shows the time segments as they are unfold on the book, therefore aligned with text offsets, while the bottom scheme shows their reshaping in time tracks. This annotation scheme of time tracks considers the charac-

ters involved in the action (the main actors with or without a specific name). The appearance or disappearance of a character in a sequence of events (time segments) determines a new sequence. So, the characters are key points in delimitating the sequences of the novel.

The start point is represented by the first time segment which starts the evolution of one character, and the end point ends his or her evolution through the book. Although they are noticeable much later by the reader because the narrator does not always mark them, the join and split points are also relevant as they may determine the appearance of a new time track (this is the case of a join point; two different time tracks unify when two characters meet and continue to live together forever or for a while), or of two new time tracks (a split point where one time track splits into two separate time tracks; one character may disappear or two or more actors split).

We have established a classification of time tracks: NAR is for typical narrations in which the story time flows constantly ahead, REM for flashbacks or remembers belonging to a character, SUP for suppositions or speculations, GEN for general knowledge where there is no time anchor, only statements about generally accepted things, and FIC for fictions, invented realities, like in movies, plays or novels.

Time tracks are composed of time segments (a sequence of events of one character or a stable group of characters over a period of story time that is uninterruptedly told in a span of text), temporal relations between time segments and time actors (explicitly expressed in text or inferred), the characters that are involved physically or virtually in the development and time location (the location of the development). We consider that the appearance or disappearance of a character in a sequence of events (time segments) determines a new sequence. So, the characters are key points in delimitating the sequences of events.

The temporal relations between time segments that we have established are BEFORE, IMMEDIATELY-BEFORE, AFTER, IMMEDIATELY-AFTER, and SIMULTANEOUS. BEFORE and AFTER relations are distant in time, while IMMEDIATELY-BEFORE and IMMEDIATELY-AFTER are established when two segments are placed one after or one before another. SIMULTANEOUS relations highlight two parallel sequences of events or even two connected sequences of events which are in progress.

The selected corpus for our study is a novel by Tash Aw (*Map of the Invisible Word*), Romanian and English versions, which is known for repeated returns in the past and through narratives presented from several perspectives (the narrator's perspective and the perspective of two main characters, the brothers Adam and Johan). For the annotation of novel, we developed our own annotation tool (Time Tagger) that can be easily configured with elements of the proposed scheme.

After the annotation step, we wanted to create a visualization tool that may catch the real order of the adventures of one or more characters in a book. This tool is intended to provide a visual representation of all the stories and substories covered by a narrator, a character or an author (for example, a time sequence or a narrative thread).

In **Figure 2**, we have an example of such a representation (several time segments with Adam as their protagonist). The time segments are displayed with their specific IDs in the chronological order: time segments on the same line are sequences of

events that immediately occur one after another or one before another, those in a descending flow are events sequences that occur one after another or one before another but not immediately, and time segments displayed one below other are parallel sequences of events or even two connected sequences of events which are in progress (they occur simultaneously).

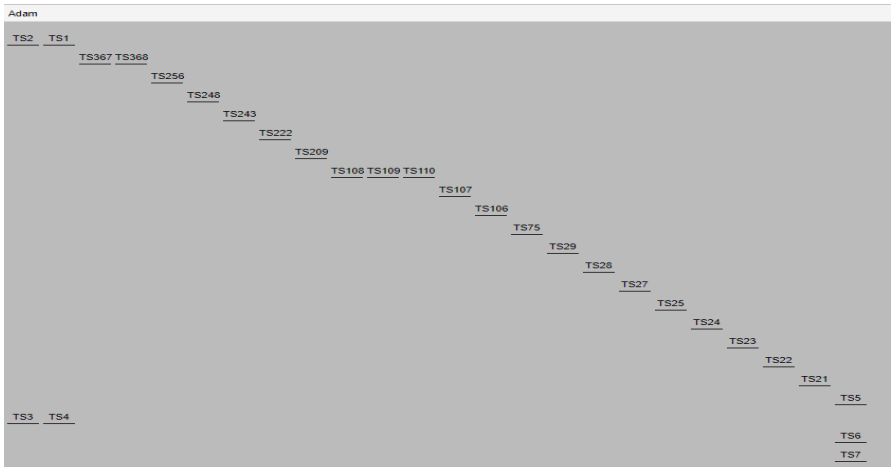

Figure 2. Representation of time segments extracted from the visualization program

Although our scheme of annotation is different than the annotation scheme proposed by the TimeML standard, our intention was to continue with an alignment process of time segments to TimeML. Thus, we have started to search for resources and tools that may help us complete our corpus to supplementary levels of annotation. At this moment, we are in an experimental stage in which we use the Tarsqi Toolkit, a set of processing components for extracting temporal information to see if we can extract time expressions, events, subordination links and temporal links from our corpus although the instruments offered by the toolkit are based on news texts.

4 Future work

The ontology of text types is a useful instrument to classify different types of texts. Starting with the proposed text types, we built an ontology by identifying different properties and relationships between these text classes (we established a set of 13 structural, lexical and morphological, stylistic and semantic features which we tested on different text types). Based on our representation of ontology, we continue to work on the visualization tool based on adaptive canvas for a web interface of the ontology with examples and results because visualising the proposed classification of texts could be very help for both specialists and readers.

And regarding our temporal model, we consider that this research presents a new scheme of temporal annotation having as a start point the conventions of TimeML

standard: it is a new way of looking at a text with elements of time analysis and text structure.

Annotating literary texts is extremely challenging as the authors and the narrators enjoy complete freedom in order to indicate a storyline to the readers; the lack of temporal indications, the presence of flashbacks, the temporal ruptures, the embedded substories can lead the action thread in a completely different time direction from the exposed order in a book.

The possibility to represent time tracks as graphs or diaphragms can contribute to a modality to restructure the story and this is a significant step in deciphering the structure of the discourse. It is a modality to make reading more interactive and to capture the attention of readers until the end of the novel (or any other literary text).

Once we implement the visualization tool of time tracks, such an instrument will make reading an interactive activity: so, the reader will not have to return to various passages in the book as long as she or he has a handy instrument that can show her or him the background of a character, the climax of a story, the present time of a story etc.

So, if we get to build the time yard diagram of a novel and this way, to recuperate the time tracks out of the sequence of time segments, we will have the certitude that such a model is deterministic, therefore can be algorithmically approached. If this will not be the case, it means that the information that we included in the XML schema of time segments is incomplete and should be revised.

Considering all the information collected through the annotation process, we can try to see it is possible to identify the place (or route) unit, the unit of characters, the time unit, to recognize the type of a temporal segment according to our classification (the typology of NAR, REM, SUP, GEN and FIC segments), to implement a tool for an automated summary on segments and to extract character narrative lines as sequences of temporal segments (time tracks) in which some characters are involved. This research can lead to the extraction of semantic content from literary texts: a sophisticated tool capable of answering questions about the content of a text.

References

1. Chandler, D.: An introduction to genre theory, Media and Communication Studies: http://www. aber. ac. uk/media/Documents/intgenre/intgenre1.html (1997).
2. Guarino, N., Oberle, D., Staab, S.: What is an Ontology? Handbook on ontologies. Springer Berlin Heidelberg, 2009. 1-17.
3. Gruber, T. R.: A translation approach to portable ontology specifications, Knowledge acquisition 5.2 (1993): 199-220.
4. Strötgen, J., Gertz, M.: Multilingual and cross-domain temporal tagging. Language Resources and Evaluation, 47(2), 269-298 (2013).
5. UzZaman, N., Allen, J. F.: Event and temporal expression extraction from raw text: First step towards a temporally aware system. International Journal of Semantic Computing, 4(04), 487-508 (2010).
6. Mazur, P., Dale, R.: Wikiwars: A new corpus for research on temporal expressions. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 913-922). Association for Computational Linguistics.

7. Laparra, E., Aldabe, I., Rigau, G.: From TimeLines to StoryLines: A preliminary proposal for evaluating narratives. ACL-IJCNLP 2015, 50.
8. Vossen, P., Caselli, T. and Kontzopoulou, Y.: Storylines for structuring massive streams of news. ACL-IJCNLP 2015, 40.
9. Shahaf, D., Yang, J., Suen, C., Jacobs, J., Wang, H., Leskovec, J.: Information cartography: creating zoomable, large-scale maps of information. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1097-1105) 2013.
10. Hu, P., Huang, M. L., Zhu, X. Y.: Exploring the interactions of storylines from informative news events. Journal of Computer Science and Technology, 29(3), 502-518, 2014.
11. Bahtin, M.M., Iliescu, N., Vasile, M.: Probleme de literatură şi estetică. Univers, 1982.