Event Models for Historical Perspectives: Determining Relations between High and Low Level Events in Text, Based on the Classification of Time, Location and Participants.

Agata Cybulska, Piek Vossen
Vrije Universiteit
Faculteit der Letteren
De Boelelaan 1105
1081 HV Amsterdam
{ak.cybulska, p.vossen @let.vu.nl}

Abstract

In this paper, we report on a study that was performed within the "Semantics of History" project on how descriptions of historical events are realized in different types of text and what the implications are for modeling the event information. We believe that different historical perspectives of writers correspond in some degree with genre distinction and correlate with variation in language use. To capture differences between event representations in diverse text types and thus to identify relations between historical events, we defined an event model. We observed clear relations between particular parts of event descriptions - actors, time and location modifiers. Texts, written shortly after an event happened, use more specific and uniquely occurring event descriptions than texts describing the same events but written from a longer time perspective. We carried out some statistical corpus research to confirm this hypothesis. The ability to automatically determine relations between historical events and their sub-events over textual data, based on the relations between event participants, time markers and locations, will have important repercussions for the design of historical information retrieval systems.

1. Introduction

The research project "Semantics of History" [1] is concerned with development of a historical ontology and a lexicon. The resources will be used in a new type of information retrieval system that can handle the time-based dynamics and varying perspectives in historical archives. History is typically a record of different realities in time and specifically focuses on the changes in reality (Ide & Woolner, 2007). Historical realities can be viewed in a subjective way, depending on the point of view of the speaker or writer. In the design of the search system of the "Semantics of History" project, we will take into consideration the reality change over time and diverse attitudes of historical text writers towards the changing reality so that they both can be used for the purpose of historical information retrieval.

In this paper, we report on a study on how descriptions of events are realized in different types of text and what the implications are for modeling the event information. The research is limited to Dutch history and we focus here on the Srebrenica Massacre, which is a recent event with a big impact on the opinions in the Netherlands.

Historical archives usually contain a mixture of news articles written shortly after a historical event happened and historical documents written with a bigger time perspective on a historical event. In those two kinds of texts the same historical events are presented in diverse ways, as exemplified by Figure 1.

Based on these two short text examples, we already notice that the Srebrenica Massacre is described in much more detail in the news fragment than in Wikipedia. In the news, some 'low level events' are presented, such as the separation of Muslim boys and men and the fact that they were taken away to a location that is very specifically described by the text author. These specific low level events together with other sub-events are part of the more general event of the Srebrenica genocide. The journalist writing the article at the time of being did not know that what was happening would later on appear to be an act of genocide. So he could not describe the events happening as such. The writer of the Wikipedia entry on the other hand, knew already that the Muslim men were taken away to be murdered. Having a time perspective on the historical event he was able to give much more background information and explanation on the event. We also get the impression that the Wikipedia writer was rather trying to show the big picture instead of concentrating on sub-events it was connected to, such as the death of a woman and a child[2].

For historical information retrieval it is crucial to map the different event representations with each other in a uniform way, regardless of the way they are expressed in

[1] The Semantics of History is funded by the Interfaculty Research Institute CAMeRA at the Free University Amsterdam as a collaboration of the Faculties of Arts and Exact Science: http://www2.let.vu.nl/oz/cltl/semhis/index.html.

[2] Of course some sub-events (mostly the most remarkable ones) can also be found in some texts written from a historical perspective but surely much less frequent than in the news and only in the context of a 'high level event' they are part of.

different text genres, allowing for full recall of information related to the same event. Within the "Semantics of History" project, a preliminary research was therefore performed on the hypothetical correlation between language use, the historical perspective and historical genre. For the purpose of creation of a historical information search system we especially researched the differences between two kinds of texts: news articles lacking time perspective and historical texts that are written with a historical perspective.

"In de brandende hitte verlieten donderdag meer dan honderd vrachtwagens en bussen volgepakt met vluchtelingen de enclave vanuit de Nederlandse VN-basis Potocari. Een vrouw en een kind kwamen te overlijden tijdens de tocht, aldus de VN.
Mannen en jongens van boven de zestien werden uit de mensenmassa gepikt en weggevoerd met onbekende bestemming. Een aantal is naar Bratunac afgevoerd, een stadje in Bosnisch-Servisch gebied ten noorden van de enclave. De Bosnische Serviërs willen onderzoeken of zij zich hebben schuldig gemaakt aan 'oorlogsmisdaden'."
News article fragment from Volkskrant published on 14 July 1995,

"Op 11 juli 1995 forceerden Servische troepen onder bevel van generaal Ratko Mladić zich met tanks de stad binnen en deporteerden en vermoordden ca. 8.000 moslimmannen en -jongens. Het waren Nederlandse troepen van Dutchbat die op dat moment de enclave theoretisch hadden moeten beschermen. Bij voorbaat was echter al bekend dat dit in de praktijk onmogelijk was. Deze actie, die in Nederland bekend staat als "het drama van Srebrenica" wordt gezien als de ergste daad van genocide in Europa sedert de Tweede Wereldoorlog."
From a Dutch Wikipedia entry: 'Het drama van Srebrenica'

Fragment from a Dutch Wikipedia entry vs. a news article fragment about the Srebrenica massacre in July 1995

"On Thursday in the burning heat more than a hundred trucks and busses packed with refugees left the enclave from the Dutch UN – base Potocari. A woman and a child passed away during the trip, according to the UN.
Men and boys over the age of 16 were separated from the crowd and taken away to an unknown destination. Some of them were transported to Bratunac, a city in Bosnian – Serb area to the north of the enclave. The Bosnian Serbs want to investigate if they were guilty of any 'war crimes'."
English translation of a Dutch news article fragment, published in Volkskrant on 14 July 1995

"On 11 July 1995 Serb troops under the command of General Ratko Mladić invaded the city with tanks and deported and murdered approximately 8,000 Muslim men and boys. At this time the Dutch troops known as Dutchbat were theoretically supposed to protect the enclave. Actually it was rather clear in advance that in practice it would not be possible. This event, known in the Netherlands as 'the Srebrenica massacre' is seen as the worst act of genocide in Europe since the Second World War. "
English translation of a fragment from the Dutch Wikipedia entry: 'Het drama van Srebrenica'

Figure 1: English translation of the Wikipedia fragment and of the news article fragment

2. Hypothesis

We believe that the difference in the historical perspective of an author corresponds with the diversity in language use and at least in some degree correlates with genre distinction. The closer to the event time a text was written the more specific and concrete the event descriptions (and events described) tend to be. The bigger the distance in time from an event the stronger the historical perspective on an event and the more general, abstract and subjective the way of event presentation.

To capture the gradual change in language use and the differences and relations between events presented in diverse historical texts over time, we use a basic event model. The event model presented below (Figure 2) is compatible with standard approaches to event modeling. For a comprehensive overview of the history of event modeling in linguistic theory the reader is referred to Tenny and Pustejovsky (2000). For resources implementing event models, we refer to FrameNet (Baker et al, 2003), SIMPLE, BSO (Pustejovsky et al, 2006), SUMO (Niles & Pease, 2001; Niles & Pease, 2003; Niles & Terry, 2004) and DOLCE (Masolo et al, 2003).

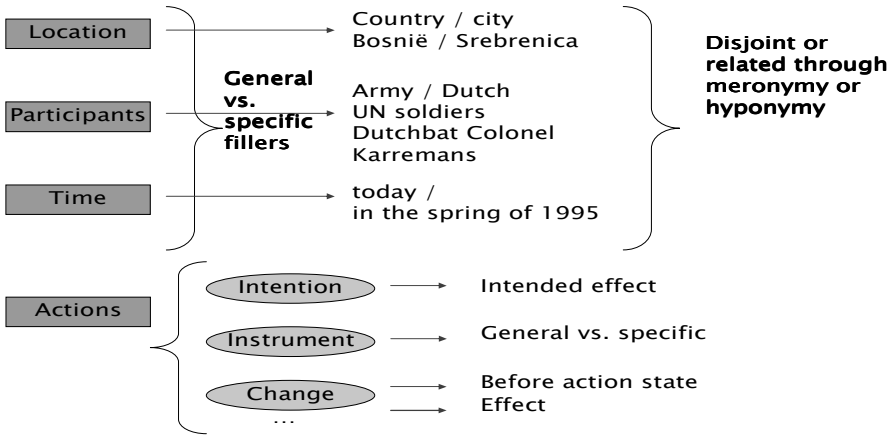

Figure 2: Minimum slots and fillers for the event model

The event model consists of 4 main slots: a location, time, participants and an action slot. The action slot in the event model might comprise of further sub-slots that might be necessary to determine the action or process performed. For now, we leave the action slot out of further consideration. We will perform more research on this part of the event model in the future. In this paper we will focus on the 3 other slots of the event model: location, time and participants.

Event descriptions in text can vary with regard to the degree of specification of the fillers for slots from the event model. Specific versus more general locations, participants and time pointers in text can be related with each other, or they can be disjoint. The significant relations between fillers for slots from the event model are those amongst the hyponymy axis (class vs. instance of a class relation such as Bosnia being an instance of the class country) and those amongst the meronymy axis (member vs. group i.e. Colonel Karremans being a member of the group of Dutch UN soldiers or part vs. whole relation such as Srebrenica being a part of Bosnia).

Furthermore, in addition to the time of an event we also have to consider another layer of time which is the time of text production that plays a crucial role in our model since it influences the time perspective which has a critical impact on generality of fillers for slots in the model. The event model after the addition of the time of writer is presented in Figure 3.

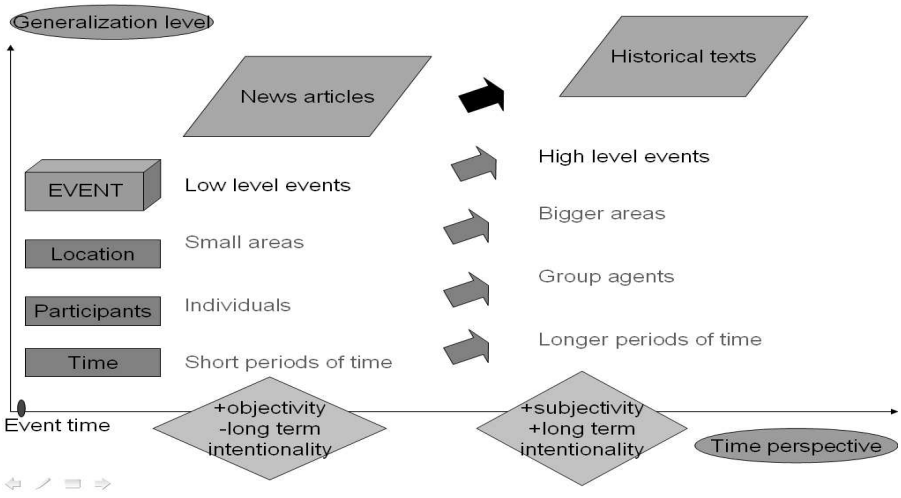

Figure 3: Correlations in the event model

Figure 3 shows how the writer's perspective (the y-axis) on an event is growing over time (the x-axis) and how it influences the language use and the level of events described in texts. In time, descriptions of events become more abstract and general, which reveals itself in the growing generality of location markers, time pointers and event participants. Also over time with the growing perspective on an event, the degree of attitude and causality (often also subjective) in text become higher as well.

On the other side, the change in historical perspective of a text author also reveals itself in favoring particular kinds of text genre. For pragmatic reasons shortly after an event happened, it is being 'reported' in news articles. With time passing by, an event will not be recent news any more so new writings about it will not

appear in the news section but views on an event from the past tend to be expressed in other more appropriate types of historical writing or other kinds of journalistic writing.

3. Corpus Composition

For the purpose of our research we created a sample 'Srebrenica corpus' which consists of Dutch texts on the Srebrenica massacre in July 1995[3]. The sample corpus consists of three sub-corpora (see Table 1). The first sub-corpus contains 26 online texts written about the Srebrenica Genocide from a historical perspective (published some years after the event). The 'historical' sub-corpus will be compared against two news sub-corpora containing news articles from two Dutch newspapers. The first news component consists of 26 articles from *Volkskrant* and the second one contains 26 articles from *Het Parool*[4]. All news articles included in the corpus were published between July 7 and 17, 1995.

Sub-corpora	No. tokens	No. texts
Volkskrant	16573	26
Texts with historical perspective	20742	26
Het Parool	13481	26

Table 1: Structure of the Srebrenica corpus

The news sub-corpora include articles representing different newspaper genres; most of them are news articles but some articles belong to other journalistic genres such as: profile article, chronology, analysis, opinion, commentary or a report[5]. The main condition for the inclusion into this component of the corpus was the short time period between the act of writing and the Srebrenica Massacre.

The corpus component representing texts, written about the events in Srebrenica from a historical perspective, consists of educational texts[6], Wikipedia entries and newspaper articles written at least few years after the massacre happened[7]. Also a number of parliament pieces[8] was included into the sub-corpus with historical perspective on the events in question.

[3] Srebrenica was captured on 11 July 1995.

[4] The news articles were acquired through http://academic.lexisnexis.nl/vu/ .

[5] Text type classification according to the *Volkskrant* – archive.

[6] An important source was www.entoennu.nl.

[7] Three dossiers on the topic were included into the corpus e. i. a dossier of the Dutch Royal Library from www.kb.nl. Some of the other sources were: www.nos.nl, www.nu.nl, www.anno.nl, www.groene.nl/home.

[8] Acquired at www.parlament.com.

4. Corpus Processing and Research

To validate our hypothesis we carried out some statistical research on the Srebrenica corpus. We conducted a contrastive analysis of event descriptions in the three components of the Srebrenica corpus with tools developed for the KYOTO project[9]. KYOTO is a platform for semantic processing of text according to a uniform conceptual model. It uses a pipeline-architecture of linguistic processors that generate a uniform semantic representation of the text in the so-called Kyoto Annotation Format (KAF)[10]. KYOTO has been tested for 7 different languages. KAF can be used to represent events with participants, locations and time periods.

For the purpose of our research Srebrenica corpus was processed by means of the KYOTO – architecture. First, the corpus was tagged with PoS- information; it was lemmatized and syntactically parsed by means of a dependency parser for Dutch - the Alpino-parser[11]. Next, word sense disambiguation was performed[12] and the corpus was semantically annotated with labels from the Dutch WordNet[13]. Finally, from each sub-corpus a hierarchy of terms was extracted by means of the KYOTO term extractor – Tybot[14].

Through the mapping of terms with WordNet synsets, it is possible to structure the full list of terms as a text-specific WordNet subtree, as is shown in Figure 4 for some time expressions. The words in big fonts (black) are Dutch equivalents for *week* and *month* from the Dutch WordNet, whereas the expressions in smaller font are specific terms and phrases for *weeks* and *months* detected in the news corpus *Volkskrant*. Note that the term extractor of KYOTO also includes general words in the term database.

The semantic classification of the terms makes it possible to quickly annotate terms for events, locations, time and participants. For example, all terms referring to soldiers of the Dutch troops and the Serb troops are grouped by a few synsets for *troops, army, soldier, etc.*, similarly for locations and time-expressions. For each labeled term, statistical data was available in KYOTO on their frequency in each sub-corpus. These statistical data on the individual terms were then generalized for each of the event-aspects: locations, time, and participants, to which they were grouped. The results have been used for the corpus analysis.

[9] More information about the 'Knowledge Yielding Ontologies for Transition-based Organization' - project can be found at www.kyoto-project.eu . See also Vossen et al (2008a).

[10] Kyoto fact Annotation Format is described in Bosma et al (2009).

[11] http://www.let.rug.nl/vannoord/alp/Alpino/

[12] For word sense disambiguation the UKB system (http://ixa2.si.ehu.es/ukb/) was used. For more information the reader is referred to Agirre & Soroa (2009).

[13] For more information see Vossen et al (2008b).

[14] For more information on the KYOTO- Tybot the reader is referred to Bosma and Vossen (2010).

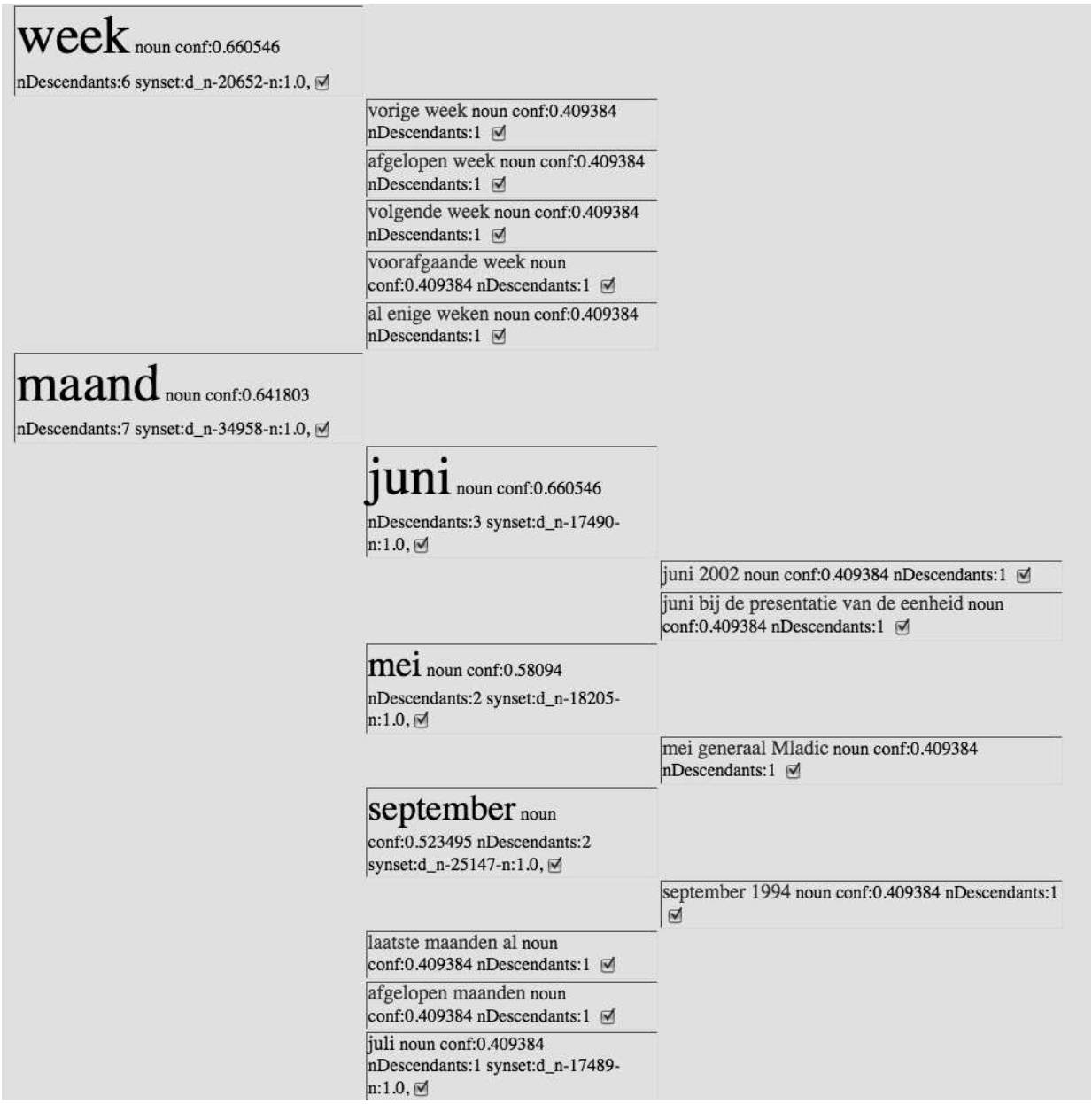

Figure 4: Term hierarchy for time expressions extracted from the news corpus *Volkskrant*

Table 2 presents the statistical information on tokens and concept types in the three components of the corpus. Also, the type token ratio was calculated per sub-corpus.

Sub-corpora	No. tokens	No. concepts extracted	Type-token ratio
Volkskrant	16573	1863	8,89
'Historical' texts	20742	1393	14,89
Het Parool	13481	1497	9,00

Table 2: Concept type vs. token statistics

The general corpus statistics from Table 2 confirm our predictions. The type token ratio for the 'historical' sub-corpus amounts to almost 15 and therefore it is remarkably higher in comparison with the type-token ratio of the both news sub-corpora which equals to 9. Those statistics show that, as expected, in texts written from a historical perspective the number of word types used is much lower than the number of types used in the news sub-corpora.

Next, we will look into statistics with regard to types and tokens used to denote locations, time and participants of the Srebrenica Massacre. To get some insights into the frequency of tokens and types referring to participants, time and location of the Srebrenica Genocide the participants, location and time markers were manually labeled as such[15] and for each corpus component the frequency of concept types and tokens was calculated.

For the purpose of the statistical analysis also per each sub-corpus normalized frequency counts of types and tokens were generated. Furthermore, the type-token ratio was counted and for the sake of completeness the percentage of unique types in the corpus was calculated [16]. Table 3 presents the results of the statistical analysis[17]

The corpus research with statistical techniques revealed that there is a clear difference in the type-token ratio for locations, time markers and participants between news and historical texts. In both news sub-corpora the type-token ratio for all elements of event descriptions is 1.1 while for 'historical' texts, as expected, the occurrence frequency of particular elements of event descriptions is higher while their diversity is lower than in the news what is especially clear when you look at the remarkably high type-token ratio of the time markers: 2.1 and participants: 1.8. Table 3 shows that the number of word types used in the news is remarkably higher than the number of types used in the sub-corpus written from historical perspective. The normalized frequencies of types confirm the hypothesis as well. Also, as expected, there are no big differences between normalized tokens frequency counts for time markers and participants of the three sub-corpora[18].

Furthermore, around 90% of all event description elements in the news occur uniquely per sub-corpus against only ca. 50% of time pointers and participants and less than 80% of locations in 'historical' texts. The percentage of the event descriptions, which reoccur per sub-corpus, in the news amounts only to ca. 10 % and is remarkably lower than in the texts written from a historical perspective: ca. 30 –50% of event descriptions are used more than once.

5. Conclusion

In this paper we showed that different historical perspectives correspond at least in some degree with historical genre diversity and correlate with variation in language use. The diversity of language use makes it difficult for an average search system to find all the information available that is semantically connected to a query but formulated in a different way. On the other hand, regularities in the language use within a genre type open new possibilities for automatic information retrieval from news articles and historical texts.

To capture differences between event representations in diverse text types, we defined an event model and we carried out some corpus research to confirm our hypothesis on the topic. Contrastive corpus analysis made clear that texts lacking historical perspective include many more specific, infrequent individual participants and markers of time and place than texts written from a time perspective, which remain rather general in their event presentation and use much more general and frequently reoccurring place and time markers as well as group participants in their description.

[15] The manual tagging was based on the semantic tagging of all concepts by means of the KYOTO-pipeline.

[16] The number of types was subtracted from the number of tokens and the difference was turned into a percentage as a result of which the percentage of uniquely occurring types in the corpus was calculated.

[17] In the future the statistical results will be further improved by inclusion of frequency counts for proper names, geo-names and adverbial time and place pointers.

[18] The relatively diversified normalized token frequencies of locations can be explained by pragmatic objectives. The events in Srebrenica happened in a relatively small area and so the variation possibilities with regard to linguistic expressions are limited by pragmatic constraints. In the future we will look into descriptions of events with a broader scope than the Srebrenica Massacre and thus which took place on a bigger geographic area.

Frequency	Sub-corpus	News sub-corpus *Volkskrant*	News sub-corpus *Het Parool*	Texts with historical perspective
Time markers	Types	84	44	36
	Tokens	95	48	75
	Relative freq. types	0,045	0,029	0,026
	Relative freq. tokens	0,006	0,004	0,004
	Type-token ratio	1,1	1,1	2,1
	Unique types %	88,4 %	91,7 %	48 %
Locations	Types	111	100	71
	Tokens	124	108	91
	Relative freq. types	0,06	0,067	0,051
	Relative freq. tokens	0,007	0,008	0,004
	Type-token ratio	1,1	1,1	1,3
	Unique types %	89,5 %	92,6 %	78 %
Participants	Types	464	379	236
	Tokens	519	412	416
	Relative freq. types	0,249	0,253	0,169
	Relative freq. tokens	0,031	0,031	0,02
	Type-token ratio	1,1	1,1	1,8
	Unique types %	89,4 %	92 %	56,7 %

Table 3: Frequency of locations, time markers and participants in components of the Srebrenica corpus

6. Future Work

In our future work we would like to use the observed relations between low and high level of actors, time and location modifiers to automatically identify event descriptions and to automatically determine relations between historical events and their sub-events, across different genres of text. Determining the relations between events over textual data will have important repercussions for the design of historical information retrieval system that is the ultimate goal of the "Semantics of History" project.

7. Acknowledgements

This research was funded by the interfaculty research institute CAMeRA *(*Center for Advanced Media Research) of the VU University of Amsterdam: http://camera.vu.nl.

8. References

Agirre, E. and A. Soroa, 2009, "Personalizing PageRank for Word Sense Disambiguation", in: *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics,* (EACL-2009), Athens, Greece.

Baker, Collin F., Fillmore, Charles J. and Beau Cronin, 2003, "The Structure of the Framenet Database", in: *International Journal of Lexicography*, Volume 16.3: 281-296.

Bosma, W., Vossen, P., Soroa, A. , Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and C. Apiprandi, 2009 "KAF: a generic semantic annotation format.", in *Proceedings of the GL2009 Workshop on Semantic Annotation*, Pisa, Italy, Sept 17-19, 2009.

Bosma W. and P. Vossen, 2010, "Bootstrapping language neutral term extraction", in: *Proceedings of the 7th international conference on Language Resources and Evaluation*, (LREC2010), Valletta, Malta, May 17-23, 2010.

Ide, N. and D. Woolner, 2007, "Historical Ontologies", in: Ahmad, K, Brewster, C., and Stevenson, M. (eds.), *Words and Intelligence II: Essays in Honor of Yorick Wilks,* Springer, 137-152.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. and A. Oltramari, 2003, "WonderWeb Deliverable D18: Ontology Library", ISTC-CNR, Trento, Italy.

Niles, I. and A. Pease, 2001, "Towards a Standard Upper Ontology", in: *Proceedings of FOIS 2001*, Ogunquit, Maine, pp. 2-9.

Niles, I. and A. Pease, 2003, "Linking Lexicons and Ontologies Mapping WordNet to the Suggested Upper Merged Ontology", in: *Proceedings of the International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada.

Niles, I. and A. Terry, 2004, "The MILO: A general-purpose, mid-level ontology", in: *Proceedings of the International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada.

Pease A., C. Fellbaum, and P. Vossen, 2008, "Building the Global WordNet Grid", in: *Proceedings of the 18th International Congress of Linguists,* (CIL18), Seoul, Republic of Korea, July 21-26, 2008.

Pustejovsky, J., Havasi, C., Littman, J., Rumshisky, A. and M. Verhagen, 2006, "Towards a Generative Lexical Resource: The Brandeis Semantic Ontology", in: *Proceedings of the Fifth Language Resource and Evaluation Conference*.

Tenny, C. and J. Pustejovsky, 2000, "A History of Events in Linguistic Theory" in: *Events as Grammatical Objects*, C. Tenny and J. Pustejovsky (eds.), 2000, CSLI Publications.

Vossen P., E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tescon, 2008a, "KYOTO: A system for Mining, Structuring and Distributing Knowledge Across Languages and Cultures", in: *Proceedings of LREC 2008*, Marrakech,

Morocco, May 28-30, 2008.
Vossen P., Maks I., Segers R. and H. Van der Vliet, 2008b, "Integrating lexical units, synsets and ontology in the Cornetto Database", in: *Proceedings of LREC 2008*, Marrakech, Morocco, May 28-30 May 2008.
Vossen P., W. Bosma, E. Agirre, G. Rigau and A. Soroa, 2010, "A full Knowledge Cycle for Semantic Interoperability", in: *Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in conjunction with the Second International Conference on Global Interoperability for Language Resources*, (ICGL 2010), Hong Kong, January 15-17, 2010.