Recognition and Normalization of Temporal Expressions in Serbian Texts

Jelena Jaćimović
University of Belgrade
Faculty of Dental Medicine
Dr. Subotića 8, 11000 Belgrade, Serbia
+381646144435
jjacimovic@rcub.bg.ac.rs

ABSTRACT

This paper presents a system for recognition and normalization of temporal expressions (TEs) in Serbian texts according to the TimeML specification language. Based on a finite-state transducers methodology, local grammars are designed to recognize calendar dates, times of day, periods of time and durations, to determine the extension of detected expressions, as well as to normalize their values, interpreted in ISO format. The results of a preliminary evaluation demonstrate usefulness of this method in both the recognition and the normalization phases.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text analysis*; F.1.1 [**Computation by Abstract Devices**]: Models of Computation – *Automata*

General Terms

Human factors, Performance

Keywords

Temporal expression recognition, temporal expression normalization, ISO-8601, finite-state transducers

1. INTRODUCTION

Processing of temporal expressions (TEs) has received increasing attention in Natural Language Processing research community over the past fifteen years. The Message Understanding Conferences (MUCs) in 1996 and 1998 have played a significant role, but their evaluations covered only recognition of TEs, while a novel contribution towards the normalization of TEs was made in 2000 [1]. The first exercise evaluating system performance that deals both with recognition and normalization of TEs was the Time Expression Recognition and Normalization (TERN) 2004 competition. With the rapid increase of electronic information and very frequent TEs occurrences, precise temporal representation of a text has become extremely important for many applications, such as machine translation, information extraction, question answering, etc.

Due to various existing interpretations of time within free text, recognition and normalization of TEs in narrative text represent a particular challenge that is far from being a simple task for automatic text processing systems. For example, in Serbian the same temporal information can be written in different forms: *13:45 časova* '13:45 hours', *1:45 popodne* '1:45 in the afternoon', *15 do dva popodne* 'quarter to two in the afternoon', and many others. Furthermore, lexical variants, such as *sat* and *čas* for a temporal unit 'hour', are also widespread. Along with some other Slavic languages, Serbian is a highly inflected and a free word order language with a complex number system in which, beside singular and plural, paucal also exists. Since numerals agree in gender and number with the nouns they modify, temporal expressions *jedan sat* 'one hour', *dva sata* 'two hours' and *pet sati* 'five hours' use three different inflected forms of a noun *sat* 'hour' – nominative singular, paucal and genitive plural, respectively.

A number of resources have been developed for the processing of temporal information in English texts, as well as in other languages, such as French, Italian, Spanish, German, Chinese, etc. Previous efforts regarding recognition of TEs in Serbian have achieved quite promising results [2]. This article details the ongoing development of a system for TEs recognition and normalization in Serbian texts.

2. TASK DEFINITION

Any natural language possesses several mechanisms for expressing temporal information, which can be grouped into three large categories: TEs, events and the temporal relations that hold among times and events [3]. As natural language phrases, TEs give information about when something happened, how long something lasted, or how often something occurred. Present task is limited to recognition of TEs denoting calendar dates (1), times of day (2), periods of time (3) and durations (4) in newspaper texts and normalization of their values, interpreted in ISO format:

(1) *8. decembra dve hiljade jedanaeste* '8th December two thousand eleventh' → "2011-12-08"
7. VIII 2008. godine → "2008-08-07"
leta 1995 'summer of 1995'→ "1995-SU"

(2) *pet minuta do ponoći* 'five minutes to midnight' → "T23:55"
12:55 časova '12:55 hours' → "T12:55"
17. februara, u 7 časova i 55 minuta uveče '17th February, at 7 hours and 55 minutes in the evening' → "XXXX-02-17T19:55"
subota u dva sata popodne 'Saturday at two hours in the afternoon' → "XXXX-WXX-6T14:00"

(3) *od podneva do pet časova popodne* 'from noon to five hours in the afternoon' → "T12/17"
između devet i 12 meseci 'between nine and 12 months' → "P9/12M"

(4) *tri nedelje* 'three weeks' → "P3W"

This work also covers modified or quantified TEs (5), such as:

(5) *početkom godine 2009.* 'early in the year 2009' → START 2009
manje od dva meseca 'less than two months' → <P2M

BCI'12, September 16–20, 2012, Novi Sad, Serbia.

Local Proceedings also appeared in ISBN 978-86-7031-200-5, Faculty of Sciences, University of Novi Sad.

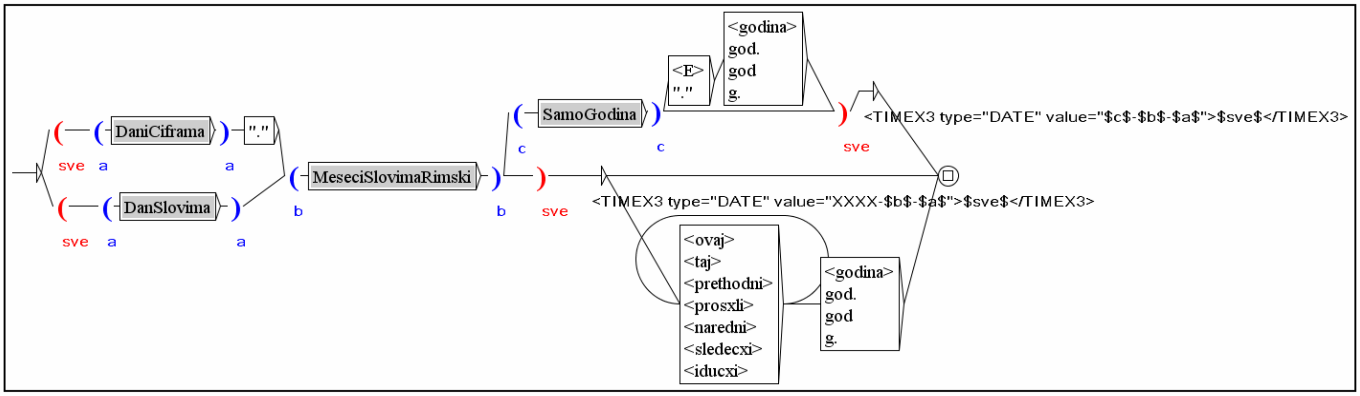

Figure 1: One path from a collection of FSTs that recognizes complete and incomplete date expressions.

Relative expressions (e.g. *juče ujutru* ‘yesterday morning’, *pre dve nedelje* ‘two weeks ago’), event anchored expressions (e.g. *35 minuta posle udesa* ‘35 minutes after the accident’) and sets of times (e.g. *mesečno* ‘monthly’) are not yet taken into consideration.

3. SYSTEM ARCHITECTURE

Serbian general-purpose lexical resources are developed with the Unitex corpus processor [4]. The role of those electronic dictionaries and dictionary Finite State Transducers (FSTs) is pre-processing of text and text tagging from a morphosyntactic point of view [5]. After this pre-processing, local grammars are applied to a text tagged with lemmas, grammatical categories and semantic features. Those local grammars, in the form of Unitex FSTs or automata, are designed to recognize TEs within an input text and to determine the extension of detected TEs, as well as to normalize their values. The required output is the recognized TEs embedded in XML tags with appropriately assigned values of attributes, according to the TimeML annotation guidelines, specified in [6, 7].

3.1 Markable Temporal Expressions

At this moment, markable TEs include absolute expressions (e.g. *3. aprila 1999. godine* ‘3rd April 1999’, *10:45 časova* ‘10:45 hours’, *zima 2007.* ‘winter 2007’) and durations (e.g. *četiri dana* ‘four days’, *79 minut*a ‘79 minutes’).

Words or particular configurations of numeric expressions whose meanings convey the concepts of time, date and duration are taken as lexical triggers and their presence in the input text discovers markable TEs. Possible triggers considered by the system include:

- nouns (e.g. *sat* ‘hour’, *čas* ‘hour’, *sekunda* ‘second’, *godina* ‘year’, *vek* ‘century’, *ponoć* ‘midnight’, *subota* ‘Saturday’, *januar* ‘January’)
- specialized time patterns (e.g. 7:15, 13.01.2009., 1992, *1980-tih* ‘1980s’)
- numbers (e.g. 4 (as in ‘She arrived at 4’), *tri* ‘three’, 6th (as in ‘He arrived on the 6th’)).

The full extent of a TE depends on the context surrounding of the detected lexical triggers. To this aim, nouns as well as noun phrases are considered as relevant information. Prepositional phrases cannot represent TE and thus they are not included in the extent of the tag (e.g. *posle* ***14 časova*** ‘after **14 hours**’, *tokom* ***poslednjih 5 godina*** ‘over **the last 5 years**’). At this moment, both temporal range expressions (3) and conjoined expressions (6) are not tagged separately.

(6) *14. i 15. februara 1992.* ‘14th and 15th February 1992’ → “1992-02-14 AND/OR 1992-02-15”

24. ili 25. avgusta ‘24th or 25th August’ → “XXXX-08-24 AND/OR XXXX-08-25”

3.2 Base Structure

The first objective was to establish the most frequent variant forms of dates and times represented in Serbian and to build their corresponding finite state automata. The usual representation of a calendar date, written using digits (Arabic or Roman), letters or both of them, is the day that is followed by the month and the year (1). In order to first track down the most certain patterns, a transducer that recognizes complete calendar dates was produced. Besides, incomplete date expressions in which year, month or day are omitted (e.g. *25. maja* ‘May 25th’, *aprila 2009.* ‘April 2009’) or can be inferred (e.g. *juna prošle godine* ‘last year on June’, *stigao je 6-og ovog meseca* ‘He arrived on the 6th this month’) were also considered (see Figure 1). Recognition of the expressions in which only the year is mentioned was also included, even if the word *godina* ‘year’ is neither stated in full nor abbreviated form (e.g. *rođen je 1976.* ‘He was born in 1976’).

Formal expressions denoting time of day (2), written using digits, letters or both of them, were also recognized. Furthermore, in regards to context analysis, it was possible to track down some time patterns after which words *čas* or *sat* ‘hour’ were not appeared (e.g. *predavanje je počelo u 8* ‘lecture began at 8’). Collection of those FSTs also covers time of day in combination with the names of days (see the last example in (2)), as well as modified time expressions (e.g. *oko 9 sati i 35 minuta* ‘about 9 hours and 35 minutes’, *oko dva sata popodne* ‘about two o’clock in the afternoon’).

Produced transducers were applied to the text to recognize patterns described in the input alphabet. When the pattern was matched, the output alphabet specified the action to be taken. For instance, FST *Datum* in Figure 1 recognizes some possible date patterns that consist of a day (written using digits or letters) followed by month (written in letters or Roman digits) followed by year (written using digits) or phrases, such as *ove godine* ‘this year’, *prošle godine* ‘last year’, *sledeće godine* ‘next year’, etc.

The output alphabet contains information on the type and value of the TE described in the input, enclosed by XML tags <TIMEX3> and </TIMEX3>. Those recognized units becomes variables $a$, $b$ and $c$, respectively, and will be used in the output as values of a tag attribute **value**.

The recognition and normalization of multi-word numerals that often appear in duration expressions was done by dictionary FSTs [5], in order to correctly tag numerals composed using both digit and alphabetic representation (e.g. *5 milijardi i 70 miliona* '5 billion and 70 millions'). Their lemmas could be retrieved from those applied dictionaries and used in <TIMEX3> output as values of a tag attribute **value**.

Therefore, rules applied in this system were grouped into possible types of TEs (DATE, TIME, DURATION) and represent a combination of an expression rule, a normalization function and the type information.

3.3 Description of the Annotation Scheme

Each detected TE was marked up with the <TIMEX3> tag, which may contain the following attributes: **type**, **value** and **mod**. At this moment, other optional attributes described in the TimeML annotation guidelines [7] remain beyond the scope of the current version of the system. The convention of indicating tag names and attribute values in all upper case (e.g. DATE, APPROX) and attribute names in lower or mixed case (e.g. type, mod) was respected.

As the non-optional attribute, **type** may have the following values: DATE (calendar time), TIME (time of the day or a combination of calendar date and time of the day) or DURATION (explicit durations).

The attribute **value** contains the normalized form of the detected date, time or duration, derived from the ISO 8601 standard format for representing time values [8]. Points of time were expressed as a string patterns YYYY-MM-DDThh:mm:ss (year-month-dayThour:minute:seconds) and may be truncated from the right (e.g. March 2002 was interpreted as 2002-03). For the unknown or vague parts of the value a placeholder character X was used (see the second example in (2)). For the representation of the normalized values, week-based format was also used – YYYY-Www-D (year-Wweek_number –day_of_the_calendar_week; see the last line in (2)). In order to separate components in the representation of time intervals, solidus [/] was used (see example (3)). Durations were expressed as a string patterns Pn, where P is used as a duration designator and n indicates one or more digits (see example (4)).

Combination of calendar dates and times-of-day were also represented with values in the ISO format. In case the text includes some reference to the specific date (7), **value** attribute must also contain the date like the following:

(7) *u 9:30, 3. januara 2007.* 'at 9:30, 3rd January 2007' → type="TIME" val="2007-01-03T09:30".

The optional **mod** attribute was used together with other attributes in order to capture temporal modifiers that change or clarify the interpretation of **value** in some way. Possible values for **mod** used at this moment are illustrated in (8).

(8) *početkom 2007.* 'early 2007' → type="DATE" value="2007" mod="START"

polovinom februara 'mid-February' → type="DATE" value="XXXX-02" mod="MID"

krajem novembra 2009. 'late November 2009' → type="DATE" value="2009-11" mod="END"

petak oko 9 časova 'Friday about 9 o'clock' → type="TIME" value="XXXX-WXX-5T09" mod="APPROX"

više od dve godine 'more than two years' → type="DURATION" value="P2Y" mod="MORE_THAN"

skoro deset dana 'nearly ten days' → type="DURATION" value="P10D" mod="LESS_THAN".

The XML output of local grammars designed for recognition and normalization of TEs is presented in the following tagged text:

Nakon <TIMEX3 type="DURATION" value="P23D">23 dana</TIMEX3> *bezuspešnog patroliranja, U-24 se* <TIMEX3 type="DATE" value="XXXX-12-16">*16. decembra*</TIMEX3> *vraća u Konstancu, gde će ostati do* <TIMEX3 type="DATE" value="1943-01-18">*18. januara 1943. godine*</TIMEX3>*, kada polazi u novo patroliranje.* 'After 23 days of fruitless patrolling, on December 16th U-24 returns to Constance, where it will stay until January 18th 1943, when a new patrol starts.'

4. EVALUATION RESULTS

The previously described system for normalization of TEs has been evaluated on a set of 6 articles from Serbian Wikipedia: *German submarine U-24, German submarine U-28, German submarine U-13, German submarine U-29, German submarine U-19* and *German submarine U-558*. These chosen texts were not used in the development of FSTs and represent completely unseen material containing many occurrences of TEs (Table 1).

Table 1: Articles Used for Evaluation

Text	Words	Date	Time	Duration
U-24	1,115	60	12	1
U-28	1,574	51	19	1
U-13	1,010	36	10	1
U-29	1,846	51	19	3
U-19	1,303	58	16	4
U-558	2,639	30	34	13
Total	9,487	286	110	23

The FSTs performance has been evaluated with respect to recognition, bracketing and normalization tasks. For that reason, a new attribute **provera** 'check' has been added to each XML tag. Possible values of this attribute were the following: OK/OK (TE was correctly recognized, full extent was correctly determined, correctly assigned normalized value), OK/NOK (TE was correctly recognized, full extent was correctly determined, but normalized value was not correct), UOK (TE was partially recognized correctly, full extent was not correctly determined (e.g. longer patterns denoting temporal ranges that were not included in FST - <TIMEX3 provera="UOK" type="DATE" value="1944-04-07">7. april 1944</TIMEX3>. - jul 1944. '7th April 1944. – July 1944.)), UOK/E (TE was partially recognized correctly, full extent was not correctly determined, because of the incorrect input), NOK (TE was partially recognized correctly, full extent was not correctly determined for some other reasons), MISS/E (TE was not recognized, because of the incorrect input) and MISS (TE was not recognized for some other reasons).

The overall evaluation of the system is presented in Table 2 and Table 3.

Table 2: Evaluation Data

Check	DATE	TIME	DURATION	Total
OK/OK	260	100	16	376
OK/NOK	0	0	0	0
UOK	11	5	3	19
UOK/E	7	1	0	8
NOK	0	1	0	1
MISS/E	1	1	0	2
MISS	7	2	4	13
Total	286	110	23	419

An error analysis shows that the main source of errors and missed TEs (lines UOK and MISS in Table 2) was the occurrence of combined temporal expressions that were not included among the FSTs rules. There were no false recognitions (line NOK), except for one case regarding time expressions. To all correctly recognized expressions were added correctly assigned normalized values (line OK/OK), which indicates that this method could be useful for both the recognition and the normalization phases. FSTs performance showed precision and recall rate of 0.946 (Table 3). Although duration expressions achieved the lowest F-measure, priority is given to precision over recall.

Table 3: Performance Measures for Recognition of TEs

TEs	Precision	Recall	F-measure
DATE	0.935	0.970	0.952
TIME	0.935	0.971	0.952
DURATION	0.842	0.800	0.821
Total	0.931	0.962	0.946

5. CONCLUSIONS AND FUTURE WORK

This paper presented the system for recognition and normalization of TEs in Serbian natural language texts, based on a finite-state transducers methodology. This approach is effective and competitive with respect to other techniques and makes it possible to include further knowledge easily [9-11]. As a rule-based temporal tagger, in the ACE TERN 2004, Chronos system [10] achieved the highest F-measure of 0.926, with precision and recall rates of 0.976 and 0.880, respectively. The system presented in this work can be compared with Chronos, since it is also based on a single module that performs both the recognition and normalization tasks.

The evaluation of the presented system is conducted on a small set of articles, but the results are quite good and in correspondence with the results obtained in previous system evaluation of TEs recognition task [2]. Nevertheless, improvements are needed in order to increase precision, which may affect further processes of temporal analysis. Future research in temporal processing is also needed to complete the tagger, in particular for recognition of relative expressions and sets of time, as well as events and temporal relations that hold between temporal entities. In order to improve the performance of this system, it would be very useful to apply transducers on the text in a precise order, as a cascade. This simple and effective way of organizing FSTs may greatly increase precision and speed of the system, as well as ability to manage priority between patterns.

6. REFERENCES

[1] Mani, I. and Wilson, G. Robust Temporal Processing of News. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (Hong Kong, 2000), 69-76.

[2] Krstev, C., Vitas, D., Obradović, I. and Utvić, M. E-Dictionaries and Finite-State Automata for the Recognition of Named Entities. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing* (Blois, France, July 2011). Association for Computational Linguistics, 48-56.

[3] Marsic, G. *Temporal processing of news: annotation of temporal expressions, verbal events and temporal relations*. Doctoral Thesis. University of Wolverhampton, 2011.

[4] Paumier, S. *Unitex 2.1 User manual*. 2011.

[5] Krstev, C. *Processing of Serbian - Automata, Texts and Electronic Dictionaries*. Faculty of Philology, University of Belgrade, Belgrade, 2008.

[6] *ISO/DIS 24617-1 Language Resources Management - Semantic Annotation Framework (SemAF) - Part 1: Time and Events (SemAF-Time, ISO-TimeML)*. International Organization for Standardization, Geneva, Switzerland, 2009.

[7] Pustejovsky, J., Bunt, H., Lee, K. and Romary, L. ISO-TimeML: an International Standard for Semantic Annotation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010* (Paris, France, 2010, 2010). ELRA, 394-397.

[8] *ISO 8601 Data Elements and Interchange formats - Information interchange - Representation of Dates and Times*. International Organization for Standardization, Geneva, Switzerland, 2004.

[9] Kolomiyets, O. and Moens, M.-F. KUL: Recognition and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (Uppsala, Sweden, 2010). Association for Computational Linguistics, 325-328.

[10] Negri, M. and Marseglia, L. *Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004*. Tecnical Report WP3.7. Information Society Technologies, 2005.

[11] Friburger, N. and Maurel, D. Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313, 1 (Feb 2004), 93-104.