Implementation of Croatian NERC System

Božo Bekavac
Department of Linguistics
University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
bbekavac@ffzg.hr

Marko Tadić
Department of Linguistics
University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
marko.tadic@ffzg.hr

Abstract

In this paper a system for Named Entity Recognition and Classification in Croatian language is described. The system is composed of the module for sentence segmentation, inflectional lexicon of common words, inflectional lexicon of names and regular local grammars for automatic recognition of numerical and temporal expressions. After the first step (sentence segmentation), the system attaches to each token its full morphosyntactic description and appropriate lemma and additional tags for potential categories for names without disambiguation. The third step (the core of the system) is the application of a set of rules for recognition and classification of named entities in already annotated texts. Rules based on described strategies (like internal and external evidence) are applied in cascade of transducers in defined order. Although there are other classification systems for NEs, the results of our system are annotated NEs which are following MUC-7 specification. System is applied on informative and noninformative texts and results are compared. F-measure of the system applied on informative texts yields over 90%.

1 Introduction

To produce a Named Entity Recognition and Classification (NERC) system for a lesser spread Slavic language like Croatian could be a task which differs a lot from the task of building such a system for a language like English, German or French. Compared to them, Croatian language has more elaborated inflectional system and more freedom in the word order within the sentence. Besides, the resources and tools needed for producing such a system (POS/MSD tagger, sentence segmentator, chunker, lexicons or gazetteers etc.) are not widely available.

But still we can say that even in languages with that kind of structural properties like Slavic languages have, named entities (NE) form a subset of natural language expressions that demonstrates relatively predictable structures. It could be questioned whether the relatively free word order in Croatian also covers the named entities (how much it influences their internal structure and their position in a sentence structure). What we also aim at in this paper is to investigate the possibility to describe NE with relatively simple rule-based systems i.e. whether it is possible to describe and classify NE in Croatian using regular grammars.

The next part of the paper describes basic methodology of our system. The third part presents strategies for NERC which have been converted into rules. The fourth part describes the architecture of the system while the fifth gives the results. The conclusion describes also possible future directions.

2 Methodology

This NERC system for Croatian is based on hand-made rules encoded in transducers which are applied in a cascade (Abney, 1996). The reason for selecting this method was simple. Since this is the first NERC system for Croatian, and there were no previous solutions for any particular NE class, we had to split the general NERC problem to a set of

Balto-Slavonic Natural Language Processing 2007, June 29, 2007, pages 11–18,
Prague, June 2007.

smaller locally manageable problems covering not just broad NE classes, but also their subclasses which were recognized by characteristic patterns. In such a way the set of rules could be kept under control and modules covering different parts of a problem could be called when needed in the runtime. In the same time the development time is shorter and the system is more consistent.

Every transducer in our system represents a local grammar (Gross, 1993) dedicated to the description of a part of a sentence i.e. local linguistic expression. The orientation to a local description where the simpler (and more certain) cases are solved first, followed by more complex ones, gives more precision to the whole system. This “island of certainty” principle (Abney, 1996:11) is also used in our NERC system.

The system uses the principle of the “longest match” as any other NERC system: in the case of more than one possible expression recognized by rules several different rules, system chooses the longest one. In this way potentially ambiguous NEs are being dynamically disambiguated (see e1 where *Madunić Ltd.* would be recognized and classified as organization NE because of the principle of the “longest match” which also included *Ltd* and thus avoiding matching only family name *Madunić* from the lexicon of names).

3 Strategies

In this section we will discuss the basic strategies that have been used in different NERC systems and their applicability for Croatian.

3.1 Internal and external evidence

The simple NER could be done by direct match of text with the list of NEs. Even if we previously solve the problem of inflection, such an approach would result with a lot of errors. In the example

> **e1:** Znali smo da je Madunić d.o.o. u vlasništvu njegova oca. (*We knew that Madunić Ltd is a property of his father.*)

the expression *Madunić* could be wrongly recognized and classified as a family name. Even that result can be questionable since it may happen that this very name is not in the list of family names. Better results could be gained by using more information i.e. features which already exist in NEs. One of features for personal names are titles such as *dr.*, *mr.*, *prof.*, *ing.* etc., for company names characteristic strings are *d.o.o.* (Ltd), *d.d.* (S.A., GmBH) etc. Such explicit strings are called *internal evidence* (McDonald 1996:22) and usually form a part of NE.

On the other hand the example such as:

> **e2:** Danas je stiglo pismo iz poduzeća "Đuro Đaković". (*A letter from the firm "Đuro Đaković" arrived today.*)

would yield simple person name *Đuro Đaković* if the contextual information of NE (i.e. string *poduzeća*/*the firm* and usage of quotes) is not taken into account. NEs often refer to certain classes such as institutions, hospitals, schools, persons, etc. Such contextual feature is called *external evidence* (McDonald 1996:22) and its recognition is mostly used as a classification criterion i.e. class membership proof. In the case illustrated by the following example:

> **e3:** U klinici za infektivne bolesti "Dr. Fran Mihaljević" tog je dana bila gužva. (*It was crowded that day at the clinic for infectuous diseases "Dr. Fran Mihaljević".*)

the external evidence is often decisive for NERC. In e3 the internal evidence (*Dr.*) represents a strong argument for a person NE, but only contextual external evidence (*klinici/the clinic* and quotes) gives the right solution.

The external evidences are crucial for NERC in any language but they also have an important role during the system development. They can be useful when a list of names is not complete – an external evidence is taking the role of an additional proof. They can also reduce the need for elaborated internal evidence checking when rules are being build.

The internal and external evidence is being used by all NERC systems such as LTG (Mikheev et al. 1999), FASTUS (Hobbs et al. 1997), Proteus (Yangarber, Grishman, 1998).

3.2 Dynamic lexicon

Sometimes during the processing there is a need for storing information which are relevant only for a current text/discourse/document. Such information are usually stored in a dynamic lexicon where temporarily relevant information are stored and used for the processing of a current document. Dynamic lexicon entries are being collected from the confident contexts and usually are being used for tagging words which could be NEs but there is not enough external evidences for that.

Dynamic lexicon could store all possible variants of a NE (a person) such as the full name and family name including middle initial, only family name, only name, only initials including all inflectional word-forms etc. In the case of companies, it could include the long company name, its shorter version and/or acronym. Distribution of acronyms shows that they frequently appear without internal and/or external evidences which are present with the full name (e.g. instead of the full name *Investicijsko-komercijalna banka*, in the text there is only *Banka* or only *IKB*). In such cases all tokens forming an NE and all their combinations are stored in the dynamic lexicon (Mikheev et al. 1999:5). In our case it would be also *Investicijsko-komercijalna, komercijalna banka, Investicijska banka*, and also an acronym derived from the first letters of all tokens (*IKB*).

Dynamic lexicon are used by a numer of NERC systems such as ones described in (Mikheev et al. 1998), (McDonald 1996) and (Piskorski et al. 2000).

3.3 Global word sequence checking

This strategy is used for solving complex ambiguities (Mikheev, 1999). The initial position in the sentence is one of such ambiguous spots. If the NE is complex e.g. has a conjuncted structure, its solving can be quite a difficult task. The following example from the newspaper can explain this:

> **e4:** Osiguranje Zagreb i Primošten potpisali su ugovor o suradnji. (*Insurance Zagreb and Primošten countersigned an agreement on cooperation.*)

> **e5:** Osiguranje Ivići i Horvatu nije isplatilo naknadu. (*Insurance didn't pay the benefit to Ivić and Horvat.*)

The token (*Osiguranje*) which in e4 is a part of NE (*Osiguranje Zagreb*) is also a common noun and is capitalized since it is in the initial sentence position. The second NE (*Primošten*) is from the list of locations but it could be also a part of conjunction (*Osiguranje Zagreb i Osiguranje Primošten*) which is shortened or forms a unique NE (*Osiguranje Zagreb i Primošten*). In the e5 there is no ambiguitiy since *Iviću* and *Horvatu* are person NEs and being in dative case clearly show that they do not belong to the same NE with *Osiguranje* (being in nominative case).

Conjunction *i* ('and') can be syntactically interpreted in two ways: it can serve as a connector of two separate NEs (*Pliva i INA*) or can be a part of NE (*Buhić i sinovi; Vodoopskrba i odvodnja*). This cases can be solved with a strategy that presupposes that at least there will be one unambiguous position for the same NE in the text. Solving the e4 example could be formulated in several steps. 1) all possible subsets of expression (*Osiguranje Zagreb i Primošten; Osiguranje Zagreb; Osiguranje Primošten; Zagreb i Primošten; Primošten…*) should be stored in a dynamic lexicon; 2) if any of this substrings is detected in the text in an unambiguous position:

> **e6:** Kapital Osiguranja Zagreb uvećan je tri puta. (*The capital of the Insurance Zagreb is enlarged three times.*)

> **e7:** Tvrtka Primošten d.d. izbjegla je stečaj. (*The firm Primošten d.d. avoided the bankrupcy.*)

the system can test that they are separate NEs and resolve the role of conjunction.

A proper solution for categorising *Primošten* is derived from this as well, since the coordinative conjunction *i* will usually connect the NEs from the same category (Mikheev, 1999).

This strategy is used in systems by Mikheev et al. (1999) and Wacholder (1997).

3.4 One sense per discourse

Ambiguous tokens, where the same string can refer to a common noun in common usage or as a part of NE, are quite common in texts (e.g. a token *Sunce* in initial sentence position can be a common noun but it has been recorded that it can also be a name of investment fund or insurance company).

Since texts are meant to be understood by readers (even when shortening and compressing procedures are used by authors) it is very rare that the same token has different meanings within the same text. Gale, Church and Yarowsky (1992) formed a hypothesis that ambiguous words have a strong tendency of keeping a single meaning in the same text/discourse. It has been experimentally proven up to 98% of cases. Therefore, detecting at least one unambiguous position for an ambiguous word enables the system to successfully solve all other ambiguous positions for this word.

3.5 Filtering of the false candidates

Specific type of problem for NERC systems pose expressions which have a structure similar to NE, but are not NEs:

> `e8:` Pripreme za Atenu 2004 približavaju se završetku. (*Preparations for the Athens 2004 are coming to the end.*)

> `e9:` Pogled nam se pružao na cijelu Atenu. (*A view to the whole Athens was in front of us.*)

In e8 string *Atenu 2004* refers to the Olympic games held in Athens 2004 and not to location NE. According to MUC specification, this should not be marked as NE. In e9 *Atenu* refers to location and should be marked as NE.

There are two possible solutions for elimination of this cases: 1) a context should be expressive enough that it can be covered by a special rule; 2) a list of false NE candidates i.e. NE-like expressions which have to be eliminated from the further processing.

It is better to discard the false NE candidates at the beginning (Karkaletis et al 1999:130) because it reduces the need for further processing and testing. The false NE candidates should not have to be deleted from the text, a better solution is to mark them with a special tag which will be deleted just before output but in the same time it will signal to the system to avoid the processing of that part of text.

Processing of false NE candidates is described thoroughly in (Stevenson, Gaizauskas, 1999:293).

4 Architecture of the system

For developing, testing and applying our NERC system we were using Intex, a well known development environment for making formal descriptions of natural languages using FSTs and their immediate application on large corpora in real-time (Silberztein 2000:8).

Our system was designed to allow the modular processing of Croatian on three levels: 1) token (single-word units) segmentation; 2) sentence segmentation; 3) multi-word units (collocations, syntagms). These modules were designed for this system but they can be used individually in any other system for processing Croatian.

Lists of personal and family names are also important for this system. We were using a list of 15,000 male and female personal names accompanied by 56,000 family names registered in the Republic of Croatia (Boras; Mikelić; Lauc 2003:224). This list was expanded to a full word-form list for every name according to the MulTextEast specification for lexica (Erjavec et al. 2003).

The rules were manually developed and tested on a subcorpus of Croatian National Corpus (Tadić, 2002) which size was 60 million of tokens of newspaper texts. The rules were coded as Finite State Transducers using Intex's graphical interface.

The system (see figure 1) consists of several sequenced modules which are applied after the tokenizaton and sentence segmentation:

1. Lexical processing: application of lexicons of common words and proper names. Unrecognized tokens are further processed with transducers which are based on characteristic endings for MSD categorization.

2. Rules (phase 1) which have the highest certainty i.e. process unambiguous text segments are being applied after the preprocessing stage. In this manner a large part of all NEs is being detected thus giving the firm anchors for the rules (phase 2);

3. Lexicon filtering: some lexical entries are highly ambiguous and make application of relaxed rules even more complex (e.g. *Kina* in Croatian can be a common noun and location NE as well. Filtering such highly frequent and ambiguous common words significantly increases results in the second phase.

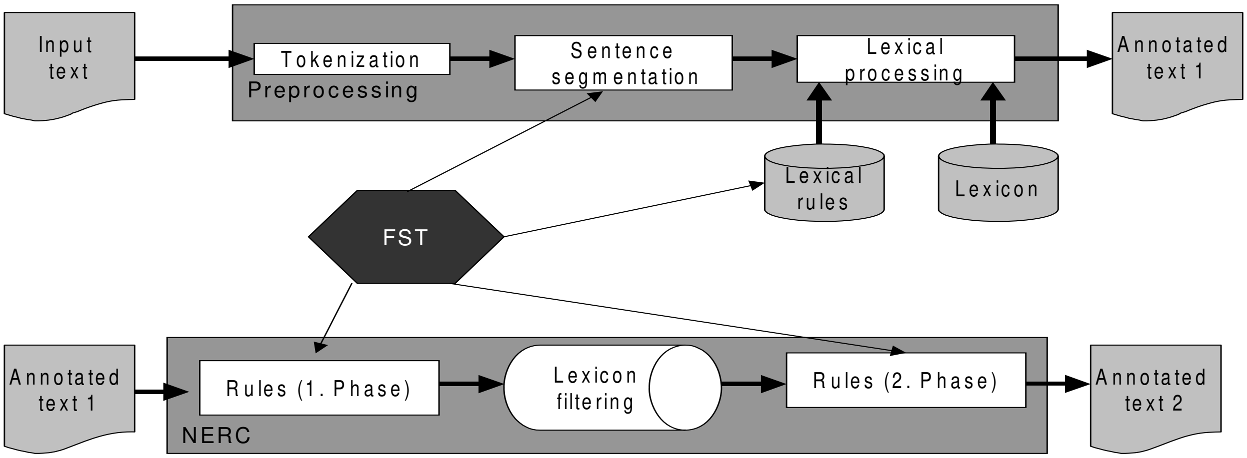

Figure 1. The general architecture of the system.

4. Rules (phase 2): all unrecognized NEs in phase 1 (mostly because of lack of supportive co-text information) are processed with new rules which are relaxed. Constraints are relaxed, but thanks to filtered lexicon precision are still rather high.

Since the overall number of rules is 106 and the description of their precise ordering and mutual interdependence would surpass the limitations of this article, we would like to exemplify the general format of the rules with the rule for detecting person NEs which include external evidence such as function of that person. Since functions can appear before or after the person NE, this rule has been stored as a separate local grammar which is being called as needed.

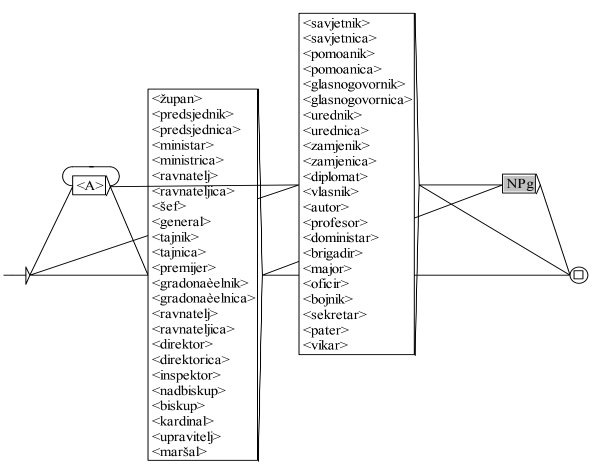

Figure 2. Graph for functions (funkcije.grf).

Beside the function name, an attribute <A> can appear on the left and NP in genitive case [NPg] can appear on the right of function name.

This local grammar ([funkcije] in grey) is being called in cascade from two other grammars for person NE detection such as:

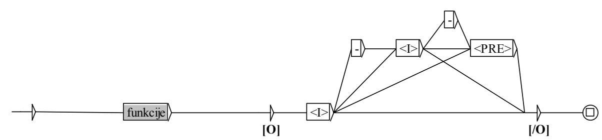

Figure 3. Graph for functions + names.

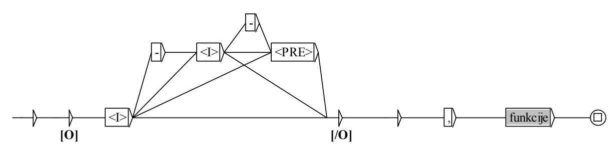

Figure 4. Graph for names + functions.

In figures 3 and 4 <I> represents a personal name recognized from the list of personal names while <PRE> represents a capitalized token. [O] and [/O] are tags that system inserts for person NE annotation. In this way potentially ambiguous NEs like *Predsjednik Microsofta* and *Predsjednik Šeks* could be resolved since only *Šeks* belongs to a list of personal names. The grammar in figure 3 can recognize cases such as:

```
et Hrvatskoj, isto kao i američkom ministru [O]Ronaldu Brownu[/O] koji je s
roprivredi Bosne, ističe generalni direktor [O]Mijo Brajković[/O]. On nagla
astrojstvo, a desna ruka generalnog direktora [O]Jana Bobosikova[/O] prekju
še od godinu dana pisala nadbiskupu [O]Josipu Bozaniću[/O] upozoravajući ga
lamenta Vaclava Klausa i predsjednika Češke Republike [O]Vaclava Havela[/O]
```

while the grammar in figure 4 can recognize cases such as:

```
sportaša i nakon što su [O]Aleksandra Mindoljević[/O], predsjednica žirija
avi aktivnom politikom. [O]Andrija Hebrang[/O], ratni ministar zdravstva,
im biznisom. Riječ je o [O]Davoru Šternu[/O], bivšem generalnom direktoru
obode. Međutim, [O]Rahim Ademi[/O], general hrvatske vojske još èeka odluku
rak [O]Hans Dietrich Genscher[/O], bivši njemaèki ministar vanjskih poslova
```

All local grammars for detecting personal NEs are being called from a grammar on upper level:

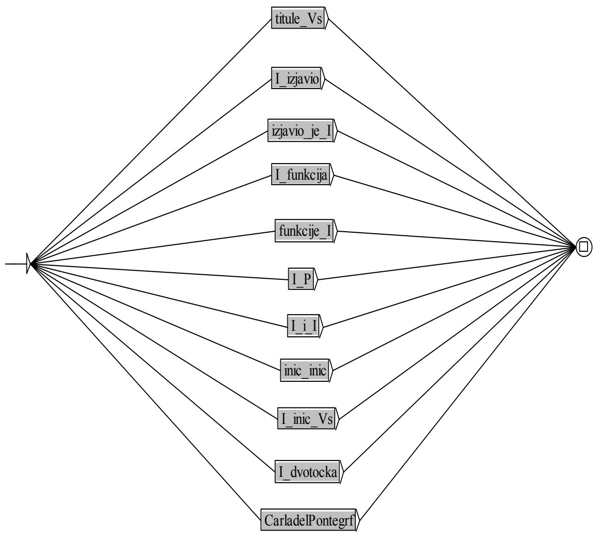

Figure 5: Graph with all person NE graphs

Similar set of rules and modular local grammars has been developed for other NE categories.

The order of applying rules (i.e. local grammars) plays important role in our NERC system. There are at least two reasons for that.

1) Certain rule can be valid for a NE which can be part of a larger NE. Rules for organization NE detection should be applied prior to rules for person NE detection. In this way correct categorisation is being achieved (e.g. *Đuro Đaković holding d.d.* where a person NE should not be used and subsumed under larger organization NE). Even if both grammars are applied simultaneously, still the principle of "longest match" would yield the correct categorisation (Poibeau, 2000). The same ordering should be kept in mind for other types of NEs which could be subsumed (e.g. dates or locations within the names of streets etc.).

2) The degree of certainty is decisive for rule ordering: the most certain NEs are being processed at the beginning and thus lowers the ambiguity also within the same category.

5 Results and discussion

Our NERC system for Croatian was tested on two types of texts: newspaper articles from *Večernji list* (economy and internal affairs, 350 articles from 2005-01, 137.547 tokens) and two textbooks from the history of arts and culture (143.919 tokens) (Maković, 1997; Žmegač, 1998). The results for newspaper texts are given in Table 1, while results for textbooks are given in Table 2.

F-measure of the whole system calculated as average from F-measures of all categories is 0.92. Since all NE categories are not equally represented in texts, more realistic measure of system efficiency can be acquired by counting all NEs that current version of a system with this set of rules should detect and categorize in a text. In this case F-measure drops to 0.90 which is still very good result.

The same rules applied to another genre (textbooks) show a significant drop in the accuracy of the system. Precision is still at 0.79 but recall is at 0.47 thus resulting with F-measure at 0.59. The most serious drop is in personal and location names. Possible explanation could be that in textbooks used for testing there is a lot of unknown, possibly foreign, names but this has to be checked in detail on more different genres.

Compared to a similar system for NERC in French texts (Poibeau; Kosseim 2001:148), where also Intex was used as a development environment, we got similar results. System developed for French yielded 0.9 for informative texts and 0.5 in noninformative texts (prose).

The example of the input and output from our system can be seen at http://hnk.ffzg.hr/nerc/.

Theoretically syntactic rules in Croatian do allow central embedding in NPs thus splitting them in two separate strings. If we apply this rule to a NERC domain, we could think of a construction which consist of function and personal name:

> **e10:** **bivši hrvatski predsjednik*, koji je stvorio hrvatsku državu, *Franjo Tuđman*...(**former Croatian president, who founded a Croatian state, Franjo Tuđman*...)

	Person	Organization	Location	Percentage	Currency	Time
Precision	0.95	0.93	0.98	0.99	0.99	0.94
Recall	0.69	0.86	0.93	0.99	0.99	0.90
F-measure	0.79	0.89	0.95	0.99	0.99	0.92

Table 1: Results for newspaper articles

	Person	Organization	Location	Percentage	Currency	Time
Precision	0.65	0.69	0.61	0.95	0.92	0.91
Recall	0.35	0.38	0.31	0.66	0.61	0.53
F-measure	0.46?	0.49	0.41	0.78	0.73	0.67

Table 2: Results for textbooks

In practice constructions of this type were never detected even in a very large corpus (>100 Mw). This led us to a conclusion that in spite the relatively free word order in Croatian, for NERC systems regular grammars could be sufficient instead of stronger formalism such as context-free grammars. NEs are local phenomena in sentences and are usually kept in one constituent. It looks like the free word order allows recombination of constituents (scrambling) while withing the constituents it is not allowed and they could be locally recognized by regular grammars. Although context-free grammars encompass regular ones, the development time for regular grammars, particularly if they are built as small-scale local grammars which are cascaded later, is much shorter and developers have stronger control over of each module, its input and output.

6 Future directions

Although it features in some areas quite promising results, this system if far from being complete. Our future directions could be: 1) testing the system on a whole different range of genres with possible rule adaptation for each genre; 2) widening the list of person and family names to include foreign names; 3) thorough analysis and typology of most typical errors; 4) include also other NEs classification schemes which go beyond MUC-7 specification; 5) since this system highly depends on Intex runtime library under which it has been designed, it is not possible to distribute it as a stand-alone application. We would like to reprogram the whole set of rules on a different platform or programming language. In this way this system can became a core of a web-based service for NERC in Croatian which is also one of our intentions.

Acknowledgments

This work was partially supported by the Ministry of the Science, Education and Sports of the Republic of Croatia within the project 130-1300646-0645 and partially by Flemish and Croatian governments within the joint CADIAL project.

References

Abney, Steven. 1996. *Partial Parsing via Finite-State Cascades*, Journal of Natural Language Engineering 2 (4):337–344.

Damir Boras, Nives Mikelić, Davor Lauc. 2003. *Leksička flektivna baza podataka hrvatskih imena i prezimena*, Modeli znanja i obrada prirodnog jezika – Zbornik radova, Radovi Zavoda za informacijske studije (vol. 12):219–237.

Tomaž Erjavec (ed.). 2001. *Specifications and Notations for MULTEXT-East Lexicon Encoding. Edition Multext-East/Concede Edition*, March, 21, p. Available at [http://nl.ijs.si/ME/ V2/msd/html/].

Friburger, Nathalie; Maurel, Denis. 2004. *Finite-state transducer cascades to extract named entities in texts,* Theoretical Computer Science, 313(1):93–104.

William Gale, Kenneth Church, David Yarowsky. 1992. *One Sense per Discourse*, Proceedings of the 4th DARPA Speech and Natural Language Workshop, Harriman, NY:233–237.

Maurice Gross. 1993. *Local grammars and their representation by finite automata*, Data Description, Discourse (ed. M. Hoey), Harper-Collins, London:26–38.

Jerry R Hobbs, Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, Mabry Tyson. 1997. *FASTUS: A cascaded finite-state transducer for extracting information from natural language text*, Finite State Devices for Natural Language Processing, (ed. Roche, E.; Schabes, Y.), MIT Press, Cambridge, MA:383–406.

Vangelis Karkaletsis, Georgios Paliouras, Georgios Petasis, Natasa Manousopoulou, Constantine D. Spyropoulos. 1999. *Named-Entity Recognition from Greek and English Texts*, Journal of Intelligent and Robotic Systems, 26(2):123–135.

Maković, Zvonko. 1997. *Vilko Gecan*, Matica hrvatska, Zagreb.

David McDonald. 1996. *Internal and external evidence in the identification and semantic categorization of proper names*, Corpus Processing for Lexical Acquisition, chapter 2, ed. Boguraev; Pustejovsky, The MIT Press, Cambridge, MA:21–39.

Andrei Mikheev, Claire Grover,Marc Moens. 1998. *Description of the LTG system used for MUC-7*, Proceedings of the 7th Message Understanding Conference (MUC-7), Fairfax, Virginia

Andrei Mikheev, Claire Grover,Marc Moens. 1999. *Named Entity Recognition without Gazetteers*, Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics, Bergen:1–8.

Andrei Mikheev. 1999. *A Knowledge-free Method for Capitalized Word Disambiguation*, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics:159–166.

Jakub Piskorski, Günter Neumann. 2000. *An Intelligent Text Extraction and Navigation System*, Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (RIAO'00), Paris

Thierry Poibeau. 2000. *A Corpus-based Approach to Information Extraction*, Journal of Applied Systems Studies, 1(2):254–267.

Thierry Poibeau, Leila Kosseim. 2001. *Proper Name Extraction from Non-Journalistic Texts*, Computational Linguistics in the Netherlands 2000: Selected Papers from the Eleventh CLIN Meeting, W. Daelemans, K. Sima'an, J. Veenstra, J. Zavrel (ed.), Rodopi, Amsterdam:144–157.

Max Silberztein. 1999. *INTEX: a Finite State Transducer toolbox,* Theoretical Computer Science #231:1, Elsevier Science

Max Silberztein. 2000. *INTEX Manual*. ASSTRIL, Paris

Mark Stevenson, Robert Gaizauskas. 1999. *Using Corpus-derived Name Lists for Named Entity Recognition*, Proceedings of the sixth conference on Applied natural language processing, Seattle, Washington, Morgan Kaufmann Publishers Inc.:290–295.

Marko Tadić. 2002. *Building the Croatian National Corpus*. LREC2002 Proceedings, Las Palmas, ELRA, Pariz-Las Palmas, Vol. II:441-446.

Nina Wacholder, Yael Ravin, Misook Choi. 1997. *Disambiguation of Proper Names in Text*, Proceedings of the Fifth Conference on Applied Natural Language Processing:202–208.

Roman Yangarber, Ralph Grishman. 1998. *NYU: Description of the Proteus/PET system as used for MUC-7 ST*, Proceedings of the 7th Message Understanding Conference (MUC-7), Fairfax, Virginia.

Žmegač, Viktor. 1998. *Bečka moderna*, Matica hrvatska, Zagreb.