# A System For Recognition of Named Entities in Greek

Sotiris Boutsis[1,2], Iason Demiros[1], Voula Giouli[1], Maria Liakata[3],
Harris Papageorgiou[1], Stelios Piperidis[1,2]

[1] Institute for Language and Speech Processing
Artemidos 6 & Epidavrou, 151 25, Athens, Greece
tel: +301 6875300, fax: +301 6854270
{sboutsis, iason, voula, xaris, spip}@ilsp.gr

[2] National Technical University of Athens

[3] Cambridge University
ml257@cam.ac.uk

**Abstract.** In this paper, we describe work in progress for the development of a Greek named entity recognizer. The system aims at information extraction applications where large scale text processing is needed. Speed of analysis, system robustness, and results accuracy have been the basic guidelines for the system's design. Pattern matching techniques have been implemented on top of an existing automated pipeline for Greek text processing and the resulting system depends on non-recursive regular expressions in order to capture different types of named entities. For development and testing purposes, we collected a corpus of financial texts from several web sources and manually annotated part of it. Overall precision and recall are 86% and 81% respectively.

## 1 Introduction

In this paper, we present a system that recognizes and classifies named entities (NE) in Greek text. The system has been developed in the framework of the EPET II "oikONOMiA" project, which aims at the construction of a pipeline integrating NE recognition, shallow parsing, and co-reference resolution technologies. The pipeline will analyze text to produce a shallow semantic representation suitable for template filling in scenario based information extraction (IE) applications.

Natural Language Processing (NLP) systems performing information extraction have gained the focus of attention of both the academic and the business intelligence community. NERC is the first task in the information extraction task series. Several factors contribute to its complexity. Name-list based recognition is not adequate, since unknown names should be dealt with in addition to names appearing in the lists. Moreover, known names may be of several types; commonly used Greek names can be of type person, organization, location, or none of the above. Moreover, the name classification schema can vary significantly across domains and applications. Thus, there are two aspects in NERC: 1) recognition and classification of known names, and

2) spotting and classification of new names. It should be noted that the creation, adaptation, and maintenance of name databases comes at a significant cost; new text needs to be scanned for names or name aliases, which should be linked to the entities they refer to. This is common in dynamic news scanning and routing.

We followed the MUC-7 NE task definition with certain adaptations. We capture organization, person and location names (ENAMEX), date and time expressions (TIMEX), percent and money expressions (NUMEX). The system is composed of a series of basic language technology building blocks for Greek developed in ILSP. The tools are modular with streamed I/O which enables their combination in a pipeline. A common Tipster-like annotation and data representation model underlies the whole application.

An initial finite state preprocessor performs tokenization and sentence boundary identification. A part-of-speech Brill tagger trained on a manually annotated corpus and a lexicon-based lemmatizer carry out morphological analysis and lemmatization. A lookup module matches name lists and trigger words against the text, and, eventually, a finite state parser recognizes NE's on the basis of a pattern grammar. A corpus of 130.000 words was used to guide system development.

System evaluation and testing was carried out against a manually annotated corpus of 20,000 words. Performance was measured with the recall (R), precision (P), and F-measure (F = 2PR / (P+R)) scores. The system achieves P=86%, R=81% and F=83%. Systems participating in MUC-6 and MUC-7 typically report F-measures around 90%, approaching human performance. We have to note, however, that our system was tested in a more diverse corpus than the MUC data set. Present performance is encouraging, but there is certainly room for improvement.

## 2 Background

Several successful systems for large-scale, accurate named entity recognition have been built. The majority of the systems operate on English text and follow a rule-based and/or probabilistic approach, with hybrid processing being the most popular.

The NYU system for MUC-6 [11] uses sets of regular expressions which are efficiently applied with finite state techniques. The system records the initial appearance of each name and its type; subsequent appearances of substrings of previously seen names are recorded as aliases. The F-measure is 80%. IsoQuest's NetOwl pattern based system [15] has been commercialized and performs around 90%. The NERC system developed in DFKI [17] for German text processing is based on FST's and performance ranges between 66% and 87% for different NE types.

The LaSIE system used in MUC-6 and MUC-7 [9] processes the input text by performing list-based matching and parsing with a special proper name grammar produced by hand. The LaSIE parser is a bottom-up Prolog chart parser. LaSIE's F-measure is 92%. An approach similar to the one in LaSIE is taken by NCSR Demokritos [13] for Greek and scores 73% and 97% for Recall and Precision respectively. Rule-based NE recognition is also followed by Umist in FACILE [3]. The MITRE Alembic system [1] relies on sequences of phrase rules, both hand-crafted and auto-

matically learned through the application of Brill's error-reduction learning algorithm [5]. The system achieved 85% success rate in MUC-6.

A probabilistic language model built from a training corpus is employed in the Kent Ridge Digital Labs system [20]. Nymble [2] is another statistical approach to NERC using a variant of the standard Hidden Markov Model. It achieves an F-measure of 91% in English and 90% in Spanish.

The NYU MENE system [4] for MUC-7 is based on maximum entropy (ME) modeling. ME modeling facilitates the combination of diverse pieces of contextual evidence for the estimation of the probability of a linguistic class, and consequently lends itself to NERC. The system has been trained on a manually annotated 270K word corpus, makes use of a broad array of dictionaries, and contains no hand-generated patterns. MENE exhibits performance of 92% for the dry-run test and 84% for the formal test. The LTG system makes use of several stages of rules and pre-trained ME models [16], achieving an F-score of 93%.

There have been several efforts to apply decision-tree techniques to the NERC task. A. Gallippi approached multilingual NERC [10] using an initial core set of linguistic features and a decision tree classification scheme. A system optimized for English (F=94%) has been ported to Spanish (F=89%) and Japanese (F=83%). Sekine et al. ([18], [19]) describe a system using a decision tree to classify names in Japanese. The CLR/NMSU team propose [7] two NE recognition systems for MUC-6. The first is a data intensive method that uses human generated patterns. The second uses training data to develop decision trees.

## 3 Specifications

Specifying the annotation schema for the Greek NERC task, we followed the MUC-7 guidelines [6]. In particular, we cater for the identification of NE's of types ENAMEX (PERSON, ORGANIZATION, LOCATION), TIMEX (TIME and DATE) and NUMEX (MONEY, PERCENT). A brief summary of our guidelines is given here under:

We mark entities appearing in the text with their full-name, an abbreviated/reduced form of this name (e.g. "Εθνική Τράπεζα της Ελλάδος/National Bank of Greece – Εθνική / National"), or a word/phrase - usually a metonymy - consistently used to describe it (e.g. "Ηρακλής (soccer team) - ο Γηραιός", "Χρηματιστήριο Αξιών Αθηνών / Athens Stock Exchange – Χρηματιστήριο / Stock Exchange – Σοφοκλέους / (the street where ASE is located)"). Of course, simple pronominal or nominal references to NE's are not marked. NE's connected through part-whole and possessor-possessed relations are marked independently, e.g. "Το [org Τμήμα Ανάλυσης και Μελετών /org] της [org Εγνατίας ΑΧΕ /org] / The [org Research Department /org] of [org Egnatia Securities /org]". Quotes are included in the NE when they are embedded in it, or when they cover it exactly.

**Person:** It is quite common for a company owner's name to appear in the company title. Thus, caution should be taken to correctly identify whether a person name refers to a person or a company, e.g. "ο κ. [person Μυτιληναίος /person] παρουσίασε τους

αναπτυξιακούς στόχους της [org Μυτιληναίος /org] μέχρι το τέλος του έτους / Mr. [person Mytilinaios /person] presented the growth target of [org Mytilinaios /org] for this year". As it is specified in the MUC-7 guidelines, titles such as "κ., κος, κον / Mr.", "κα./Miss,Ms", "πρόεδρος/president", "διευθύνων σύμβουλος/CEO", etc. are not marked as part of the NE. Also we do not mark person names included in the names of prizes, products, methods etc. E.g. "τεστ Παπανικολάου/pap test", "βραβεία Ωνάση/Onasis awards".

**Organization:** Councils and committees are marked as NE's only when they are written with their first letters in capital, e.g. "[org Υπουργικό Συμβούλιο /org] / [org Council of Ministers /org]", "[org Διοικητικό Συμβούλιο /org] / [org Board of Directors /org]". NE's of type location are included in an organization name only when they function as NP modifiers in genitive, e.g. "η [org Τράπεζα της/det Ελλάδος /org] / the [org Bank of Greece /org]". On the contrary, location names in complement position of prepositional phrases modifying organizations are not included in the organization NE's, e.g., "η [org Ελληνική Πρεσβεία /org] στα/prep [loc Τίρανα loc] / the [org Greek Embassy /org] at [loc Tirana /loc]". Organization designators, e.g. "εταιρεία / [company, society], οργανισμός / organization", are included in the organization name only when they are written with a capital first letter. E.g. "εκδόσεις [org Σάκουλα /org] / [org Sakoula /org] publications" vs. " [org Εκδόσεις Ερμής /org] / [org Ermis Publications /org]". Only "υπουργείο / ministry" and "χρηματιστήριο / stock exchange" are excluded from this rule, e.g. "[org υπουργείο Εξωτερικών /org] / [org ministry of Foreign Affairs /org]", "[org χρηματιστήριο της Φρανκφούρτης /org] / [org Frankfurt stock exchange /org]". Company prefixes and suffixes, e.g. "Αφοί", "A.E.", "A.X.E.", etc. are included in the organization name when present.

**Location:** According to the MUC-7 guidelines, location names used to refer to organizations are marked as locations: "Η [loc Ιταλία /loc] νίκησε τη [loc Βραζιλία /loc] / [loc Italy /loc] won [loc Brazil /loc]". In contrast to the MUC-7 guidelines, locative specifiers accompanying location names are always included in the named entity, e.g. "[loc αεροδρόμιο Αθηνών /loc] / [loc Athens airport /loc]", "[loc λιμάνι του Πειραιά loc] / [loc Piraeus port /loc]". Adjectives modifying a location name are included in the named entity only when they are written with a capital first letter. E.g. "[loc Βόρειος Αμερική /loc] / [loc North America /loc]", "βόρειο [loc Αιγαίο /loc] / north [loc Aegean /loc]". Adverbs are not included in the named entity unless they are part of the formal name, e.g. "πρώην/adv [loc Σοβιετική Ενωση /loc] / former [loc Soviet Union /loc]", but "[loc πρώην/adv Γιουγκοσλαβική Δημοκρατία της Μακεδονίας /loc] / [loc former Yugoslavic Republic of Macedonia /loc]".

**Date and Time:** Following the MUC-7 guidelines, we mark absolute date and time expressions, e.g. "[date Παρασκευή 23 Ιουλίου 1999 /date] / [date Friday 23 July 1999 /date]", "[time 10 μ.μ. /time] / [time 10 pm /time]". We also mark relative expressions indicating a specific date or time, e.g. "[date σήμερα /date] / [date today /date]", "[date χθές /date] / [date yesterday /date]", but not vague expressions that do not point to a specific date or time, e.g. "πριν μερικές ημέρες / a few days ago". Decades and centuries are marked, too. Names of seasons, months, days, holidays, and heads with date/time modifying expressions are included in the date/time. E.g. "[date δεκαετία του '80 /date] / [date the 80's /date]", "πριν την [date Πρωτοχρονιά του 2000 /date] / before the [date new year's day of 2000 /date]", "το [date οικονομικό

έτος 2000 /date] / the [date fiscal 2000 /date]", "το [date σχολικό έτος 2000 /date] / the [date school year 2000 /date]". Also "[time 10πμ ώρα Ελλάδας /time] / [time 10 am Greek time /time]". Expressions such as "αρχή/beginning", "τέλη/end", "μέσα/mid" are marked with the date following them. We do not mark nouns expressing duration, e.g. "περίοδος [date 1990-1995 /date] / period [date 1990-1995 /date]". Unlike MUC-7, temporal units such as "πρωί/morning", "απόγευμα/evening" are marked even if they are not followed by an absolute temporal expression. E.g. "[time 10 το πρωί /time] / [time 10 in the morning /time]", "το [time πρωί /time] / in the [time morning /time]".

**Money and Percent:** We mark only numeric expressions followed by a currency expression or a percent. Currency names which are not followed by a specific numeric expression are not marked. Country names post-modifying the currency name are marked too: "[money 10 εκατ. δολάρια ΗΠΑ /money] / [money 10 million USA dollars /money]". Unlike MUC-7, we do not mark monetary expressions modified by multipliers such as "αρκετά εκατομμύρια δολάρια / several million dollars". Percent ranges are marked as one entity. Approximators, e.g. "περίπου / about", are not marked. E.g. "περίπου [percent 10%-15% /percent] / about [percent 10-15% /percent]".

## 4 The Corpus

A corpus of Greek texts of ca. 12,000,000 words in total comprising articles from financial newspapers and magazines (Express, Naftemporiki, Isotimia, Oikonomikos Tahidromos, and Vima) was downloaded from the web. As we wanted to use text with a high density of named entities, only the articles with the highest percentage of words with an uppercase first letter were chosen. The selected articles formed the training and testing corpus, which amounts to ca. 150,000 words. This corpus was then manually annotated according to the annotation schema described in the "Specifications" section. A TclTk graphical user interface facilitated the manual annotation of NE's in the text. Following MUC, document sections were delimited by SLUG, DATE, NWORDS, PREAMBLE, TEXT and TRAILER tags. The annotated corpus was used for both development and evaluation: 130,000 words were used to guide system development, e.g. evaluate rule performance, while the remaining 20,000 words of text were put aside for testing purposes.

## 5 System Architecture

System architecture is illustrated in Figure 1. The main system components are: Tokenizer, POS Tagger & Lemmatizer, Name Lookup, and NE Parser. All processing modules share a common Tipster-like [12] data model that facilitates efficient interoperation and addition of new annotation. The system runs under the PC/Windows operating system.

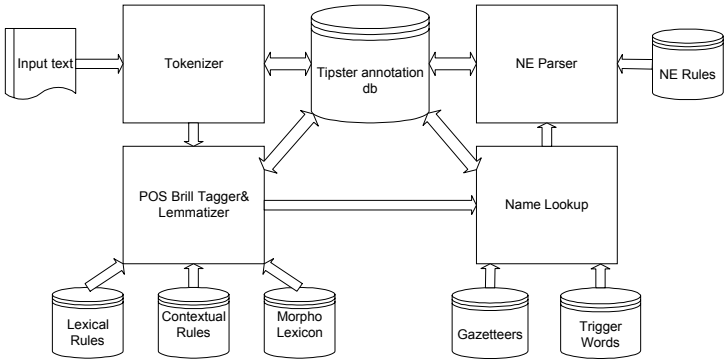


**Figure 1**: System Architecture

### 5.1 Tokenizer

Recognizing and labeling surface phenomena in the text is a necessary prerequisite for most Natural Language Processing (NLP) systems. At this stage, texts are rendered into an internal representation that facilitates further processing. Basic text handling is performed by a MULTEXT-like tokenizer [8] that identifies word boundaries, sentence boundaries, abbreviations, digits, and simple dates. Following common practice, the tokenizer makes use of a regular-expression based definition of words, coupled with downstream precompiled lists for the Greek language and simple heuristics. This proves to be quite successful in effectively recognizing sentences and words, with accuracy up to 95%.

### 5.2 Part-of-Speech Tagger & Lemmatizer

We use the Brill tagger [5] trained on Greek text. Rules were automatically learned from a manually annotated Greek corpus of 250K words. We use the PAROLE tagset, which, conforming to the guidelines set up by TEI and NERC, captures the morphosyntactic particularities of the Greek language. There are 584 different part-of-speech tags, so the usually reported Brill tagger accuracy is degraded down to 90%. First, the tagger assigns initial tags, looking up in a lexicon created from the manually annotated corpus during training. A suffix-lexicon is used for initially tagging unknown words. 799 contextual rules are then applied to improve the initial phase output. After part-of-speech tagging has taken place, the lemmas are retrieved from a Greek morphological lexicon containing 70K lemmas.

### 5.3 Name lookup

At this stage, a set of static pre-stored names and regular expressions are matched against the tokenized, tagged, and lemmatized text in order to identify known named entities and trigger words.

We compiled lists of person, organization, and location names, combining material from several different sources such as yellow pages, company lists and place name lists available from the Athens Stock Exchange, the Technical Chamber of Greece, the Hellenic Telecommunications Organization, the National Statistical Service of Greece, etc. The name lists were also enhanced with names extracted from 130,000 words of manually annotated text. After all additions, the company name list had 1,059 entries, the location name list 793 entries, and the person name list 1,496 entries.

Furthermore, we formed lists of words, multi-words and regular expressions which are indicative of the existence of named entities in their surrounding, such as company designators, person titles, currency units, occupations, etc. This was done by automatically extracting indicative words through the application of word count and mutual information statistics to windows of 3-5 words to the left and to the right of each named entity in the training corpus and then manually clustering extracted words according to their use and semantics. These clusters were manually edited and further augmented during NE grammar development. At the name lookup stage, words appearing in a cluster get a specific tag which fires corresponding rules during the parsing phase. There are also regular expressions matching more than one words. In total, we use 57 clusters containing 920 words, multiwords, and regular expressions.

Name lookup is implemented on the basis of finite state recognizers, scanning the text at high speed for the existence of strings and regular expressions appearing in the name lists and clusters.

### 5.4 NE Parser

This is the last component of the NERC pipeline and finalizes the annotation added at previous stages.

Although a name in the text may appear in one of the lists, this does not necessitate that the name is of the corresponding to the list type. Context should also be taken into account to reach a safe conclusion. For instance, a company designator following a location name, could be used to correctly recognize the preceding name as of type company. To this end, rules are applied to the output of the name lookup stage to finalize named entity typing, as well as to recognize names not in the lists. Rules operate on the basis of: names recognized at the lookup stage, capitalization information, POS tags, and tags corresponding to the clusters mentioned in the previous section. Rules are written in the form of regular expressions [14] which are compiled into finite state transducers that transform input text by inserting or removing special markers. Rules are sequentially applied to the text using longest match. We make use of the FSA6 package [21] for compiling rules into finite state transducers and a C parser for efficiently applying them on the text.

The grammar consists of 110 rules in total: 17 for person, 19 for location, 37 for organization, 23 for date, 5 for time, 7 for money and 2 for percent. There are two types of rules: simple and composite, the latter being the ordered composition of two or more rules applied at the same pass. Rules may or may not take context into account. An example of a composite rule is given below:

```
markup (
[geosign+, atdf_ge^, {cap_aj, locadj_cap}^, abbr^,
{const({'[person', '[loc'}, {'/person]', '/loc]'}), cap_word,
cap_rg}+, dig^],
'[loc', '/loc]')

o

conditional_markup_upward(
[{cap_rg, cap_word}+ ], '[loc', '/loc]',
[{geosign, indiclocverb}, as_se],
[] )

<EOR>
```

This rule recognizes structures such as “[loc Οδός Σίνα 4 /loc] / [loc 4 Sina Street /loc]” , “[loc νομός Θεσσαλονίκης /loc] / [loc prefecture of Thessalonica /loc]”, “ορυχείο στο [loc Πότι Ρουμανίας /loc] / mine at [loc Poti, Romania /loc]”, “πυρηνικός σταθμός στο [loc Τσέρνομπιλ loc] / nuclear plant in [loc Chernobyl /loc]”, “Φθάνουμε στα [loc Σπάτα /loc] / We arrive at [loc Spata /loc]”, ΄γεννήθηκε στην [loc Αθήνα /loc] / he was born in [loc Athens /loc]”, etc. The first rule marks with “[loc”, “/loc]” brackets the following: one or more geographical designators (geosign) such as “βουνό / mountain, αεροδρόμιο / airport, οδός / street” etc., optionally followed by a definite article, optionally followed by a capitalized adjective or a capitalized adjective indicative of location (locadj_cap) such as “βόρειος/north, νότιος/south” etc., optionally followed by an abbreviation, followed by one or more person NE’s and/or location NE’s and/or capitalized words, optionally followed by a digit. In this first rule the context of the NE is not taken into account. In the second rule, one or more capitalized words and/or capitalized foreign words are recognized as of type “location” only if they are preceded by a verb indicative of location such as “φθάνω/arrive” or a geosign followed by the Greek compound preposition “στον/[at, in]”. [const(X,Y) is a macro for strings starting with X and ending with Y].

System development follows an iterative process. After each run, a Java graphical interface is used by the developer to view named entities spotted in the text. The interface identifies differences between automatically and manually recognized NE’s and calculates precision and recall figures for each NE category. This facilitates fast NE grammar development.

## 6 Evaluation

30,000 words of the manually annotated corpus were used solely for evaluation. The performance of the system for each NE type is shown in Figure 2. There are no benchmarks for NE's of type time since only two time expressions appear in the test corpus. Figure 3 displays the error distribution over common error sources.

A significant number of errors (18.4%) are due to preprocessing (tokenization, tagging, lemmatization). As can be seen, the system did not perform particularly well in recognizing persons. 46% of errors in recognizing persons are due to preprocessing. For example, a sentence delimiter is sometimes inserted after initials which are naturally followed by periods. The same can happen in organization and location names containing abbreviations. Ambiguity between certain NE types (usually person – organization and location – organization) in the absence of clarifying context is a usual source of errors credited to the NERC stage itself.

We have taken action to deal with problems in preprocessing, as well as expand the NERC module so as to increase recognition performance per se. This includes fine-tuning the preprocessing chain, tailoring some aspects of preprocessing to NERC, expanding the NERC module to take into account gender information, and incorporating an NE cache. For example, let's consider the following:

```
Η [person Γερμανός / person] εξέδωσε 1.000.000 νέες
μετοχές. / [person Germanos / person] issued 1,000,000
new shares.
```

"Γερμανός" is both a person name and a company name. Here, it was mistakenly recognized as person. Ambiguity could have correctly been resolved, if gender information were taken into account. Article "Η" is feminine whereas "Γερμανός" is masculine. This seeming violation of agreement (ellipsis in fact) could have been exploited to correctly raise the ambiguity. Furthermore, the NERC module is expanded with the incorporation of an NE cache storing instances of already recognized/classified names. This will facilitate the recognition of NE's which have been encountered and classified in other parts of the text.

Spelling mistakes account for another 11.5% of the errors. There are also words with letters from both sets, since some letters are shared by the Greek and Latin alphabets, but a script is used to map characters to the appropriate character set.

When no lists of known NE's (persons, organizations, locations) were used at the lookup stage, performance dropped dramatically. Precision and Recall figures are given in Figure 4.

| NE Type | Precision | Recall | F-Measure |
|---|---|---|---|
| **Person** | 0.71 | 0.71 | 0,71 |
| **Loc** | 0.85 | 0.82 | 0,83 |
| **Org** | 0.80 | 0.72 | 0,76 |
| **Money** | 0.99 | 0.95 | 0,97 |
| **Percent** | 1.00 | 0.98 | 0,99 |
| **Date** | 0.89 | 0.84 | 0,86 |
| **Time** | ? | ? | 0 |
| **Total** | 0.86 | 0.81 | 0,83 |

**Figure 2:** NERC Performance

| NE Type | Error Distribution | | | |
|---|---|---|---|---|
| | % of errors due to preprocess-ing | % of errors due to spelling | % of errors due to ambiguity | Other |
| **Person** | 46.0% | 00.0% | 12.6% | 41.4% |
| **Loc** | 12.5% | 00.0% | 09.7% | 77.8% |
| **Org** | 09.2% | 15.7% | 06.1% | 69.0% |
| **Date** | 15.1% | 23.3% | 00.0% | 61.6% |
| **Money** | 81.9% | 00.0% | 00.0% | 18.1% |
| **Total** | 18.4% | 11.5% | 06.6% | 63.5% |

**Figure** 3: Distribution of Error

| NE Type | Precision | Recall | F-Measure |
|---|---|---|---|
| **Person** | 0.80 | 0.34 | 0.47 |
| **Org** | 0.77 | 0.36 | 0.49 |
| **Loc** | 0.82 | 0.14 | 0.23 |
| **Date** | 0.89 | 0.84 | 0.86 |
| **Money** | 0.99 | 0.95 | 0.96 |
| **Percent** | 1.00 | 0.98 | 0.98 |
| **Total** | 0.75 | 0.45 | 0.56 |

**Figure 4:** NERC performance when the name lookup is omitted

## 7 Conclusion

In this paper, we presented a Greek named entity recognizer oriented towards large scale information extraction applications. We implemented finite state techniques favoring efficient text processing and adopted a modular design allowing fast customization to the needs and particularities of specific applications. We also carried out an elaborate evaluation of the system's output and identified the design and implementation aspects we should enhance. Since work is still in progress, we expect that benchmarks will further improve; the system, however, has already reached a level of performance (F=83%) which is satisfying for many real-world applications.

## References


1. Aberdeen J., Burger J., Day D., Hirschman L., Robinson P., Vilain M. 1995. Mitre: description of the Alembic system used for MUC-6. Proceedings of Sixth Message Understanding Conference (1995)
2. Bikel D., Miller S., Schwartz R., Weischedel R.. Nymble: a high-performance learning name-finder, Conference on Applied Natural Language Processing (1997)
3. Black W., Rinaldi F., Mowatt D. Facile: description of the NE system used for MUC-7. Proceedings of Seventh Message Understanding Conference (1998)
4. Borthwick A., Sterling J., Agichtein E., Grishman R. 1997. Description of the MENE Named Entity System as used in MUC-7. Proceedings of Seventh Message Understanding Conference (1998)
5. Brill E. A corpus-based approach to language learning. Doctoral Dissertation, Univ. of Pennsylvania (1993)
6. Chinchor N., MUC-7 Named Entity Task Definition, Version 3.5 (1997)
7. Cowie J. 1995. Description of the CLR/NMSU systems used for MUC-6. Proceedings of Sixth Message Understanding Conference (1995)
8. Di Christo, P., S. Harie, C. De Loupy, N. Ide, and J. Veronis. Set of programs for segmentation and lexical look up, MULTEXT LRE 62-050 project Deliverable 2.2.1 (1995)
9. Gaizauskas R., Wakao T., Humphreys K., Cunningham H., Wilks Y. 1995. University of Sheffield: Description of the LaSIE system as used for MUC-6. Proceedings of Sixth Message Understanding Conference (1995)
10. Gallippi A., Learning to recognize names across languages. Proceedings of the 16th International Conference on Computational Linguistics (1996)
11. Grishman R. 1995. The NYU system for MUC-6 or where's the syntax. Proceedings of Sixth Message Understanding Conference (1995)
12. Grishman R., Tipster architecture design document version 2.3. Technical report, DARPA (1997)
13. Karkaletsis V., Spyropoulos C., Petasis G. Named entity recognition from Greek texts: the GIE project (1999)
14. Karttunnen L., The Replace Operator. In Finite State Language Processing, ed. Roche Em. and Schabes Yv., MIT Press (1997)
15. Krupka G., Hausman K. IsoQuest: description of the NetOwl extractor system as used for MUC-7. Proceedings of Seventh Message Understanding Conference (1998)
16. Mikheev A., Grover C., Moens M. 1997. Description of the LTG System used for MUC-7. Proceedings of Seventh Message Understanding Conference (1998)

17. Neumann G., Backofen R., Baur J., Becker M., Braun C. 1997. An information extraction core system for real world German text processing. ACL (1997)
18. Sekine S., Grishman R., Shinnou H.. A decision tree method for finding and classifying names in Japanese texts, Sixth Workshop on Very Large Corpora (1998)
19. Sekine S. NYU: description of the Japanese NE system used for MET-2. Proceedings of Seventh Message Understanding Conference (1998)
20. Yu S., Bai S., Wu P. Description of the Kent Ridge Digital Labs system used for MUC-7. Proceedings of Seventh Message Understanding Conference (1998)
21. Van Noord Gertjan and Dale Gerdemann. An Extendible Regular Expression Compiler for Finite-state Approaches in Natural Language Processing. WIA, Potsdam, Germany (1999)