NERosetta – an Insight into Named Entity Tagging

Cvetana Krstev[1], Anđelka Zečević[2], Duško Vitas[2], Tita Kyriacopoulou[3]

[1] University of Belgrade, Faculty of Philology, Studentski trg 3, Belgrade, cvetana@matf.bg.ac.rs
[2] University of Belgrade, Faculty of Mathematics, Studentski trg 16, Belgrade, vitas@matf.bg.ac.rs
[3] Université Paris-Est, Laboratoire d'informatique Gaspard-Monge, Marne-la-Vallée, tita@univ-mlv.fr

Abstract

Named Entity Recognition has been a hot topic in Natural Language Processing for more than fifteen years. A number of systems for various languages have been developed using different approaches and based on different named entity schemes and tagging strategies. We present the *NERosetta* web application that can be used for comparison of these various approaches applied to aligned texts (bitexts). In order to illustrate its functionalities, we have used one literary text, its 7 bitexts involving 5 languages and 5 different NER systems. We present some preliminary results and give guidelines for further development.

Keywords: aligned texts, named-entity recognition, named-entity scheme, META-NET

1. Motivation

Named Entity Recognition (NER) has been a hot topic in the Natural Language Processing (NLP) community for more than fifteen years. Ever since their introduction at the Sixth Message Understanding Conference (Grishman, 1996), named entities have been attracting interest of developers of various NLP applications. A comprehensive overview of NER research literature is presented in (Tatar and Cicekli, 2011). The authors point out that although most of the research in this field is still being done for English, it can be observed that this task draws attention of the research community at large, and not only of those dealing with the languages well-provided with NLP resources and tools, like German, French, Dutch and Spanish. Moreover, the languages for which some or substantial work in the NER field has been reported belong to various language families.

Another piece of evidence concerning the level of achievement in the NER field can be found in the Language White Papers Series, produced as part of the METANET project (Rehm and Uszkoreit, 2012). According to this source, at least some information extraction tools and applications (comprising NER) exist for 28 out of 30 analyzed European languages.

According to (Nadeau and Sekine, 2009), the term "Named Entity" usually refers to names of persons, locations and organizations, and numeric expressions including, time, date, money and percentage. In the last decade, this definition of basic named entities has often been redefined and refined, mostly by adding new a few major types, like "products" and "events", and several marginal types, like "e-mail addresses" and "book titles".

The majority of NER systems are monolingual systems developed for a particular language. Consequently, most of them rely on the language resources of that language and/or language-dependant methods. Recently, a number of authors have reported on multilingual NER applications (Steinberger and Bruno, 2009).

The applied methods vary from handcrafted rule-based systems that rely heavily on linguistic knowledge to machine-learning techniques. The usual approaches in machine-learning are supervised learning (Tatar and Cicekli, 2011), semi-supervised learning (Liu et al., 2011) and unsupervised learning (Nadeau et al., 2006). Rule-based systems usually rely on large-scale lexical resources and grammars, often in the form of regular expressions or FSTs (Maurel et al., 2011). Some authors benefit from combining rule-based and machine-learning approaches when developing their NER systems (Béchet et al., 2011).

In Section 2, we will discuss the variations between different NE schemes and tagging techniques that make the comparison between them difficult. In Section 3, we will present a new web application, *NERosetta* that enables a comparison of different NER approaches, while in Sections 4 and 5, we will give the first results obtained for one text in several languages tagged using several NER systems. Finally, in Section 6, we will give some concluding remarks.

2. Variety of Named Entity Schemes and Tagging Strategies

NER systems apply various approaches when defining a named entity structure, ranging from those offering just a few types, like those proposed in the MUC-6 task: ENAMEX, TIMEX, NUMEX, each having just a few attributes for further refinement (Chinchor, 1995), to those offering a named entity hierarchy which includes as many as two hundred different types (Sekine and Nobata, 2004). Useful guidelines on how to tag named entities in texts are also given in Chapter 13 of TEI Guidelines P5 (Burnard and Bauman, 2008). A balanced named entity structure in that respect is defined in the Quareo project (Rosset et al., 2011). The hierarchy of named entities in the Quaero project consists of eight top-level types: PERSONS, FUNCTIONS, ORGANIZATIONS, LOCATIONS, HUMAN PRODUCTS, AMOUNTS, TEMPORAL EXPRESSIONS, and EVENTS. All of these types, except AMOUNTS and EVENTS have one or two levels of sub-types.

Various named entity schemes differ not only by the basic set of entities they take into consideration, but also as to whether these entities are refined and to what extent.

For instance, the Quaero NE structure has only the type AMOUNT that does not have sub-types. However, many NE schemes distinguish various sub-types: MONEY, MEASUREMENT, PERCENTAGE, etc. Moreover, two-level sub-types are sometimes distinguished, like the NE scheme presented in (Maurel *et al*., 2011) in which the AMOUNT type has two sub-types – MONEY (called *valeur monétaire*) and MEASUREMENT (called *valeur physique*) – while the latter has nine sub-subtypes: for duration, temperature, distance, etc.

The same example can be used to illustrate one more difference between the various NE structures. Namely, a sub-type can be present in two NE schemes, but as a sub-type of different types. For instance, DURATION is a sub-type of MEASUREMENT, and thus AMOUNT both in (Maurel *et al*., 2011) and Quaero. The NE scheme presented in (Krstev *et al.*, 2013) relies on the international standard for semantic annotation of time and events (Pustejovsky et al., 2010), and therefore, DURATION is a sub-type of TIME, along with DATE, HOUR and SET (sets of time). Similarly, in TEI Guidelines P5 (Burnard and Bauman, 2008), DURATION is connected to TIME, but it is not at the same level as DATE and HOUR; it is rather their sub-type, since it is introduced as an attribute of the corresponding elements.

Differences in NE schemes naturally reflect the way NEs are tagged in texts. Named entity tagging proposed for the European newspapers project[1] introduces only three entities: PERSON, ORGANIZATION and LOCATION, but does not allow nested tagging, e.g. *Canada* is not tagged as a location in the organization name *Library and Archives Canada*. The same strategy is not applied in (Rosset et al., 2011), (Maurel *et al*., 2011), (Krstev *et al.*, 2013), where several levels of nesting can actually occur in a text.

The similar applies to the span of a named entity. According to the European newspapers project, titles such as *Dr.* are not part of the PERSON NE, while some other strategies include them (Burnard and Bauman, 2008), (Rosset et al., 2011), (Maurel *et al*., 2011), (Krstev *et al.*, 2013). The other aspect of tagging is connectedness of tagged entities: for instance, (Rosset et al., 2011), (Maurel *et al*., 2011) treat the FUNCTION (a person holds) as separate from a person's NAME, both of which can be connected by the PERSON entity, but can be tagged separately, as well. One example from the Quaero project illustrates this:

```
<pers.ind>
    <title>
        Son Altesse Royal le
        <func.ind>prince</func.ind>
    </title>
    <name.ind>Rainier</name.first>
</pers.ind>
Versus
Le
<func.ind>roi</func.roi>
<pers.ind>
    <name.first>Mohamed</name.first>
    <qualifier>VI</qualifier>
<pers.ind>
```

In the strategy used by (Krstev *et al.*, 2013) a person's NAME and FUNCTION are both mandatory part of the content of the tag PERSON.

In some strategies, named entities can be specified hierarchically, as illustrated by an example from TEI Guidelines (Burnard and Bauman, 2008):

```
<orgName>
    <orgName>Department of Modern
    History</orgName>
    <orgName>
        <name type="city">Glasgow</name>
        <name type="role">University</name>
    </orgName>
</orgName>
```

The same approach is taken by the Quaero project and (Krstev *et al.*, 2013); naturally, such a possibility cannot occur in the strategies that do not allow nesting, as in the European newspaper project.

Finally, NE tagging strategies differ in determining the semantic scope of each NE type and sub-type. It is interesting that both Quaero and the European newspaper projects give *The Beatles* as an example: the former strategy treats it is as a collective PERSON, while the latter treats it as ORGANIZATION.

3. Comparison of Named Entity Schemes and Tagging Strategies

In order to be able to compare various NE structures and tagging strategies primarily qualitatively, we have developed a web application dubbed *NERosetta*[2]. The comparison is performed through a reference NE scheme. At present, our reference scheme relies mostly on the Quaero project, with addition of some sub-types, e.g. for AMOUNT. We have chosen this particular scheme for several reasons: (a) it has quite an elaborate and balanced structure with respect to NE types and sub-types; (b) it is well documented; (c) the NER systems that were at our disposal for testing were developed independently from it. This reference NE scheme can be easily replaced by another one.

NERosetta does not perform NE recognition and tagging, rather it works with the documents previously tagged with some NER system. It can accept any document as long as it is a well-formed XML document with NEs tagged with XML tags. It is also presupposed that a document is segmented into paragraphs and sentences (or segments). In order to work with it, it is necessary to define the NE structure used, which is done by defining the mapping to the reference scheme, for instance: **amount.money** → money.exact (*70.000 miliona evra*) and **amount.money** → money.range (*(između) 5,5 i šest miliona dolara*). As a rule, each mapping is followed by an example. It should be noted, however, that this is performed only once for each structure and need not be repeated for each document in the system that uses it.

In general, the established mapping between the newly defined NE scheme and the reference scheme is many-to-many.

[1] http://www.europeana-newspapers.eu/focus-on-newspaper-refinement-quality-assessment-in-belgrade/

[2] http://www.korpus.matf.bg.ac.rs/nerosetta/

1. In some cases, an entity type or a sub-type is mapped to one and only one type or a sub-type in the reference structure. Such is the case with the type MONEY in the NE scheme Stanford NER 7 (Finkel et al., 2005), which is mapped only to the sub-type AMOUNT.MONEY in the reference NE structure and vice versa.
2. There are cases where an entity type or a sub-type is mapped to more than one type or sub-type in the reference scheme. Such is the case with the type ORGANIZATION in the NE scheme Stanford NER 7, which is mapped to two sub-types in the reference NE structure: ORG.ADM and ORG.ENT.
3. Finally, there are cases where several entity sub-types are mapped to one type or sub-type in the reference scheme. Such is the case with the various measurement sub-types in the NE scheme described in (Maurel *et al.*, 2011): AMOUNT.PHY.DUR, AMOUNT.PHY.TEMP, AMOUNT.PHY.LEN, etc. that are all mapped to one sub-type in the reference NE scheme: AMOUNT.MEASURE.

NERosetta has three different modes. It can work with: (a) one particular document tagged with one NER system; (b) one document tagged with two different NER systems; (c) one document in two languages (e.g. source and translation language), each tagged with some NER system developed for the corresponding language. For working in mode (c) it is necessary for the two texts to have been previously aligned at the segment level (paragraphs, sentences, or sub-sentence). The format of alignment is described in Section 16.4.2 of the TEI P5 Guidelines (Burnard and Bauman, 2008). It basically relies on two separate files for each language in which all segments that should be aligned are labeled by unique identifiers. A third file consists of a group that links the corresponding segments. Such a format can be produced, for instance by XAlign[3]. However, it is also possible to import an aligned text in a TMX (translation memory interchange) format and *NERosetta* will split it in three files that are in the requested format.

The search is performed only by the types and sub-types of the reference scheme. Before formulating the search, a user has to decide whether she/he is looking for the exact match, the sub-types of the chosen type, and/or the super-types of the chosen type. These options are necessary in order to overcome different kinds of mappings. If only the exact match were supported, then only case 1 mapping would give the expected results. However, if one NE structure supported only a top level type (e.g. LOCATION in Stanford NER 7) and the other supported sub-types of the reference scheme (cases 2 and 3 of the mapping) the exact match would yield no results, but the option "match sub-types" solves the problem. The option "match super-types" is adequate if just some sub-types are to be included in a search. A user can select one or more types and sub-types and they are combined in a conjunction query. All the results obtained are preceded by the chosen search criteria and the corresponding mapping, so that a user can better understand the results. The aligned concordances are presented ten at a time.

A user can work with *NERosetta* at two levels. As an unregistered user, she/he can view all applied NE schemas and search the available textual resources. If a user becomes registered, then she/he can also define new NE schemas, upload new texts and/or new tagging, and delete her/his own texts. The "How to Use *NERosetta*" document is put at the disposal of all users.

The application is organized as a classic 3-tier web application: the first, presentation tier, reflects user information needs and performs data visualization; the second, logic tier is most complex one, as it serves in both directions, extracting input data and sending back the obtained results; the third, data tier is where data is stored and retrieved from a data-base system. Due to potentially big XML files that correspond to data and linking instructions, its content is not loaded into memory at once, but a stream-based parser is used. It requires knowledge of file character encoding so a user performing a file upload should provide the system with the appropriate information. As mentioned above, the application is developed with a wide set of users in mind. All the changes, such as a new file upload or addition of a new NER schema description are immediately visible to the public and ready for use.

4. Verne's novel "Around the World in 80 Days" in *NERosetta*

For the first application of *NERosetta* we have chosen Verne's novel Around the World in 80 Days" which is available through the META-SHARE repository in 18 languages – the French original and 17 translations[4]. Moreover, the 32 parallel versions aligned 1:1 at the sentence level are also available in the TMX format (4436 segments in all bitexts). The availability of this text in many languages would not have been enough if it had not been suitable for experimenting with named entity tagging, because of the nature of the text itself.

However, we restricted this experiment only to those languages for which some NER system was at our disposal. The languages we covered and the NER systems we used are:

- *French* – we have used the frMaurel system described in (Maurel *et al.*, 2011).[5] This system is based on a cascade of finite-state transducers and e-dictionaries and it is implemented in the Unitex corpus processing tool.[6] It recognizes the following top-level types: AMOUNT, EVENT, FUNCTION, LOCATION, ORGANIZATION, PERSON, PRODUCT, TIME.
- *English* – we have used Stanford NER 3 and NER 7 classifiers. The classifiers are statistical in nature and based on conditional random field information extraction systems boosted with the Gibbs sampling technique for incorporating non-local information. The algorithm is described in depth in (Finkel at al., 2005), while the most up-to-date versions can be downloaded from the official website.[7] The first classifier labels PERSON, LOCATION and ORGANIZATION, while the second

[3] A tool developed by P. Bonhomme, T. M. H. Nguyen and S. O'Rourke, http://led.loria.fr/outils/ALIGN/align.html.

[4] http://www.meta-share.eu/

[5] We have used the version of the cascade and e-dictionaries from February 2012.

[6] http://www-igm.univ-mlv.fr/~unitex/

[7] http://nlp.stanford.edu/software/CRF-NER.shtml

covers a wider set of types: PERSON, LOCATION, ORGANIZATION, TIME, DATE, MONEY and PERCENT.

- *Serbian* – we have used the srKrstev system, described in (Krstev *et al*., 2013). This system is also based on a cascade of finite-state transducers and e-dictionaries. It recognizes the following top-level types: AMOUNT, LOCATION, ORGANIZATION, PERSON, TIME. It can be used not only in the Unitex environment, but also through the *NERanka*[8] web interface. We have applied the NER system developed for Croatian (described in the next item) to Serbian too. The reasons for this were twofold: firstly, Serbian and Croatian belong to the same South-Slavic family and are therefore often regarded as closely related, and secondly, NER systems for Serbian and Croatian were developed on different principles and thus are interesting to compare.
- *Croatian* – we have used the system described in (Ljubešić at al., 2013) here denoted as hrNER. It is based on the Stanford named entity recognizer and adapted to the specific task of Croatian and Slovenian NER. The hrNER that we downloaded from the official website[9] tags three top-level types: LOCATION, ORGANIZATION and PERSON. For the reasons explained above, we have applied the NER system developed for Serbian to Croatian as well.
- *Greek* – we have used the grTita NER system presented in (Kyriacopoulou *et al.*, 2011). This system is also based on finite-state transducers and e-dictionaries applied in Unitex. The Greek NER system tags only two top-level types: PERSON and TIME.

For these five languages and the NER systems described we have produced the pairs listed in Table 1.

	Lang.	**NER**	**Lang.**	**NER**
A	French	frMaurel	English	Stanford-7
B	French	frMaurel	Serbian	srKrstev
C	English	Stanford-7	Serbian	srKrstev
D	Greek	grTita	English	Stanford-3
E	Croatian	hrNER	Serbian	srKrstev
F	Croatian	hrNER	Serbian	hrNER
G	Serbian	srKrstev	Croatian	srKrstev

Table 1. Language/NER system pairs in *NERosetta*

5. Qualitative and quantitative analysis of NER with *NERosetta*

The main purpose of *NERosetta* is to facilitate comparison between various NE schemes, strategies and systems. Additionally, it can provide insight into translation strategies for various types of named entities, as illustrated by the following example of ORGANIZATION recognition (tagged sequences are given in bold):

FR: Rowan, directeur police, administration centrale, Scotland place.

CR: Rowanu, upravitelju redarstva, središnja uprava, **Scotland yard**.

SR: Rovanu, upravniku policije, centralna uprava, trg Skotland.

EN: Rowan, **Commissioner of Police**, Scotland Yard:

[8] http://hlt.rgf.bg.ac.rs/VebRanka/NERanka.aspx

[9] http://nlp.ffzg.hr/resources/models/ner/

Pair	**in 500**	**in both**	**in 1st**	**in 2nd**
A	127	24	24+6	24+97
%	14.60	18.90	23.62	95.28
B	156	21	21+9	21+126
%	15.00	13.46	**19.23**	94.23
C	163	103	103+18	103+42
%	14.17	63.19	74.23	88.96
D	212	131	131+70	131+11
%	12.85	61.79	**94.81**	**66.98**
E	167	69	69+22	69+76
%	14.15	41.32	**54.49**	**86.82**
F	98	29	29+61	29+8
%	16.61	29.59	91.83	**37.76**
G	148	14	14+132	14+2
%	14.23	9.46	98.65	**10.81**

Table 2. The number of concordance lines for PERSON (with sub-categories) in the first 500 segments of the text – the lowest number of concordance lines is given in bold

NERosetta can be useful for quantitative assessment of different NER systems, namely for the computation of the precision (as the ratio of all correctly retrieved NEs and all retrieved NEs), the recall (as the ratio of all correctly retrieved NEs and all NEs in a text) and the F-measure. To illustrate this, we have analyzed the tagging results obtained by various NER systems presented in Section 4 using the type PERSON as an example, because of its being the only type presented in all the systems used. *NERosetta* produced from 590 concordance lines (for the pair F) to 1,650 lines (for the pair D). We have restricted our research only to the first 500 segments of each bitext. Data in Table 2 was computed automatically by *NERosetta*: it gives the number of concordance lines in which at least one PERSON named entity was recognized, correctly or not, in the first language of the pair, in the second language and in both.

In order to compute the precision it is enough to investigate concordance lines obtained for any pair in which a chosen NER system applied to a language appear. For instance, in order to determine the precision of srKrstev applied to Serbian one could look at the concordances for B, C, E or F – the results would always be the same. We counted all PERSON NEs that were retrieved in an aligned segment.

	Lang/ NER	**Recall**	**Lang/ NER**	**Recall**
A	fr/frMaurel	20.00	en/Stanford-7	97.87
B	fr/frMaurel	**17.82**	sr/srKrstev	90.06
C	en/Stanford-7	92.62	sr/srKrstev	**86.67**
D	gr/grTita	**92.52**	en/Stanford-7	**79.90**
E	cr/hrNER	**50.00**	sr/srKrstev	87.15
F	cr/hrNER	84.91	sr/hrNER	**37.38**
G	sr/srKrstev	96.27	cr/srKrstev	**9.94**

Table 3. The recall for all Language/NER system pairs in *NERosetta* – the upper bound of the recall is in bold

The recall of the application of a NER system to a language cannot be determined so easily unless we have a version of a text that can serve as a "gold standard", which was not the case for our example text. Therefore, we have calculated the recall for a NER system applied to

a language for all pairs in *NERosetta*. The results are presented in Table 3. The lowest value of the recall for a Language/NER system pair is the upper bound of the recall, that is, the recall calculated in comparison to the "golden standard" would be less than or equal to it.

The ranking of all NER systems applied to a language according to the F-measure is given in Table 4. The systems frMaurel and srKrstev (on the Croatian text) have the low rank because they had problems with the recognition of foreign personal names. The system grTita has the high rank but its recall was estimated on the basis of just one aligned pair. On the other hand, srKrstev applied to the Serbian text has the highest rank and the recall was estimated on the basis of four aligned pairs; however, it was not compared to grTita that has the highest upper bound of the recall. Thus, the results are provisional. The final values can be obtained either by comparing the NER systems to the "golden standard", or at least by taking into account all language/NER system pairs (35 of them in our case).

	Precision	**Recall <**	**F <**
sr/srKrstev	0,950617	0,866667	0,906703
gr/grTita	0,845494	0,925234	0,883569
en/enNER-7	0,985612	0,798995	0,882546
hr/hrNER	0,9	0,5	0,642857
sr/hrNER	0,97561	0,373832	0,540541
fr/frMaurel	0,967742	0,178161	0,300922
hr/srKrstev	0,888889	0,099379	0,178771

Table 4. The precision, the recall and the F-measure for all Language/NER system pairs for PERSON NE

Nevertheless, the obtained results lead to formulation of two hypotheses: (1) The recall of a Language/NER system is inversely proportional to the recall of another Language/NER system in the aligned pair; (2) The lowest recall of a Language/NER system in the aligned pair corresponds to the lowest number of concordance lines for a Language/NER system in the aligned pair.

6. Future work

We have presented the first results of comparison of the performance of several NER systems. These results are far from conclusive – they were presented to point to the usefulness of *NERosetta* for qualitative and quantitative analysis, rather than to provide a detailed analysis of these systems. Our plans for the future are to have more users of *NERosetta*, and to work with more texts, languages and NER systems that would enable us to confirm or dismiss the formulated hypotheses.

Acknowledgments

This research was conducted through the project 178006 financed by the Serbian Ministry of Education, Science and Technological Development.

References

Béchet F., Sagot B., Stern R. et al. (2011) Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées. In : *TALN'2011-Traitement Automatique des Langues Naturelles*.

Burnard L., Bauman S. (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*.

Chinchor N. (1995) *MUC-6 Named Entity Task Definition (Version 2.1)*.

Finkel J. R., Grenager T., Manning C. (2005) Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: *Proc. of the 43nd Annual Meeting of the ACL,* p. 363-370.

Grishman R., Sundheim B. (1996) Message Understanding Conference-6: A Brief History. In: COLING, vol. 1, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 466-471.

Krstev C., Obradović I., Utvić M., Vitas D. (2013) A system for named entity recognition based on local grammars. *Journal of Logic and Computation*. doi:10.1093/logcom/exs079

Kyriacopoulou T., Martineau C., Mavropoulos T. (2011) Les noms propres de personne en français et en grec : reconnaissance, extraction et enrichissement de dictionnaire. In: Proc. of the 30th Conference on Lexis and Grammar – LGC'2011, Cyprus.

Liu X., Zhang S., Wei F., Zhou M. (2011) Recognizing named entities in tweets. In: Proc. of the 49th Annual Meeting of the ACL: Human Language Technologies, vol. 1, p. 359-367.

Ljubešić N., Stupar M., Jurić T., Agić, Ž. (2013.) Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. In *Slovenščina 2.0: empirical, applied and interdisciplinary research*, in press.

Maurel D., Friburger N., Antoine J.-Y., Eshkol I., Nouvel D. et al. (2011) Cascades de transducteurs autour de la reconnaissance des entités nommées, *Traitement Automatique des Langues*, vol. 52, no 1, p. 69-96.

Nadeau D., Turney P., Matwin S. (2006) Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity, In: 19th Canadian Conference on Artificial Intelligence, Québec City, Québec, Canada.

Nadeau D., Sekine S. (2009) A Survey of Named Entity Recognition and Classification, In S. Sekine, E. Ranchhod (eds), *Named Entities: Recognition, Classification and Use*, John Benjamins Pub. Co., Amsterdam/Philadelphia, p. 3-28.

Pustejovsky J., Lee K., Bunt H., Romary L. (2010) ISO-TimeML: An International Standard for Semantic Annotation. *7rd LREC 2010*, ELRA, Valletta, Malta.

Rehm G., Uszkoreit H. (eds) (2012) *META-NET White Paper Series*, Springer.

Rosset S., Grouin C., Zweigenbaum P. (2011) *Entités Nommées Structurées: guide d'annotation Quaero*. LIMSI–CNRS, Orsay, France.

Sekine S., Nobata C. (2004) Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy, *LREC*, Lisbon, Portugal.

Steinberger R., Bruno P. (2009) Cross/lingual Named Entity Recognition. In S. Sekine, E. Ranchhod (eds), *Named Entities: Recognition, Classification and Use*, John Benjamins Pub. Co., p. 137-164, 2009.

Tatar S., Cicekli I. (2011) Automatic rule learning exploiting morphological features for named entity recognition in Turkish, J. Information Science, vol. 37, no 2, p. 137-151.