

Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation Model

Héla Fehri

Laboratory MIRACL, University of Sfax
Route Tunis Km 10 B.P 242, Sakiet Ezziat
3021 Sfax

hela.fehri@fss.rnu.tn

Kais Haddar

Laboratory MIRACL, University of Sfax
Route Tunis Km 10 B.P 242, Sakiet Ezziat
3021 Sfax

kais.haddar@fss.rnu.tn

Abdelmajid Ben Hamadou

Laboratory MIRACL, University of Sfax
Route Tunis Km 10 B.P 242, Sakiet Ezziat 3021 Sfax

abdelmajid.benhamadou@isimsf.rnu.tn

Abstract

Recognition and translation of named entities (NEs) are two current research topics with regard to the proliferation of electronic documents exchanged through the Internet. The need to assimilate these documents through NLP tools has become necessary and interesting. Moreover, the formal or semi-formal modeling of these NEs may intervene in both processes of recognition and translation. Indeed, the modeling task makes more reliable the constitution of linguistic resources, limits the impact of linguistic specificities and facilitates transformations from one representation to another. In this context, we propose an approach of recognition and translation based on a representation model of Arabic NEs and a set of transducers resolving morphological and syntactical phenomena. The representation model is based on the feature structure independent of lexical categories.

Keywords: Representation Model, NE' recognition, NE' translation, Transducer, Local grammar.

1 Introduction

Recognition and translation of NEs are two current research topics with regard to the proliferation of electronic documents exchanged through the Internet. The need to assimilate these documents through NLP tools has become necessary and interesting.

Furthermore, the formal or semi-formal modeling of NEs can be involved in recognition and translation processes. This modeling task allows the constitution of more reliable linguistic resources. Indeed, such a modeling can represent all the constituents of a NE in a standard manner and limit the impact of linguistic specificities. In fact, a formal representation of Arabic NEs can help, firstly, in the identification of dictionaries and grammars required for any NLP application and, secondly, in the use of advanced linguistic methods of translation (i.e., transfer or pivot method). This abstraction level favors the reuse of certain linguistic resources.

The elaboration of a formal and generic representation of an NE is not an easy task because, on the one hand, we have to find a representation that takes into consideration the concept of recursion and length of NE. In fact, a NE can be formed by other NEs. So, its length is not known in advance. On the other hand, the representation to be proposed should also contain a sufficient number of features that can represent any NE independently of the domain and grammatical category. In other words, the same features must satisfy all types of NE.

Transducers (eventually local grammars) can resolve many problems of the NEs and should respect hierarchy types of NEs. However the connection between these transducers to obtain a deep analysis isn't a trivial task. So, a linguistic platform like NooJ can help us to do this kind of analysis.

It is in this context that the present work is situated. In fact, the main objective is to propose an approach of recognition and translation of Arabic NEs based on a representation model, a set of bilingual dictionaries and a set of transducers resolving morphological and syntactical phenomena related to the Arabic NEs and implemented with the linguistic platform NooJ. To reach our objective, we have to offer a representation model that takes into account the notion of recursion of NEs. We have also to specify features that describe any NE independently of the domain. Finally, the representation should be compatible with the linguistic platform NooJ (Silberztein, 2004) chosen for the implementation.

In this paper, we present, firstly, a brief overview of the state-of-the art. Then, we detail our proposed representation model. After that, we give a general idea of our resources construction and their implementation in the linguistic platform NooJ. Finally, the paper concludes with some perspectives.

2 Related Work

Research on NEs revolves around two complementary axes: the first involves the typing of NEs while the second concerns the identification and translation of NEs. As for the identification, the tagging and the translation of NEs, they have been implemented for multiple

languages based on different approaches: linguistic (Coates-Stephens, 1993), statistic (Borthwick et al., 1998) and hybrid (Mikheev et al., 1998) approaches. In what follows, we focus on the linguistic approach used for NE processing.

Regarding the recognition of NEs, we cite the work presented in (Friburger, 2002). This work allows the extraction of proper names in French. The proposed method is based on multiple syntactic transformations and some priorities that are implemented with transducers. We can cite also the work described in (Mesfar, 2007). The elaborated method is applied on a biomedical domain. Other Arabic works are dealing with the recognition of elliptical expressions (Hasni et al., 2009), compound nouns (Khalfallah et al., 2009), broken plurals (Ellouze et al., 2009) and most important categories in Arabic script (Shaalán and Raza, 2009). All these works that use the linguistic platform NooJ can be integrated in the NE processing.

Other works have been dedicated to the translation of different structure (e.g., NE) from one language to another. We can cite the work presented in (Barreiro, 2008) dealing with the translation of simple sentences from English to Portuguese. Additionally, the work of (Wu, 2008) provides a noun translation of French into Chinese. The elaborated prototype tests a limited corpus of 600 French nouns and is experimented with NooJ.

The literature review shows that the already proposed translation approaches are not well specified (e.g., lack of abstraction and genre). Each one addresses a particular phenomenon without taking into account other phenomena. We should also mention that there are few works that proposed a modeling of NEs for explicitly representing the effects of meaning within the NE and explaining phenomena like synecdoche and the metonymy (Poibeau, 2005). However, these works don't treat the concept of embedded NEs which is very important and can help to implement the recognition and the translation process of NEs.

Furthermore, all translations using NooJ platform adopt a semi-direct approach of translation, in which the recognition task is combined with that of translation. Thus, the reuse of such work has become limited, which does not promote multilingualism.

3 Proposed Model for Representing Arabic NEs

The model that we propose is used to formalize and to identify Arabic NEs. This model is inspired by formalisms based on structural features like Head-driven Phrase Structure Grammar (Pollard and Sag, 1994). Its features are inspired from the concepts "Head and Expansion" introduced by (Bourigault, 2002). The essential characteristics of the feature structure of the proposed model are:

- an element of the structure can be atomic or complex,
- an internal structure of an element is defined by its attributes and values.

In what follows, we describe the structure and features of our proposed model.

3.1 Structure and Features of the Proposed Model

Each NE has a type and is composed of two parts: one is essential and the other is extensional. The essential part is also a NE and has itself essential and extensional parts. This proves the recursion for an NE. The type of a NE "Type_EN" is usually indicated by a trigger word. The essential part is represented by the feature "Tête_EN" (head of NE) and the trigger word is represented by the feature "Mot déclencheur". The extensional part represents the final form that composes the NE. It does not admit a type because it is preceded by a lexical item "Element_EN" (element of NE) (e.g., preposition, special character). Then, it can not be considered as a NE but it can contain a NE. Its existence or non-existence doesn't affect the well-formation of the NE. This part is represented by the feature "Fin_EN" (end of NE).

The value of the feature "Tête_EN" can be atomic or structured. If it is structured, then it is composed by the features "Mot déclencheur", "Tête_EN", "Fin_EN" and "Type_EN". The "Mot déclencheur" value is simple or composed. Indeed, the trigger word can be formed by a word or a sequence of words. It can also be empty. The "Fin_EN" value can be atomic or structured. If it is structured, then it is composed by the features "Element_EN", "Tête_EN" and "Fin_EN". It can also be empty. The feature "Type_EN" value is always simple or composed but not empty. In fact, it represents one of the categories identified in the NE hierarchy. The "Element_EN" value is always

simple. The structure can be equipped with a set of principles allowing the well-formed construction and the evaluation of NE-representation.

3.2 Principles of the Proposed Model

For the presented model, two principles should be satisfied and are useful in the recognition phase. These principles are used to indicate whether a NE is well-formed or not.

Saturation Principle: A structure is called *saturated* if it can be considered as a well-formed NE. That means, it consists of a NE head ("Tête_EN") whose value is not empty. Figure 1 describes an example of a formal representation that satisfies a saturation principle.



Figure 1: Representation of the word الرياض *el Riadh*

In Figure 1, the value of the feature "Type_EN" is atomic; it is not empty. Thus, a word الرياض *Riadh* is considered as a NE whose type is *Ville*.

Non-saturation Principle: A structure is called *unsaturated* if it is not a NE and can be completed to become a NE. That means, it is formed only by a NE end ("Fin_EN") or if the value of the feature "Tête_EN" is empty. For example, in the word بالرياض *bi Riadh*, the value of the feature "Tête_EN" is empty because this word doesn't have a type. Thus, this word cannot be considered as a NE. It doesn't satisfy the saturation principle. However, it should be noted that this word can contain a NE.

The two mentioned principles allow us to avoid ambiguity between a NE-word (or set of words) and a non NE-word.

3.3 Illustrative Example

In this section, we give an example that explains how to construct NE representation. So, Figure 2 gives a formal representation of the NE ملعب الملك عبد العزيز الدولي بالرياض *Malaab el malik Abd el Aziz*

translation of the NE in Figure 3 (a) ملعب الملك عبد العزيز الدولي بالرياض Malaab el malik Abd el Aziz el doali bil Riadh gives in Figure 3 (b) the sequence of words "stade du roi Abd el Aziz international à Riadh" representing an ill-formed NE because it doesn't respect the specificities of the target language. Therefore, readjustment rules are necessary and should be associated in translation process.

4 NooJ Implementation of the Set of Transducers

The NooJ implementation of our system requires two phases process: recognition of Arabic NEs and translation in which a transliteration process is integrated. Each phase requires the construction of its proper transducers.

4.1 Phase of Recognition

The proposed representation model helps us to identify the necessary resources for the recognition and translation of NEs. In fact, each structured feature "Tête_EN" containing not empty features, other than the feature "Type_EN", is transformed into a local grammar (*place name, person name, ...*); whereas, each elementary NE (value of "Tête_EN" feature is atomic) will be transformed into a dictionary (*city name, ...*).

From the NE representation in the considered model, we have created the following transducer:

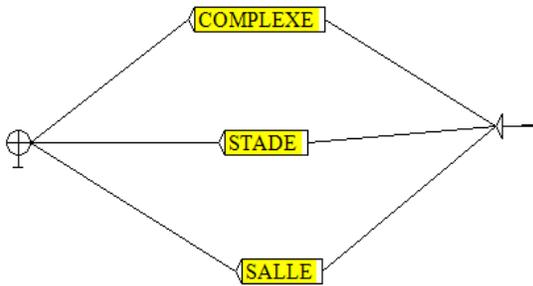


Figure 4: Main transducer of NE' recognition

The transducer of Figure 4 contains three sub-graphs. Each sub-graph represents a category identified in the NE hierarchy, especially in the category of *Place name*. This grammar allows recognition of NEs. Each path of each sub-graph represents a rule extracted in the study corpus.

In the recognition phase, we have solved the problems related to the Arabic language (e.g., agglutination) establishing morphological grammars built into the platform NooJ. This phase contains 19 graphs respecting the local grammars identified in the study corpus.

4.2 Phase of Translation

The implementation of the translation phase involves two steps as illustrated in Section 4.4: a step of word-to-word translation and another more detailed taking into account the readjustment rules and following the specificities of the target language (in our case the French language). In what follows, we describe resources designed to these two steps.

Word-to-word Translation: To implement the word-to-word translation in the platform NooJ, we built a syntactic grammar allowing the translation of each word composing a NE with the exception of words not found in dictionaries, or can not be translated (number, special character, etc.). This grammar takes as input the NE list extracted by the transducer of Figure 4 allowing the recognition. It is described by the transducer of Figure 5.

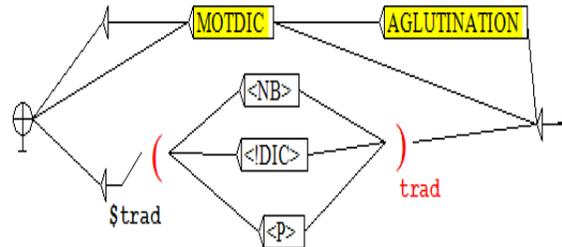


Figure 5: Transducer of word-to-word translation

The transducer of Figure 5 takes into account, in a NE, the words that keep the same values in the target language. These words can correspond to a number <NB>, a special character <P> or a word not existing in designed dictionaries <!DIC>. The sub-graph MOTDIC treats the rest of the words (existing in dictionaries) which require a specific treatment. For example, if the word to translate is a first name, then it keeps the same value in the source language; this word will be treated in the transliteration process (Fehri et al., 2009).

Translation with Readjustments: Several readjustment rules must be applied to improve the word-to-word translation step. These rules have

essentially a relationship with the order of the words composing a NE and with the agglutination. For instance, on the one hand, if a NE in the source language contains an adjective then we have to know whether this adjective belongs to the trigger word or to the noun that comes just before. For example, in the NE ملعب المدينة الدولي *malaab el madina el doali stadium of ground city*, the adjective الدولي *el doali ground* is singular and masculine, the trigger word ملعب *malaab stadium* is also singular and masculine, but the noun المدينة *el madina city* is singular and feminine. We can deduce that the adjective الدولي *el doali* belongs to the trigger word ملعب *malaab* and not to the noun المدينة *el madina* that comes just before. On the other hand, if a NE in the source language contains a noun then some rules are applied to solve the problem of contracted forms in Arabic. For example, in the NE ملعب الملك فهد *malaab el malik Fahd*, the noun الملك *el malik* is singular, masculine and definite by the letter ال *el*, so when we translate this noun we should add the preposition "du" before it. Thus, we obtain *stade du roi Fahd* (international stadium of Fahd) and not *stade roi Fahd* (international stadium Fahd).

Readjustment rules are made with syntactic local grammars in NooJ. These grammars intervene after the word-to-word translation phase. In what follows, we give an idea about transliteration process integrated in translation phase.

Transliteration process: The transliteration is done after having executed all the transducers allowing the NE' recognition and translation (word-to-word translation or with readjustments). In fact, it consists in transliterating all the non-translated words which are written in the source language (Arabic characters) using the appropriate resources. In this process, we consider rules respecting the chosen transliteration system *El Qalam* and also the transformation rules. These rules are implemented with NooJ morphological transducers. The transliteration is preceded by a vowelizing phase to avoid some problems. So, the word is vowelized before its transliteration. However, the connection between a vowel transducer and transliteration transducer can not be done in NooJ; that is why, we resort to use noojapply. Noojapply is a command-line program which can be called either directly from a "shell" script, or from more sophisticated programs

written in PERL, C++, JAVA, etc. In our work, we use C#. The transliteration step is detailed in (Fehri et al., 2009).

5 Experimentation and Evaluation

The experimentation of our resources is done with NooJ. As mentioned above, this platform uses (syntactic and morphological) local grammars already built. Table 1 gives an idea about dictionaries which we added to the resources of NooJ. In addition to the dictionaries mentioned in Table 1, we use other dictionaries existing in NooJ like dictionary of adjectives, nouns and first names (Mesfar, 2008).

Dictionaries	Number of inputs	Annotation in the dictionary
Player Names	18000	N+Joueur
Team Names	5785	N+Equipe
Sport Names	337	N+Sport
Capital and country Names	610	N+Toponym
Personality Names	300	N+Perso
Trigger words	20	N+Dec
Functions Names	100	N+Fonction

Table 1: Added dictionaries

To these dictionaries, we add some entries related to the sport domain. We also add French translations of all entries in all mentioned dictionaries. Let's note that the first name dictionary remains monolingual because its entries can be transliterated. To experiment and evaluate our work, we have applied our resources to two types of corpus: sport and education corpora. We started with sport domain since it is the subject of our study corpus.

5.1 Experimentation of Recognition Phase

To evaluate the recognition phase, we have applied our resources to a corpus formed by 4000 texts of sport domain (different of the study corpus). This corpus is collected from various newspapers (e.g., Assabah, al Anwar, al chourou9, al ahram) and Wikipedia. It contains 180000 NEs belonging to different categories of sport domain (e.g., player name, name of sport, sports term). In these

NEs, there are 40000 NEs belonging to the category *place name*. These NEs are manually identified using NooJ queries. The obtained result is given in the NooJ concordance table of Figure 6.

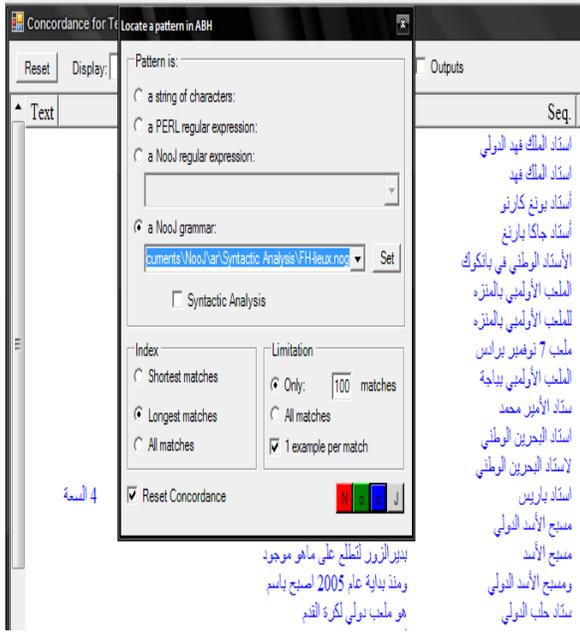


Figure 6: Concordance table of NE' recognition in the sport domain

Let's note that a NE is detected if it satisfies one of the paths described by the transducer of Figure 4. Indeed, a transducer is characterized by an initial node and one or many end nodes. If multiple paths are verified, we maintain the longest one. The obtained results are interpreted by calculating the following metrics: Precision, Recall and F-measure. Results are illustrated in Table 2.

The values of measures in Table 2 show that there are some problems that are not yet resolved. Some problems are related to the lack of standards for writing proper names (e.g., *el hamza*). This causes a silence. Other problems are related to Arabic specific concepts as metaphor. For example, we can find a NE in a text composed of a trigger word specific to the sport domain followed by a famous person name followed by a city as *ملعب الأسد بسوريا Malaab el Assad bi Souriya (stadium al Assad in Syria)*. The context where this NE appears is not in the objective to cite a stadium name but to show the skills of *President al-Assad* in the cited

subject. Such problems are rare but cause the noise.

	Precision	Recall	F-measure
Newspaper texts (Sport domain) 4000 texts (94,5 Mo)	98%	90%	94%

Table 2: Obtained results

As indicated above, we have also applied our resources to the education domain. We have collected a corpus composed of 300 texts containing university institution names. The performance measure of the obtained results gives 98% of precision, 70% of recall and 82% of F-measure. We deduce that the silence is increased. This is caused by the incompleteness of specific dictionaries to this domain and lack of some paths in the developed transducers. So our resources are applicable regardless of the domain, provided that we use the same features adopted in dictionaries we have built. It is evident that for reasons specific to the domain, we should sometimes add other paths and other sub-graphs, but we do not have to redo everything.

5.2 Experimentation of Translation Phase

The translation phase is applied to the extracted Arabic NEs during the recognition phase. Note that erroneous results are inherited. Therefore, heuristics filtering are necessary before the translation process. The obtained results of the translation phase are illustrated in Figure 7. As shown in this figure, the proper problems of this phase involve multiple translations that can be assigned to a word. For example, the selected lines in Figure 7 represent the NE' translation *ملعب مدينة الباسل الرياضية بدرعا malaab madinat el bacel el riadhiya bi deraa stadium of city Bacel sportive in Deraa*. In this NE, the word *مدينة madina* can be translated to the word "cit " *city* or "ville" *country*. NooJ displays all possibilities. In this case, the adjective can resolve this ambiguity. In fact, the adjective *الرياضية el riadhiya sportive* is generally related to the city and not to the country. Let's note that the word "باسل" *Bacel* remains in the source

language because it is a first name, so it will be transliterated later.

20095,20107,20107, 20095, ستاد بورسعيد, stade Port Said
 20353,20368,20368, 20353, ستاد جامعة بكين, stade Beijing
 ,23270,23290,23290, 23270, مسبح الأسد بدير الزور, piscine EL-Asaad Dayr az-Zur
 1,23501,23518,23518, 23501, مسبح الأسد الدولي, piscine الأسد international
 3,30369,30369, 30353, اسناد حلب الدولي, stade Alep international
 32781,32796,32796, 32781, ستاد حلب الدولي, stade Alep international
 2,32846,32862,32862, 32846, ستاد عمان الدولي, stade Amman international
 3,33098,33107,33107, 33098, ستاد عمان, stade Amman
 3,33501,33518,33518, 33501, ستاد مبارك الدولي, stade مبارك international
 4,34976,34976, 34944, ملعب مدينة الباسل الرياضية بدراعا, stade cité basel sportive Daraa
 7,34944,34976,34976, 34944, ملعب مدينة الباسل الرياضية بدراعا, stade ville basel sportive
 1,34980,34999,34999, 34980, الملعب البلدي بدراعا, stade municipal Daraa

Figure 7: Extract of results of word-to-word translation

To get a finer translation, we applied the readjustment rules that take into account the order of the constituents of NEs and the agreement of the adjective with the noun for which it is associated.

Our method provides 97% of well translated NEs while ensuring the specificities of the target language. The obtained result is promising and shows that there are some problems. These problems are related to the multiple translations assigned to a toponym (e.g., تونس *tounis* can be translated in *Tunis* or *Tunisia*).

After this evaluation, we remark that the proposed representation model facilitates the construction of the linguistic resources with the platform NooJ and the transformation from the semi-direct translation to transfer one. Indeed, we have separated the NE-recognition of their translation. In addition, it helps the promotion to the reuse of the needed local grammars. In fact, it is sufficient to change the inputs (i.e., dictionaries, morphological grammars) of the syntactic grammars for the desired results. Thus, for example, if we want to translate Arabic NE to another language other than French, the recognition module can be reused with some modifications if necessary (related to the specificities of the domain). Moreover, if we want to translate NEs from any language into French, also translation module can be reused. Indeed, this

module addresses the specificities of the French language.

6 Conclusion and Perspectives

In this paper, we have proposed an approach for recognition and translation of Arabic NEs (eventually NEs from other languages) based on a representation model, a set of bilingual dictionaries and a set of transducers resolving morphological and syntactical phenomena related to the Arabic NEs. Besides, we have described the representation model structure, its features and principles that should be satisfied. We have also given an experimentation and evaluation on the sports and education domains proving that our resources can be reused independently of the domain. The experimentation and the evaluation are done in the linguistic platform NooJ. The obtained results are satisfactory.

As perspectives, we try to improve our representation model by introducing other features related to the semantics. Furthermore, we are currently identifying heuristics filtering enabling finer NE translation.

References

- Barreiro A. 2008. Port4NooJ: an open source, ontology-driven Portuguese linguistic system with applications in machine translation. NooJ'08, Budapest.
- Borthwick A. Sterling J. Agichtein, E. and Grishman R. 1998. NYU: Description of the MENE Named Entity System as used in MUC-7. In Proc. of the Seventh Message Understanding Conference (MUC-7).
- Bourigault D. 2002. UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. TALN.
- Coates-Stephens S. 1993. The Analysis and Acquisition of Proper Names for the Understanding of Free Text. In Computers and the Humanities, Kluwer Academic Publishers, Vol. 26(5-6), Hingham, MA, p. 441-456.
- Ellouze S. Haddar K. and Abdelwahed A. 2009. Etude et analyse du pluriel brisé avec la plateforme NooJ. Tozeur, Tunisia.
- Fehri, H. Haddar K. and Ben Hamadou A. 2009. Translation and Transliteration of Arabic Named Entities. LTC Conference, Pologne, pp 275-279.
- Friburger N. 2002. Reconnaissance automatique des noms propres. PhD thesis, university of François Rabelais.

- Hasni E. Haddar K. and Abdelwahed A. 2009. Reconnaissance des expressions elliptiques arabes avec NOOJ. In proceedings of the 3rd International Conference on Arabic Language Processing (CITALA'09) sponsored by IEEE Morocco Section, 4-5 May 2009, Rabat, Morocco, pp 83-88.
- Khalfallah F. Haddar K. and Abdelwahed A. 2009. Construction d'un dictionnaire de noms composés en arabe. In proceedings of the 3rd International Conference on Arabic Language Processing (CITALA'09) sponsored by IEEE Morocco Section, 4-5 May 2009, Rabat, Morocco, pp111-116.
- Mesfar S. 2007. Named Entity Recognition for Arabic Using Syntactic grammars. NLDB 2007 Paris, 28-38.
- Mesfar S. 2008. Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard.: Thesis, November 2008, University of Franche-Comté.
- Mikheev A. Grover C. and Moens M. 1998. Description of the LTG system used for MUC -7. In Proc. of 7th Message Understanding Conference (MUC-7), http://www.itl.nist.gov/iad/related_projects/muc/ 894.02/
- Poibeau, T. 2005. Sur le statut référentiel des entités nommées. Laboratory of data processing of Paris North – CNRS and University Paris 13.
- Pollard C. and Sag I.A. 1994. Head-Driven Phrase Structure Grammar. Published by the press in the University of Chicago, Edition Golgoldmittu, Chicago, LSLI.
- Shalan K. and Raza H. 2009. NERA: Named Entity Recognition for Arabic. Published in Journal of the American Society for Information Science and Technology, Volume 60 Issue 8.
- Silberztein M. 2004. NooJ : an Object-Oriented Approach. In INTEX pour la Linguistique et le Traitement Automatique des Langues. C. Muller, J. Royauté M. Silberztein Eds, book of the MSH Ledoux. Presses University of Franche-Comte, pp. 359-369.
- Wu M. 2008. La traduction automatique français-chinois pour les groupes nominaux avec NooJ. Budapest.