

Recognition and translation Arabic-French of Named Entities: case of the Sport places

Abdelmajid Ben Hamadou

MIRACL, University of Sfax
Sfax, Tunisia.

abdelmajid.benhamadou@isimsf.rnu.tn

Odile Piton

University Paris1
Panthéon Sorbonne, France.

piton@univ-paris1.fr

Héla Fehri

LASELDI, University of Franche-Comte
Besançon, France.

MIRACL, University of Sfax
Sfax, Tunisia.

hela.fehri@fss.rnu.tn

Abstract

The recognition of Arabic Named Entities (NE) is a problem in different domains of Natural Language Processing (NLP) like automatic translation. Indeed, NE translation allows the access to multilingual information. This translation doesn't always lead to expected result especially when NE contains a person name. For this reason and in order to ameliorate translation, we can transliterate some part of NE. In this context, we propose a method that integrates translation and transliteration together using the linguistic NooJ platform that is based on local grammars and transducers.

In this paper, we focus on sport domain. We will firstly suggest a refinement of the typological model presented at the MUC-6 Conference. We will describe the integration of an Arabic transliteration module into translation system. Finally, we will detail our method and give the results of the evaluation.

Keywords: Named Entities (NE), Typological Model for NE, NE retrieval, Automatic Translation of NE, Morpho-Syntactic Analysis

1 Introduction

The recognition of named entities is an area of current research (Ehrmann 2008) given the proliferation of electronic documents exchanged through Internet and the need to treat them by means of NLP tools. In this context, several works for identification and marking have been carried out especially for Latin languages, and for English (Daille 2000). The language platforms

dedicated to specific domains such as the medical field are developed (Hamon 2007). On the other hand, the translation of named entities from one language to another (Piton 2003) (Grass 2000) (Maurel 2007) opens new perspectives because it may be the basis of new applications including in the domains of multilingual access to information, the annotation/indexing of documents and distance teaching of languages. Little work has been dedicated to NE of Arabic (Mesfar 2007), (Mesfar 2008), (Fehri et al. 2008).

This article focuses on the recognition and translation from Arabic into French of NE. We are particularly interested in the names of athletic venues: stadiums, arenas, pools, tracks.... We propose a typological model specific to sport domain, refining the model presented at conferences MUC-6 and an approach of recognition and translation. Based on rules, the implementation of this approach was performed using the NooJ platform. The rest of the article is organized as follows. We begin by detailing the typological model of NE of athletic venues names. Then, we make a parallel between the grammars of NE in the two studied languages. This parallel highlights the problems of Arabic-French translation for the studied domain. Finally, we detail the approach proposed for the simultaneous extraction and translation of NE.

2 A typological model of NE of sport venues

Typing NE is relatively recent. It has been clearly specified for the first time at the conference MUC-6. The three basic types that have been defined as a hierarchy (Poibeau 2005) are: ENAMEX which includes three subcategories of

proper names: Persons, Organizations and Locations, NUMEX which includes the numerical expressions -percentages, quantities and monetary values - and TIMEX representing dates and the durations.

Refinements have been proposed. The Prolex project (Tran 2006), whose purpose is to enable automatic processing of NE, has presented a modeling of the domain of the ENAMEX type. This modeling has defined the “conceptual proper name” and the “prolexeme” defined as “a set of variants (aliases), quasi-synonyms and morphosemantic derivatives”. An ontology has been specified taking into account the typology of Bauer (1998). This project has identified a two-level hierarchy (Grass 2002) made of types and super types. The super types are four: Anthroponyms (human feature i.e. names of per-

sons), Toponyms ((locative feature i.e. places names: city, country,..), Ergonyms (artifact i.e. objects and manufactured products) and Pragmonyms (event feature) and the types are thirty. The names of athletic venues that we are dealing with, correspond to the super type Toponym. The typological model that we propose is the result of a study of different forms of venues (stadiums, swimming pools, ice rink, ski slopes, ...) on corpus and lists of official names of sports venues available on the Internet for Arab countries (Tunisia, Algeria, Egypt, Saudi Arabia, Syria, ..) and Francophone countries (France, Belgium, Canada, ..). Our conclusion is that the typology tree proposed by the MUC conferences is unsuitable for our objective, and that is why we have to add a new type.

```

- <Exemples>
- <Exemple1>
  <LieuDeSport> استاد الملك فهد الدولي بالرياض = Stade Roi Fahd de Ryadh </LieuDeSport>
- <Catégories>
  <CategorieLieuDeSport> استاد = Stade </CategorieLieuDeSport>
  <Ethnonyme> الملك فهد = Roi Fahd </Ethnonyme>
  <Toponyme> الرياض = Ryadh </Toponyme>
</Catégories>
</Exemple1>
- <Exemple2>
  <LieuDeSport> ملعب مدينة تشرين الرياضية = stade de la cité sportive Tchrine </LieuDeSport>
- <Catégories>
  <CategorieLieuDeSport> ملعب = Stade </CategorieLieuDeSport>
  <LieuDeSport> مدينة تشرين الرياضية = cité sportive Tchrine </LieuDeSport>
  <CategorieLieuDeSport> مدينة = Cité </CategorieLieuDeSport>
  <Pragmonyme> تشرين = Tchrine </Pragmonyme>
  <Adjectif> الرياضية = sportive </Adjectif>
</Catégories>
</Exemple2>
</Exemples>
  
```

Figure 1 The typological model proposed retailer relationships that may exist between the base types.

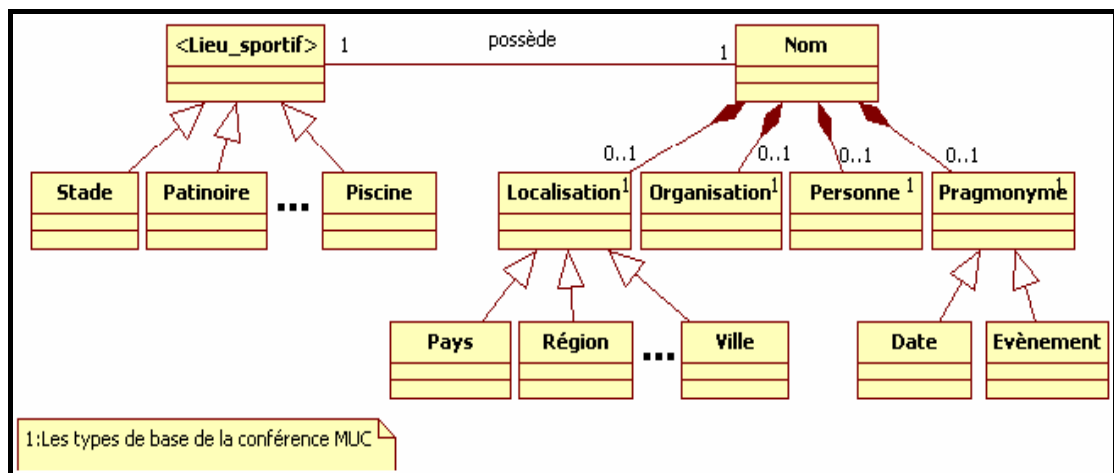


Figure 2: Typological Model of NE sports venues

We particularly stress the compositionality. Indeed, the name of an NE may include another with which it is connected. This inclusion raises the problem of the extraction the NE. Let's illustrate this by two examples:

3 Problems of recognition of Arabic NE

3.1 Agglutination problem

Arabic is an agglutinative language. Most words are formed by joining morphemes together. Indeed, textual forms are made up of the agglutination of prefixes (articles: definite article ال = the, prepositions: ل =for, conjunctions: و = and), and suffixes (linked pronouns) to the stems (inflected forms: لعبوا = لعبوا + وا = they play). In general, to obtain the different decompositions of a textual form, a morphological grammar is needed.

3.2 Determination problem

Some constituents of athletic venues names are always determined (in the case of adjectives). Others may be determined or not determined without rules governing these situations. This is true for toponyms:

City name directly following the category: ملعب صفاقس = Sfax stadium where the name is not determined and ملعب الرياض = Stadium Riyadh where the name is determined.

It is the same for countries and regions such as: ملعب المغرب = stadium of Morocco and ليبيا = Libya stadium.

This can also be met for anthroponyms: ملعب شتيفي غراف = Steffi Graf Stadium is undetermined while ملعب الشاذلي زويتن = Stadium = the Chadli Zouiten (Sport Tunisian) is determined

The treatment of this problem requires the inclusion of a feature in the dictionary that indicates that the noun is or is not determined.

3.3 The problem of proper names

Unlike the Latin languages, Arabic proper names do not begin by capital letters (upper case does not exist in Arabic). They are therefore difficult to identify. Thereof, their length is not known in advance, and may depend on the traditions of the region in which the person was born. Thus, in athletic venues names, they can be written as a single name (ملعب الأسد = El-ASAD stadium) or name and surname (ملعب الطيب المهيري = Taieb MHIRI stadium) or name and surname preceded by a title of nobility (ملعب الملك عبد الله = King Abdullah Stadium) In some regions it can be followed by "son of" (ولد) (ملعب سحيم بن حمد = Sahim Bin Hamad Stadium). Furthermore, it is not possible

to put in a dictionary all the NE with all their variants of writing. The transliteration could be a compromise because it can be applied generally.

3.4 The syntactic problems

in syntax, the grammatical building of NE in general, and those of sports venues in particular is rich and varied:

The length of NE (or the number of components) cannot be known in advance and is variable. To complete the sense and make it non-ambiguous, we tend to add an additional word (City, Olympic National,) or the name of the town, or the city name followed by the name of the country, or the kind of the game played in the sport venues: e.g. الملعب الأولمبي بالمنزه = Stade Elmenzah; ملعب مدينة الباسل الرياضية بدرعا = Stadium of the city El-Bacelli sport in Deraa.

The same type of component can be found in different positions: it is mainly the case of the adjective that does not always follow the name to which it reports: ستاد عمان الدولي = Amman International Stadium; استاد الوطني في بانكوك = National Stadium in Bangkok.

The position of toponyms is also variable. It can follow the category of sport venue or be at the end: ستاد حلب الدولي = Aleppo International Stadium; استاد الملك فهد الدولي بالرياض = King Fahd International Stadium in Riyadh.

1 Problems of Arabic-French translation of NE

The translation of NE from Arabic into French is not a trivial task. Several problems must be treated to produce a valid translation. We try to summarize here these issues:

- Triggers ambiguities of the source language. Example the word "stadium" which is a trigger name of athletic venues is also used to name sport club like بعامل يسوتلا = Tunisian Stadium.

- The gender (masculine or feminine) is not always the same for the Arabic word and for its translation in French. These features have an influence on the agreement within the NE. Example: the word مسبح is masculine, while its translation "pool" is feminine.

- The translation of proper names and city names from one language to another depends on the existence of an exonym (a local name for a foreign place, like Londres in French for London); otherwise, we should transliterate.

- Ambiguity between country names and capital names in Arabic: e.g. the toponym تونس "Tunes" can be translated as "Tunisia" or as "Tunis". It is also the case of الجزائر which can be translated as "Algiers" or as "Algeria".

- The place of adjectives in the NE is not the same for both languages: example ملعب الملك عبد العزيز الأولمبي = Malaab Al-Malik Abdelaziz al-oulimpi which is translated into French by: King Abdelaziz Olympique Stadium.

- The order of determiners and prepositions is not always the same. For the Arabic language, if the name of cities follows the category directly, there is no particle (ملعب موناكو), while in French we use 'of' better than 'in' (stadium of Monaco).

- The translation of dates is also problematic when it is expressed in the origin text according to Hijri calendar (Hijri). Changes become necessary.

4 An approach to recognition and translation

Our approach of recognition and translation is based on rules. It is based on a balance between grammar and lexical representation. Grammar is the set of rules of arrangement of lexical components to form an NE. The lexical components are limited to types of the typological model presented above (see samples in Table 1). They are represented as lemmas with the corresponding French translation and with a bending model. The list of lexical components is based on the typological model that we have proposed. It shows the components that make up the athletic venues names.

This list is as follows:

- Names of cities
- Names of countries
- Adjectives associated with athletic venues (Olympic, national, municipal, and sporting)
- The names of politicians and athletes
- The categories of sports venues (stadium, swimming pool, ice rink...)
- Common nouns associated with sports: fraternity, freedom, progress...

The bilingual dictionaries have the features that are relevant for each language. The Arabic dictionary has the features needed to identify the noun phrase, and of which we have detailed the specific grammar. They are the features, Fonction, Cat_Geo, Toponyme, Ville (City), Pays (Country), Perso (Person), and LieuSport (Sport venues). The 'Fonction' feature concerns the

functions occupied by the person: e.g. Amir, president or king. The 'Cat_geo' feature indicates a geographical category, e.g. republic, city, region. The 'Toponyme' feature indicates a place name. A second characteristic specifies whether it is a city or a country; The 'Perso' feature indicates that it is a name of a personality; that can be a leader or an athlete. The dictionary also indicates the inflections.

The features listed in the French dictionary are the features useful for the generation of a noun phrase. Besides the classical features such as gender and number, we register the feature DETZ which characterizes names without determiner. The apostrophe feature affects the entries beginning with a vowel or a mute h, and requiring the use of the apostrophe after 'de' and after a determiner: i.e. d' or l'. The French dictionary also indicates the inflections of French lemmas. The use of flexion and of the file Properties.def -that describes the name of the features- allow to place in the graph the conditions concerning the features of the word in French, as well as the features of the word in Arabic.

Arabic Dictionary
Functions names
ريدم, N+Fonction+FLX=A1+FR=directeur
مكلم, N+Fonction+FLX=مكلم+FR=roi
names of geographical categories
مجمهورية, N+FLX=عراق+Cat_Geo+Toponyme+FR=république
مكلم, N+FLX=عراق+Cat_Geo+Toponyme+FR=royaume
Geographical names
تونس, N+PR+s+Pays+Toponyme+FR=Tunisie
تونس, N+PR+s+Ville+Toponyme+FR=Tunis
رياضي, N+PR+s+Ville+Toponyme+FR=Riyadh
Personalities' names
إليزابيث, N+PR+Perso+f+s+FR=Elisabeth
حبيب بورقيبة, N+PR+Perso+m+s+FR="Habib Bourguiba"
Common nouns; name of sporting places; triggers
مركز, N+LieuSport+FLX=مركز+FR=centre
حمام, N+LieuSport+FLX=حمام+FR=piscine
ملاعب, N+LieuSport+FLX=ملاعب+FR=stade
Common nouns
أخوة, N+FLX=عراق+FR=amitié
شجعان, N+FLX=شجعان+FR=armée
adjectives
وطني, A+FLX=A1+FR=national
أولمبي, A+FLX=A1+FR=olympique
بلدي, A+FLX=A1+FR=municipal
دولي, A+FLX=A1+FR=international
demonym adjectives
تونس, A+Toponyme+FLX=A1+FR=tunisien

مصر, A+Toponyme+FLX=A1+FR=égyptien
French Dictionary
names
stade, N+FLX=Ballon
amitié, N+apostrophe+FLX=ABH-O201
roi, N+FLX=ABH-Oroi
Emir, N+apostrophe+FLX=ABH-Oemir
armée, N+apostrophe+FLX=Table
adjectives
municipal, A+FLX=76
olympique, A+FLX=31
national, A+FLX=76
names of geographical categories
république, N+FLX=Table
ville, N+FLX=Table
royaume, N+FLX=Table
Personalities' names
Habib Bourguiba, N+PR+m+s
Elisabeth II, N+PR+f+s
Fahd, N+PR+m+s
Geographical Proper names
Riyadh, N+PR+FLX=ABH-O201
Sfax, N+PR+FLX=ABH-O201
Tunis, N+PR+FLX=ABH-O201
Tunisie, N+PR+FLX=ABH-O201
Indonésie, N+apostrophe+PR+FLX=ABH-O201
Jakarta, N+PR+FLX=ABH-O201
Maroc, N+PR+FLX=ABH-O199
Bangkok, N+PR+FLX=ABH-O199
Jaka Baring, N+PR+FLX=ABH-O199

Table 1: Extracts of French and Arabic dictionaries

The local grammars for recognition and translation of sports venues use other more elemen-

tary grammar: a morphological grammar for decomposition of words glued in Arabic language; inflectional grammars for each language, a grammar for recognizing dates that represent historical events in Arabic.

The process of recognition and translation of sports venues requires 3 basic steps:

1. Recognition of Arabic NE
2. Translation (literally) of the components of the French NE with the associated characteristics
3. Reorganization of the components in accordance with the grammar of French. (Necessity to change the words order and change or addition of specific connectors). These steps can be combined in two different ways: by consolidating the two early stages and by linking the third or by combining the three steps (1 +2 +3) into one. The advantage of the first alternative is that it is more portable than the second. We chose the second alternative, particularly for reasons of simplicity of implementation because it requires the definition of the general format suitable for other languages and our control of NooJ platform of development does not allow us to retrieve information generated by the transducers of recognition and automatic sequencing of reorganization.

The recognition and translation that we propose is based system on Transducers implementing the grammars above indicated which combine triggers which may be internal evidence (i.e. part of the NE) or external (i.e. announcing a NE). The transducer for the recognition and translation is given in Figure 3.

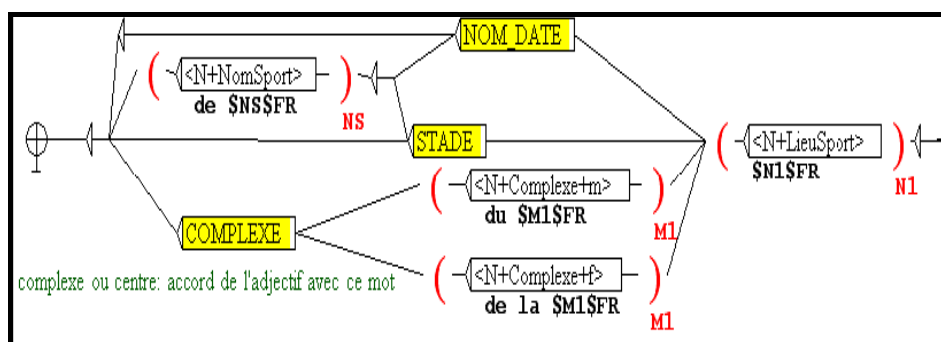


Figure 3: Main transducer for the recognition and translation of names of sports venues

As we stated above, adjectives pose a problem for both recognition and translation. To solve this problem, we devoted a special transducer. This transducer allows the translation of the adjective before processing components that precede it in the source NE that can be a toponym, an ethnonym, or a pragmonym. Thereof the result will

respect the position of adjectives in a French noun phrase (see Figure 4). The graph in Figure 5 shows an example of translation of place names and country names. This graph takes into account the presence or absence of determiner before the names of countries according to gender (masculine or feminine).

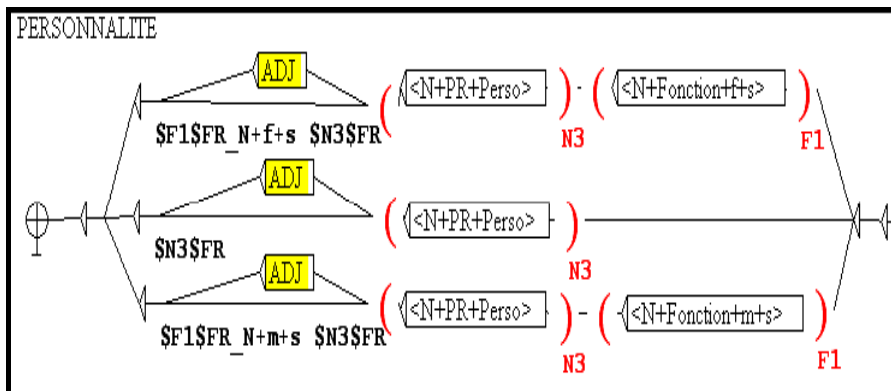


Figure 4: Transducer treating adjectives

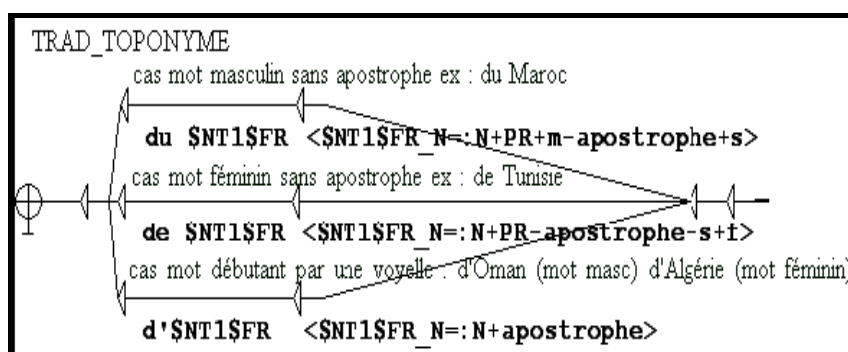


Figure 5: Transducer of translation of names of countries

بفح في مدينة هوشي	أسناد الجيش/stade de l'armée
بفح أسناد الملك فهد الدولي	أسناد الملك فهد الدولي بالرياض/stade international roi Fahd de Riyadh
في الجزء الشمالي الشرق...	أسناد الملك فهد الدولي/stade international roi Fahd
، حيث تم تصميمه على...	أسناد الملك فهد الدولي/stade international roi Fahd
الكبير الواقع بالعاصمة ال...	الأسناد الدولي/stade international
على 23 غرفة خاصة لمح...	أسناد الملك فهد الدولي/stade international roi Fahd
يحتل سفح أسناد الملك فهد	أسناد الملك فهد/stade roi Fahd
أعلى سفح ملعب في العالم	أسناد الملك فهد الدولي/stade international roi Fahd
على انه واحد من أهم	أسناد الملك فهد الدولي/stade international roi Fahd
بفح في وسط جاكرتا في	أسناد بونغ كارنو/"stade "Bung Karno
هو أسناد متعدد الإستخدام...	أسناد جاكا بارنج/"stade de "Jaka Baring
في نابلند هي ملعب متعدد	الأسناد الوطني في بانكوك/stade national de Bangkok
بحمام الألف ملعب ريا...	الملعب البلدي/stade municipal
محتويات 1	أسناد باريس/stade de Paris
بكتسي طابعا مهما في حال	المسبح البلدي/piscine municipale
: هو أول سناد في المملكة	سناد عمان الدولي/stade international de Amman
في العاصمة عمان عاصم...	سناد عمان/stade de Amman
ل 35 ألف متفرج ، ويعد من	سناد عمان/stade de Amman

Figure 6: A sample of the concordance obtained for the evaluation corpus

5 Experimentation and Evaluation

To test and evaluate the proposed approach, we collected a corpus made up of journalistic articles and lists of official naming of sports venues available on the Internet. The corpus that we used for this initial evaluation consists of a hundred texts. Figure 6 shows the results obtained in the form of concordances as provided by the platform NooJ.

The evaluation metrics we used are Recall, Precision and Fmeasure. Let's remember that the recall measures the quantity of relevant responses of a system compared to the ideal number of responses; Precision is the number of relevant answers of the system among all the answers he gave and the F-measure is a combination of a single value of measures of Precision and Recall for penalizing the very large inequalities between these two measures. The values obtained in the evaluation of our work are:

	Precision	Recall	F-measure
Sports venues	97%	95%	96%

In addition to issues already raised, we encountered a first problem mainly caused by the lack of standards for writing proper names, especially those transliterated from foreign proper names and caused by the use of signs delimiters like the dash and the parenthesis at the end of names of sports venues, to specify the geographical location. The second problem is due mainly to the absence of a specific tool available in NooJ, able to recognize the particles agglutinated with compound nouns (see the example in Figure 6: de الملعب البلدي بحمام الأنف = municipal stadium Hammam Lif whose recognition is incomplete).

6 Conclusion

In this article we have presented an approach for recognition and translation from Arabic to French names of sports venues. We have particularly noted the problems posed by the recognition. Some of them are specific to the Arabic language. These problems have been largely resolved, but some deserve special consideration • in particular the transliteration of the proper names and the abbreviations and acronyms. Methodologically, we want to test the separation of the recognition step of the translation. This is in order to reuse the recognition for translation to languages other than French.

References

- DAILLE B., MORIN E. (2000) Reconnaissance automatique des noms propres de la langue écrite: les récentes réalisations. *Traitement automatique des langues*, 3/2000, ATALA/Hermes Sciences, Paris Vol. 41, 601-621.
- DAILLE B., FOUROUR N. MORNE E. (2000) Catégorisation des noms PROPRES : Une étude en corpus. *Cahiers de grammaire*, Vol. 25, 115-129.
- EHRMANN M. (2008). Les Entites Nommées, de la Linguistique au Tal : Statut théorique et méthodes de désambiguïsation. Thèse 2008, Paris 7.
- FEHRI H., HADDAR K., SILBERZTEIN M. et BEN HAMADOU A. (2008), Reconnaissance automatique et analyse sémantique d'entités nommées en Arabe, Conférence NOOJ'08, Budapest.
- GRASS TH. (2000). Typologie et traductibilité des noms propres de l'allemand vers le français. *Traitement automatique des langues*, 3/2000, ATALA/Hermes Sciences, Paris Vol. 41, 643-669.
- GRASS TH., MAUREL D, PITON O. (2002). Description of a Multilingual Database of Proper Names. Actes de *PorTAL 2002*, LNAI 2389, 137-140.
- HAMON TH., NAZARENKO A., POIBEAU TH., AUBIN S., DERIVIÈRE J. (2007). A Robust Linguistic Platform for Efficient and Domain specific Web Content Analysis. *RIAO 2007 - Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*, Pittsbrigh : United States
- 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. Thèse, novembre 2008, Université de Franche-Comté
- MAUREL D., VITAS D., KRSTEV S., KOEVA S. (2007), Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian, Les langues slaves et le français : approches formelles dans les études contrastives *Bulag*, 32 , 55-72.
- PITON O., GRASS TH., MAUREL D. (2003) Linguistic Resource for NLP: Ask for "Die Drei Musketiere" and meet "Les Trois Mousquetaires". *8th International Conference on Applications of Natural Language to Information Systems*, 2003, Burg (Spree-wald), Natural Language Processing and Information Systems- NLDB'2003, *Lecture Notes in Informatics*, GI-Edition, Antje Düsterhöft, Bernhard Thalheim (Eds), p. 200-213 , ISBN 3-88579-358-X

POIBEAU TH. (2005). Sur le Statut référentiel des entités nommées, *TALN 2005, Vol. 1/2005* 173-182.

SILBERSTEIN M. (1993). Dictionnaires électroniques et analyse automatique de textes. Le système INTEX. Masson Ed. Paris Milan Barcelone Bonn.

SILBERSTEIN M. (2005) NooJ's dictionaries, *Proceedings of LTC*, Poznan University.

TRAN M., MAUREL D., (2006). Un dictionnaire relationnel multilingue de noms propres *TAL Vol. 47 – n°3/2006*. 115-139.