

Linguistic resource for NLP: Ask for “Die Drei Musketiere^{*}” and meet “Les Trois Mousquetaires^{*}”.

Odile Piton, Thierry Grass, Denis Maurel

Université Paris I
12 place du Panthéon
75005 Paris France
piton@univ-paris1.fr
grass@univ-tours.fr
maurel@univ-tours.fr

Abstract: Our work concerns proper names or "named entities" in NLP, in a multilingual context and in the spirit of the action "Technolangue" launched in France by the Ministry for Research, the Ministry for the Culture and the Communication and the Ministry for the Economy, Finances and Industry in 2002, with the purpose of production, validation and diffusion of linguistic resources. Our objective is to create a multilingual electronic dictionary intended to record such terms in several languages as well as the bonds between them, and

We know that toponyms follow strict rules imposed by international or national organizations (UNGEGN) intended to standardize them. This is not true for the other categories of proper names. In order to create an operational tool that is able to help translation, we must study the proper names, their grammar and the semantic relations between them. We propose ontology based on a taxonomy inspired by the works of Bauer and others. We took as a starting point the work of Mel'cuk on the lexical functions as well as work of G. Miller on the *WorldNet* system. The basic model includes semantic or lexical relations. These relations make it possible to locate a proper name in a lexical network and to provide responses on request.

Introduction

Production, diffusion, automatic treatment and exploitation of electronic information are very largely conditioned by the availability of adapted linguistic resources and powerful software components. The importance of linguistic resources is undeniable. However, the essential effort of scientific NLP community concerns until now dictionnairic resources of common nouns and specialized terminological resources. We propose, through this project, to introduce knowledge on the proper names.

^{*} The Three Musketeers

The aim is the constitution of a multilingual relational dictionary of proper names (it is a database) and the realization of tools of assistance to the extraction and the maintenance of these dynamic knowledge bases.

Tired of being mistaken for these people and anyone else who might share his name, Alan Berliner, the filmmaker from New York -- not to be confused with Belgian filmmaker Alain Berliner -- decides to rid himself of the dreaded Same Name Syndrome. His solution: invite all the Alan Berliners in the world over to his house for dinner¹.

„Ich bin ein (Alan) Berliner“ Filmmaker Alan Berliner has a new version of the name game, but there's a twist. You have to be named Alan Berliner to play. [...] Berliner is the 17,639th most common last name in the USA, he says. Smith is the most common, while Johnson is No. 1 for African-Americans. Rodriguez (and variations) is the most common Latino surname, and Cohen (and variations) is the most common Jewish surname. Chang is the most common worldwide. Bill Keveney, USA TODAY 06/26/2001

Example 1: "I am an Alan Berliner"

Example 1 tells a true story. It is a counterexample of some generally admitted properties of human proper names: it does not concern an unique person and it shows that if you need it, you can use the determiner “a” or “the” and the word Berliner can have a “s” when in the plural: an Alan Berliner = *ein Alan Berliner*, all the other Alan Berliners = *alle anderen Alan Berliners*.² This is an introduction to our work that concerns proper names, and an invitation to adopt a definition that includes pure proper names and descriptive proper names [Jo94]. We treat of proper names and their derivatives [Eg02], [EMB98], in a multilingual context: the question is “what is to be registered, what are the properties and information useful to translate proper names”?

Our work has been initiated at the University François Rabelais of Tours by Denis Maurel (the Prolex project [Ma96]). It concerns proper names and the constitution of such resources for NLP [PMB99], due to the fact that they are usually missing in paper dictionaries [Re77] or in electronic ones [Ma95]. Our approach is based on both the work of classifications suitable for onomastics [Za68], [Bau85], [Gra00] and for NLP [Pa96]. We have noted that proper names have specificities that make them difficult to be translated. Translation must not be seen as two corresponding lists of words or expressions. Proper names have many forms going from initial forms to full developed forms, including sometimes several abbreviations. Moreover this combines with specific rules like capitalization or determination that can be different in different languages. It is strategic to complete the study to be able to register relevant syntactic and semantic information. Our first part relates remarks on proper names in a multilingual context, our second part will describe our taxonomy and ontology. The third part presents the evolving from a monolingual DB to a multilingual DB. In this paper, we are going to limit ourselves to French and German.

¹ Text found in German: (Alan (NOBODY'S BUSINESS) Berliner hat sich Gedanken über (seinen) Namen gemacht - und alle anderen Alan Berliners zum Essen eingeladen).

² This is a German/English example, in French we would say “Un Alan Berliner”, and “Les Alan Berliner(s)”.

1 Proper Names in a Multilingual Context

We have to study what a proper name is in all the target languages. It is not a surprise that they do not fit together. We will not develop here what proper names are in French [GG86]. It can be referred in other studies in NLP [DM00], [Fr02], [Gra00], [MPE00]. We will only point out some characteristics that are important in a computational and multilingual context.

Names can they be parted between proper names and common nouns? Many common nouns are coming from proper names by an operation called *antonomasia*, or by *catachresis*. Some names are both common and proper names: for example the proper name “Harpagon” has provided the common noun “harpagon”. Of Greek origin, the word *harpagon* means raptor. There is a “return ticket” from one category to the other and back. There is no reason that lexicalization is exactly the same at the same time in different languages.

1.1 About Proper Names in French

The frontier between common nouns and proper names is not obvious. As an example: North takes a capital letter when it acts as the cardinal point, but not when it acts as a direction. Sun, Ground and Moon are proper names only when they act as stars, “the Earth turns around the Sun”, but they are common nouns in the expressions: “*tomber par terre*” - to fall to the ground and in “*nous avons eu deux jours de soleil*” - we've had two sunny days. French dictionaries have two parts (Larousse) or two volumes (Petit Robert). One of them is for usual words and the other for proper names. This last part is not a real dictionary: most time gender and number of words are not indicated. You have to read the text and if you find a non-ambiguous sentence, you can guess what they are. But more often than not, you will not be able to. When a proper name is lexicalised as a common noun, it enters the usual dictionary as a headword [Me99], for example *sandwich*, *ampère*, *frigidaire*³...

- *Variability of morphological forms.*

It is usual to find proper names according to morphological forms. We point out that it is not very reliable. For example, we looked for “Mont-Blanc” and “Tour Eiffel”. in the TLFi⁴. “mont-blanc” is registered as a headword, because it is the name of a cake, but it is not registered as the highest point of the French Alps⁵, nevertheless we have found it in 47 explanations or examples, with four different writings⁶.

³ Resp. name of Lord Sandwich; name of a scientist; name of a trade mark.

⁴ Trésor de la Langue française informatisé: Computerized dictionary: 100,000 headwords, 270,000 definitions and 300,000 examples. It does not register proper names as headwords, but they can be found in sentences given as examples. Most examples are chosen in French literature.

⁵ It is a trade mark too: « Lait Mont-Blanc » from 1923 in Rumilly (Savoie).

⁶ It cannot be excluded that some writings in TLFi are mistakes.

For the different writings, we write the number in Table 1 and we give the name of one French author who used it. We do the same for the two writings of “Tour Eiffel”.

Mont-Blanc	30	Hugo, <i>La Légende des siècles</i> , t. 5, 1877.	mont Blanc	6	In “Le Littré”
Mont Blanc	10	Sainte-Beuve, <i>Pensées et maximes</i> , 1869.	tour Eiffel	20	E. Triolet, <i>Le Premier accroc coûte deux cents francs</i> , 1945
Mont blanc	1	Chateaubriand., <i>Voyage au Mont blanc</i>	Tour Eiffel	10	Colette, <i>Fanal bleu</i> , 1949.

Table 1: “Mont-Blanc” and “Tour Eiffel” in TLFi

It is noticeable that great authors do not use the same forms. Let us catch a glimpse on the rules.

- Hyphen and Capitals in French

Proper names have to be listed in many documents. Their place depends on alphabetical order. And alphabetical order of composite terms depends on hyphen. So hyphen obeys precise rules with consequences on capitalization.

Organizations entitled to give directives raise many problems in connection with capital letters, the practices differ according to languages and French has the appearance of an exception. But this exception is not well known. We present some citations from the European Union, from Nato, and from the *Office québécois de la langue française*.

An example you can find: le nouveau Directeur Adjoint de l’Office est Monsieur X, ancien Directeur du Développement et des Relations Internationales du Groupe «École Supérieure de Commerce de Reims»⁷ .instead of: le nouveau directeur adjoint de l’Office est M. X, ancien directeur du développement et des relations internationales du groupe «École supérieure de commerce de Reims».

Example 2: European Union⁸

The "Guide of use of the French capital letter" and the "Guide to Capitalization in English" intend to specify the rules of use of the French and English capital letters. The two guides begin by: “This annex is not intended to replace grammar and specialized books, which should be referred to when in doubt. **Nevertheless such authoritative books do not always agree on the use of capitals in some specific cases.**” In the four pages-guide, it is noted that the adjective located after the name takes a capital letter if it is connected to him by a hyphen, the word « Nations Unies » that is written with two capital letters without any hyphen, constitutes an exception of use within this organization, as well as the word « Etat-major ». (Staff).

Example 3: Nato²

⁷ The new Deputy manager of the Office is Mr Guy Haug, old Directing Development and International relations of the Group "Reims management School".

⁸ <http://eur-op.eu.int/code/fr/fr-4100200.htm>

... « le fait d'écrire l'*Institut national de la recherche scientifique* avec une majuscule initiale sur le premier mot ou sur tous les mots (l'*Institut National de la Recherche Scientifique*) ne change pas le sens..... Ce phénomène n'est pas considéré comme une « maladie » en anglais, en espagnol, en italien ou en portugais, c'est «normal». Le français, pour sa part, fonctionne autrement.¹⁰ »
Charte des droits fondamentaux de l'Union européenne, Charter of Fundamental Rights of the European Union, Carta dei diritti fondamentali dell'Unione europea, Charta der Grundrechte der Europäischen Union, Carta de los Derechos Fundamentales de la Unión Europea, Charta um Chearta Bunúsacha den Aontas Eorpach, Carta dos Direitos Fundamentais da União Europeia Handvest van de grondrechten van de Europese Unie

Example 4: The *Office québécois de la langue française*¹¹

When you write in English, the use of capital letters is different than in French. We will not list the rules, we just notice that it is not said that the capitals are relevant to proper names, they can be a mark of deference, or the beginning of a sentence, or used to write something in capitals without any regard to proper names. It has been noticed that some parts of composite terms need always a capital [BZ00]. We need a “Capitalization feature” for these parts.

1.2 About Proper Names in German.

In the German grammar there is no dichotomy between proper and common nouns, but there is one between concrete and abstract words. The *Duden Universalwörterbuch* includes some proper names, but the *Wahrig* does not. The result is that lot of proper names can't be found in dictionaries or encyclopedias. Capitals are used for all names and part of adjectives that derive from proper names, and for some terms of composite proper names. The official rules are established by the “amtliche Neuregelung”. The specific points concerning German have to be studied and registered. We must add that bilingual dictionaries register common nouns as well as some proper names.

We register some information that is relevant: such as gender, number and determiner, but for some words the gender (or the number) can be unknown. We register the capitalization if some terms have a part that always needs a capital [BZ00]¹².

Let's observe different examples in German of a French proper name: “*Ein Stern im Michelin ist ein besseres Kapital*”, here “*der Michelin*” means the tourist Guide “*le Michelin*”, while in “*Michelin hat einen Reifen hergestellt*”, “*Michelin*” is without determiner and means the enterprise. In “*Ich habe Michelin-Reifen getestet*”, “*Michelin-Reifen*” means Michelin tires and can be translated by “*des michelins*” in French, or by “*des pneus Michelin(s)*” It is not homonymous in French, because the determination disambiguates between the three meanings. In French there is a fourth one: “*les Michelins*” that is used (by journalists) for “employees of Michelin-Enterprise”.

⁹ <http://www.nato.int/docu/stanag/aap015/aap15.htm>

¹⁰ “the fact of writing National Institute of Scientific Research with an initial capital letter on the first word or all the words does not change the meaning. ... This is not regarded as a “disease”, in Spanish, English, Italian or in Portuguese where it is “normal”, but not in French.

¹¹ <http://www.olf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/faq/249a.html>

¹² Invariants between lists coming from various origins and versions: English, French, German.

German language has three genders. Some names don't accept a determiner (it's true in French too, but not exactly for the same words): "*Deutschland*" in German, while "*l'Allemagne*" needs one determiner in French. Proper names are usually either in the singular or in the plural, but because of declension, we have to take care of the genitive. When the term is not translated, we need a "default case". We choose the nominative and the singular [Gra02]: "*Er wohnt in der Alten Bergstraße*" = "*Il habite dans la Alte Bergstrasse/ il habite la Alte Bergstrass/ il habite Alte Bergstrasse*". Personal names are sometimes used with particular civilities (Prof., Dr) in German. It is different in French. This leads to study local grammar of person names [FM01] and the use of civilities, the German composites etc... Many categories of proper names have specific grammar.

1.3 Translation

- *Connotation?* There is a discussion about denotation and connotation of proper names. Connotation is in relation to beliefs and ideas and to the symbolic meanings that an expression may carry. Mill says that genuine proper names have denotation and not connotation, and that "general names", or general terms, have connotation. With respect to what we have said before, we think that there is not a partition of words. Proper names have more or less connotation, our dictionary will have to register connotation for some of them. When people think of proper name, usually they think of name of people and of toponyms. But there are other kinds of proper names and there are sometimes many ways to express a proper name, and some ways have connotation: *Paris / Paname*. We want to register the connotation to be able to give an equivalent in another language. It is an important aspect of translation.

- *Translation or adoption?* We have observed an important point: as regards proper names, translating has evolved. We noticed that in many domains, and it is particularly true in toponymy. Many organisms (national and international organisms) work for the normalization of toponymic datas¹³. UNGEGN¹⁴ internationally promotes the use of the same names, so nowadays, new toponyms are less and less translated, but they are "adopted" (transliterated if necessary). In the old days, it was just the contrary, the names were translated, new words were created: *Munich* for *München*, *Londres* for London. This evolution is the same for many proper names: name of politicians, of artists, film titles: *Apocalypse Now* for example. They look like more and more in French and German.

¹³ "As fundamental to the need for global standardization of geographical names, UNGEGN promotes the recording of locally-used names reflecting the languages and traditions of a country. UNGEGN's goal is for every country to decide on its own nationally standardized names through the creation of national names authorities or recognized administrative processes. With the wide dissemination of the nationally standardized forms through gazetteers, atlases, web-based data bases, toponymic guidelines, etc., UNGEGN can promote the use of these names internationally."

¹⁴ United Nations Group of Experts on Geographical Names (UNGEGN)

- *Partial or total translation?* When a word is a composite name, consisting of a generic element and a specific element, or of a composite specific element, the translation of the “generic+specific” can be partial or total. For example *îles* (islands) *Falkland* = *Falklandinseln*, here only the generic island is translated. In other cases the specific word is translated too: *lac de Constance* = *Bodensee*. The generic can be a composite term and the specific too. As regards the words that derive from proper names, they are a big part of our database. To be true we need them because part of composite names use a word while other forms use a derivative (*France, République française*).

- *Initial form or developed form?* Some points that we observed so far the names of companies or associations often have several quasi-equivalent forms. Initial forms can need a development. The translation can require the use of the long form or an explanatory paraphrase preceding the initials, at least at the time of the first appearance of the proper name, for example: *SNCF* = *die französischen Staatsbahnen* “*SNCF*”.

- *Truncated forms.* Let us also say a word of ambiguity due to the use of reduced forms [BM96]. For example, in French, the majority of the composite names of cities are truncated when the context (extra linguistics: geographical proximity, notoriety...) allows it. In France, for example, 17 cities are named Neuville. For another 68 of them the composite name starts with Neuville! In the same way, in German, *Frankfurt* can be the short form of two different towns, *Frankfurt am Main* and *Frankfurt an der Oder*.

- *Substitution?* Some aspects cannot be translated without a sentence. In French a country can be indicated by its capital town, a city can be indicated by one of its attributes, Prime Minister can be replaced by Matignon. Such possibilities are not the same in German. A translation into German requires a substitution and the relation of synonymy that exists disappears with the translation.

These various forms are registered in our database, or can be computed from it.

2 Taxonomy and Ontology

2.1 Conceptual Classes

From our studies we have conclude that there exists sets of words that have similar properties, and we want to identify these sets. We studied various classifications concerning the proper names: [Bau85], MUC [Ch97], [Pa96], [Za68] and [Ba01]. Taxonomy that we created results from the absence of a satisfactory typology, adapted to the linguistic treatment. Our classification of the proper names is coherent from the linguistic point of view. We choose to define a two levels taxonomy [GMP02]:

a) Hypertypes, in fact the traditional syntactico-semantic features are: **anthroponymes** (human feature), **toponyms** (locative feature), **ergonyms** (inanimate feature) and **pragmonyms** (event feature).

b) Types, more precise subdivisions, that can be part of one, two or three hypertypes.

Anthroponyms are: patronymic, first name, dynasties or people's names, divinity, mythical or fictive personal names, firms names, associations or political parties, artistic ensembles or sporting clubs names, public or private institution names, content items such as universities, hospitals, institutions, names of international and non governmental organizations, names of inhabitants of a country, a town or a region.

Toponyms are: country & region name, islands, town or village name, name of a group of countries, quarter, road or street name, building name, parks and gardens, monuments, bridges, theatres, hydronyms (water areas, river or stream names), geonyms (natural geographical sites, mountains, glaciers, caves, plains or forests), celestial objects names (include planets and asteroids, galaxies), fictive or mythical places names, public or private institution names which are also collective humans.

Ergonyms are: something concrete or abstract produced by a human being: brand name or trade mark, Firms names which are also considered as collective humans, work names like books, films, fictive or mythical object name, vessels names.

Pragmonyms are: meteorological event name, historical or political event name, sporting or cultural event name feast name.

2.2 Ontology

Our database of proper name is a set of data with an organisation. Ontology is "explicit formal specifications of the terms in the domain and relations among them" [Gru93]. It consists of relevant information, about semantic relationship. It will be used later to compute requests. This ontology is a set of properties that makes semantics (a part of semantics) able to be computed (used with a computer). These properties are for example: "is a subclass of", "is a", "is a part of", "has parts", "is close to", etc. Some of them are transitive. To define the semantic relations in the multilingual database, we took as a starting point the work of Mel'cuk [Mel 84,88,92] on the lexical functions as well as the work of G. Miller [Mi90] on the *WordNet* system and its ramifications (*EuroWordNet*). We have kept its terminology and add some relations that are specific to translation. It is important that our ontology should be compatible.

Some *semantic relations* exist in the two languages: synonymy (and quasi-synonymy), polysemy, hierarchical links, Cap. Example of Synonymy: (*République fédérale d'Allemagne, Allemagne; Confédération suisse, Suisse*) [PM97], (*Bundesrepublik Deutschland, Deutschland; Schweizerische Eidgenossenschaft, Schweiz*). For quasi-synonyms we register features (familial, very short name, pseudonym, emphatic name, poetic name, pejorative etc...). The polysemous features concern different aspects of one single term. For example *Tschernobyl = Tchernobyl* can be think as a disaster that is to say a catastrophic event, or as a town. We duplicate.

Hierarchy is registered by meronyms and holonyms. A meronym is the name of a constituent part of something. A holonym is the name of the whole of which the meronym names a part. For example: *Tours* \subset *Indre-et-Loire* \subset *Région Centre* \subset *France*. An hyponym is the specific term used to point out a member of a class. This is a link between a proper name and its type and between types and hypertypes.

The *lexical fonction* Cap, meaning ‘chief’, is studied by Mel’cuk. In our case, it can concern a special link between a region and a capital town, between enterprises and subsidiaries. In French this link gives a quasi-synonym that can be used in international politics.

Specificity to one language: Two semantic relations, homonymy and expanding are usually specific to one language. There are a lot of homonymies between toponyms and other words [PM01]. In French the word “*France*” is either the name of a country or a first name. As a country name, it is translated into “*Frankreich*” and as a first name it is not modified. Victoria Falls (*die Victoria Falle, les chutes Victoria*) is the name of well-known falls, but it is also the name of a town besides in Zimbabwe, and this name is not translated, neither in German nor in French. In the example below (extracted from *WordNet* that has registered some proper names) the homonymy lasts, because most of these names are alike in French and in German.

1. Victoria, Queen Victoria -- (Queen of Great Britain and Ireland and Empress of India)
2. Victoria -- (goddess of victory; counterpart of Greek Nike)
3. Victoria, Victoria Falls -- (a waterfall in the Zambezi River on the Zambian border)
4. Victoria -- (a town in southeast Texas southeast of San Antonio)
5. Victoria, capital of Seychelles -- (port city and the capital of Seychelles)
6. Victoria -- (a state in southeastern Australia)
7. Victoria -- (capital of the Canadian province of British Columbia on Vancouver Island)

Example 5: The noun “**Victoria**” in *WordNet*

As regards expanding, it concerns the different composite terms that can be built on the specific: *PSA / PSA-Gruppe = groupe PSA; Indre-et-Loire / departement d'Indre-et-Loire*. The expanding should not be confused with a long form of the proper name (which is not translated), as *Deutsche Lufthansa AG / Deutsche Lufthansa* or more simply *Lufthansa*, these three expressions are all synonymous. The expanding plays a fundamental part in search and extraction [MaD96], [FM01].

3 Lexical resource and tools for NLP

3.1 Database of French proper names and transducers

From 1996 we had made a French dictionary that includes more than 323,000 words or expressions and 55,000 relations [MP99]. We have created a cascade of transducers for automatic recognition and typing on French language firstly [GM00], [FM01], [Fri02].

Transducers are applied with the development environment *Intex* on pre-processed texts. Pre-processing (division of the text in sentences, tagging with dictionaries) uses electronic dictionaries of single or compound words. For each word they register syntactic information like noun, verb, gender, number or person..., and semantic features like human, animate, concrete, toponym. They can be freely chosen. To improve the dictionary coverage, it was necessary to make a dictionary of proper names. Owing to our database *Prolex*, we have performed an electronic dictionary of toponymic proper names and inhabitant names *Prolintex*.

Table 2 gives the results of a study on one year of the newspaper *Le Monde* with a two-level finite state transducer cascade. (The first stage searches for names thanks to dictionaries of first names, place names and occupation nouns, the second level locates left and right contexts which indicate presence of proper names). Results can be used to improve the dictionaries coverage.

	Persons names	Occupation names	Place names
First level cascade			
Recall	84,7	78	95,2
Precision	97,5	94,7	95,0
Second level cascade			
Recall	9,0	10,1	1,3
Precision	84,5	77,3	66,7
Two level cascade results			
Recall	93,7	88,0	96,4
Precision	96,1	92,3	94,5

Table 2: Results for two level FST cascades

The transducers are applied with respect to priority order. All this is resulting of study which takes account of knowledge on proper names, as exposed before, and of properties of the corpus (for example: journalistic or literary text, transcription of oral communication...). So tools must be adaptable. Transducers have such good properties because they can be combined from others transducers.

3.2 From monolingual to multilingual resource for NLP

As a conclusion, we decided that the next step should be a general coverage of all categories of proper names, and that we should perform the study in a multilingual prospect.

Thierry Grass made a study on a German French thesaurus. “We can compare the pages of a French encyclopedia with those of a German one. *Brockhaus Enzyklopädie*, *Diercke Weltatlas*, *Duden Lexikon*, *DTV Atlas zur Weltgeschichte*, *Encarta Enzyklopädie*, *Grosser Atlas der Welt*, *Lexirom*, *Meyers Enzyklopädie*, for German and *Atlas classique Larousse*, *Encyclopédie Encarta*, *Encyclopédie géographique Quillet*, *Dictionnaire encyclopédique Larousse*, *Grand atlas historique Larousse*, *Petit atlas Bordas*, *Petit Robert 2* for French” [Gra02]. This has given us about 13000 most usual proper names. All these proper names have a link with a type and at least one hypertype.

In Table 3 we show some examples of data with translation into French. It has been made from German to French, but it can be used from French to German.

German	French	Hypertype	Type	Short generic translation
Falklandinseln	îles Falkland, Malouines	toponym	geonym	groupe d'îles britanniques au sud-est de l'Amérique du Sud
Falkland-Inseln	îles Falkland, Malouines	toponym	geonym	groupe d'îles britanniques au sud-est de l'Amérique du Sud
Europäische Bank für Wiederaufbau und Entwicklung	Banque européenne pour la reconstruction et le développement (BERD)	ergonym	entreprise	banque créée en 1990
Eumaios	Eumée	anthroponym	divinity	porcher d'Ulysse dans l'Odyssée
Erklärung der Menschen- und Bürgerrechte	Déclaration des droits de l'homme et du citoyen	pragmonym	historical and political event	

Table 3: Examples of data with translation into French.

We present in Figure 1 an interface for consulting.

Masque :	Consultation :	Traduction :
Détermination : indifférent	France [Ajouter] [Corriger]	Frankreich
Genre : indifférent	Détermination : Genre Nombre : Type : OUI FS Pays	
Type : Pays	SYNONYMIE Hexagone [Modif]	
Commence par : f	CHEF Paris [Modif]	Paris
Contient :	expansion : capitale DE LE	
Termine par :	EXPANSION [Modif]	
Non traduit : indifférent	DERIVATION Français [Modif]	

Figure 1: Interface for looking words up in DB.

The new database registers information with respect to the language, and transducers are adapted to different languages. The focus is not only to register some pairs of languages as exposed here, but to obtain multi-language resource as a natural result. Besides we intend to merge information on proper names in the general set of tools used for NLP. This requires the collaboration of specialized organisms as listed below. We have too the possibility to construct special tools on requesting, for example specialized monolingual proper name dictionaries adapted to special applications.

We have argued about a French-German dictionary, but other languages are expected: English, Polish, Arabic... We give a sight of the organisms that are involved in the project directed from Tours by Denis Maurel: French universities, *CNRS – FRE 2546* « Analyses de corpus linguistiques, usages et traitement » has a French Arabic thesaurus of about 3000 words and links, that it will develop to 30 000; *Systran S.A.* (information and translation technology) intend to operate translation, transliteration, merging and integrating in translating tools and evaluation; *Synapse Développement* has the purpose to incorporate the multilingual dictionary with check spelling, content analyzer and search engine, and *Exalead* will act for indexing of web sites.

Conclusion

If it is difficult to agree on what is a proper name, one can moreover wonder whether this property is preserved by translation. It is obvious that the “set of proper names” that we consider don’t fit neither the German nor the French definition of proper name. It is an entity able to include the proper names of different languages and their derivatives, as well as relevant terms expressions, according to the principal relations concerning an expression. We think that the double aim: proper names plus translation will determine a domain larger than proper names strictly speaking.

When the proper name comes from a third language, will the term obtained be the same one if one carries out a direct translation from the third language or if one translates it from the language of the text in the target language? This will be to study.

As specific tools, dictionaries of proper names could be integrated into a kit of minimal linguistic resources (such that elaborate within the framework of the Outilex project), for a given language and, probably, for a given period. Indeed, one can imagine that there is a category of proper names whose notoriety is relating to a national cultural inheritance (Molière, Paris, Napoleon Bonaparte...), whereas others seem to have a notoriety relating to current developments: the main part of proper names found in the study of a year of a journalistic corpus can appear useless a few years later. This specific “BLARK¹⁵” of proper names thus would consist of a nucleus plus modules located in time. Owing to our classification, BLARKs can be limited to some categories or be more general.

¹⁵ Basic LAnguage Resources Kit

Bibliography

- [Ba01] Ballard, M.: *Le nom propre en traduction*, Paris, Ophrys, 2001.
- [Bau85] Bauer, G.: *Namenkunde des Deutschen*, Bern, Germanistische Lehrbuchsammlung Band 21, 1985.
- [BM96] Belleil, C.; Maurel, D.: *Traitement informatique des ambiguïtés dans la reconnaissance des noms propres liés à la géographie*, Bulag, n°21, 1996, 29-50.
- [BZ00] Bodenreider, O.; Zweigenbaum, P.: *Stratégie d'identification des noms propres à partir de nomenclatures médicales parallèles*, Paris, Hermès, 2000, TAL, 41-3, 727-757.
- [Ch97] Chinchor, N.: *Muc-7 Named Entity Task Definition*, 1997, Website: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html#appendices.
- [DM00] Daille, B.; Morin, E.: *Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations*, Paris, Hermès, 2000, TAL, 41-3, 601-622.
- [Eg02] Eggert, E.: *La dérivation toponymes-gentilés en français : mise en évidence des régularités utilisables dans le cadre d'un traitement automatique*, thèse de doctorat en linguistique, cotutelle des universités de Tours et Münster, décembre 2002.
- [EMB98] Eggert, E.; Maurel, D.; Belleil, C.: *Allomorphies et suppléments dans la formation des gentilés. Application au traitement informatique*, 1998, *Cahiers de Lexicologie*, n°73, 167-179.
- [Fr02] Friburger, N.: *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*, thèse de doctorat en informatique, Université de Tours, 2002.
- [FM01] Friburger, N.; Maurel, D.: *Finite state transducer Cascade to extract Proper Nouns in French text*, 2nd Conference on Implementing and Application of Automata, Pretoria (South Africa), Edition : LNCS, 23-25 juillet 2001.
- [GM00] Garric, N.; Maurel, D.: *Désambiguïssation des noms propres déterminés par l'utilisation des grammaires locales*, *Revue française de Linguistique appliquée*, 2000, volume 5-2, 85-100.
- [Gra00] Grass, T.: *Typologie et traductibilité des noms propres de l'allemand vers le français*, Paris, Hermès, 2000, TAL, 41-3, 643-669.
- [Gra02] Grass, T.: *Quoi ! Vous voulez traduire "Goethe" ? - Essai sur la traduction des noms propres allemand - français*, Bern, 2002, Peter Lang, collection Travaux Interdisciplinaires et plurilingues en LEA.
- [GMP02] Grass, T.; Maurel, D.; Piton, O.: *Description of a multilingual database of proper names*, PorTal 2002, Faro, Portugal, 23-26 juillet, à paraître in *Lecture Notes in Computer Science*.
- [GG86] Grevisse, M.; Goosse, A.: *Le Bon Usage*, Duculot, 1986, Gembloux, Belgique.
- [Gru93] Gruber, T. R.: *() A translation approach to portable ontologies*. *Knowledge Acquisition*, 1993, 5(2):199-220.
- [Jo94] Jonasson, K.: *Le nom propre. Constructions et interprétations*, Duculot, 1994, Paris.
- [MaD96] MacDonald, D.: *Internal and external evidence in the identification and semantic categorisation of Proper Names*, *Corpus Processing for Lexical Acquisition*, 1996, 21-39, Massachusetts Institute of Technology.
- [MLC95] Maurel, D.; Leduc, B.; Courtois, B.: *Vers la constitution d'un dictionnaire électronique des noms propres*, *Linguisticae Investigationes*, 1995, volume 19:2, 355-368.
- [Ma96] Maurel, D.; Belleil, C.; Eggert, E.; Piton O.: *Le projet PROLEX, séminaire Représentations et Outils pour les Bases Lexicales, Morphologie Robuste de l'action Lexique du GDR-PRC CHM*, 164-175, Grenoble, 13-14 novembre 1996.
- [MP99] Maurel, D.; Piton, O.: *Un dictionnaire de noms propres pour Intex : les noms propres géographiques*, *Linguisticae Investigationes*, 1999, volume 22, 279-289.
- [MPE00] Maurel, D.; Piton, O.; Eggert, E.: *Les relations entre noms propres : lieux et habitants dans le projet Prolex*, *Traitement automatique des langues*, 2000, Vol. 41-1, 623-641.

- [Mel84,88,92]Mel'cuk, I.: (1984-I, 1988-II, 1992-III), Dictionnaire explicatif et combinatoire du français contemporain, Les presses de l'Université de Montréal.
- [Me99] Menoni, G.: Ces noms propres qui sont devenus communs. Les revues pédagogiques de la Mission Laïque Française Connaissance du français, . janvier 1999, n° 35.
- [Mi90] Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K.: Introduction to *WordNet* : an on-line lexical database, International Journal of Lexicography, 1990, n°3, 235-244.
- [Pa96] Paik, W.; Liddy, E. D.; Yu, E.; McKenna, M.: Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval, Corpus Processing for Lexical Acquisition, 1996 61-73, Massachusetts Institute of Technology.
- [PM97] Piton, O.; Maurel, D.: Le traitement informatique de la géographie politique internationale, Colloque Franche-Comté Traitement automatique des langues (FRACTAL 97), Besançon, 10-12 décembre 1997, in Bulag, numéro spécial, 321-328.
- [PMB99]Piton, O.; Maurel, D.; Belleil, C.: "The Prolex Database : Toponyms and Gentiles for NLP". Applications of Natural Language to Information Systems. Proceedings 4th,International Conference NLDB 99 ,Juin 1999, Klagenfurt, Autriche , G. Friedl, H. C. Mayr (eds.) Österreichische Computer Gesellschaft, Band 129, 233-237.
- [PM01] Piton, O.; Maurel, D.: Les Noms Propres Géographiques et le Dictionnaire Prolintex 4ème Workshop of INTEX Users, Bordeaux 11-12 Juin 2001, à paraître dans PUFC (Presses Universitaires de Franche-Comté).
- [Re77] Rey, A.: Le lexique : images et modèles. Du dictionnaire à la lexicologie, Paris, 1977, Armand Colin.
- [Za68] Zabeeh, F.: What's in a Name, An Inquiry into the Semantics and Pragmatics of Proper Names, La Haye, Martinus Nijhoff, 1968.