[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

A Multilingual Onomasticon As a Multipurpose NLP Resource

Svetlana Sheremetyeva, Jim Cowie, Sergei Nirenburg and Remi Zajac
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 880003
{lana,jcowie,sergei,zajac}@crl.nmsu.edu

Abstract

The paper specifies the generic structure of a large-scale multilingual onomasticon (proper names knowledge base) for English, Spanish, French, German, Chinese, Japanese, Korean, Arabic, Persian, Serbo-Croatian, Russian and Turkish, A taxonomy of proper names is developed with all their variants and senses for each of the languages involved. This work will create a language resource for a variety of computational linguistics applications.

1. Introduction

This paper describes some preliminary results of ongoing development of a large-scale multilingual onomasticon for English, Spanish, French, German, Chinese, Japanese, Korean, Arabic, Persian, Serbo-Croatian, Russian and Turkish. Processing proper names is a crucial feature of any NLP system which deals with real text, as in many types of documents proper names are very frequent. For example, in the TIPSTER and MUC corpora almost 50% of the vocabulary consists of proper names (Cowie 1995). The onomasticon is intended for use for a variety of computational linguistics applications: machine translation, information retrieval, extraction and summarization.

The Computing Research Laboratory is developing a complex multilingual onomasticon, a lexical DBMS for management and a tool to support the proper name acquisition. To model entries and links typed feature structures are used. The multilingual knowledge base of proper names is a part of a general multilingual lexical knowledge base being developed at CRL and has the same generic structure though some standard zones are never used. The specifics of proper names also impose some further constraints on the values of some zones (see Zajac, 1998). In this paper we focus on the issues specific to the content of onomastica.

The multilingual onomasticon under development at CRL is a cross-referenced set of monolingual onomastica and is thus composed of

1. A set of monolingual onomastica.
2. Translation relations linking these monolingual onomastica.
3. A schema which defines the structure (allowed attributes and values of these attributes) of onomasticon entries.

The monolingual onomastica, all using the same entry structure, though possibly different types of entry fillers (due to language and culture differences), have the following components:

- A monolingual onomasticon schema which defines the structure of its entries (and relations between entries).
- Monolingual onomasticon data, a set of actual onomasticon entries which are instances of the monolingual onomasticon schema.

The work to create a proper name multilingual database includes the following steps:

- developing superentry and entry structures, the number and content of dictionary zones;
- systematization of the semantic category inventory for all the languages included in the onomasticon;
- checking the adequacy of semantic categories by examining the occurrences of proper names in a large corpus;
- extraction of information from the tagged text to fill the fields of the onomasticon entries (proper names variants, morphological features, semantic categories, etc.).

2. The Onomasticon Entry

Each entry in a monolingual onomasticon corresponds to a single sense of a proper name. These entries are grouped in "superentries" based on their orthographic form, regardless of their syntactic category, pronunciation, or meaning. For example, if a superentry describes the name "New Yorker" in the English onomasticon the entries will be *New- Yorker-1* for the New York resident, and *New-Yorker-2* for the magazine. Each superentry is a tree of entries. The structure and the names of elements roughly follow the TEI specification (http://www.uic.edu/orgs/tei/) with some

crucial modifications in the specifications of semantics, translations and cross-references. In the formalism used, the SuperEntry type formally defines the structure of a superentry and the Entry type defines the structure of entries. The SuperEntry and the Entry have a set of features (zones/fields) in common that are defined by the type EntryElements (entry zones). For search purposes, the citation form is used as the primary key. The sense is used, as a secondary key and to connect aliases and translations.

The onomasticon organization can be briefly presented as follows:

```
onomasticon = {superentry}+

superentry ::= citation-form
    form-type := lemma | inflected |
phrase | abbriviation| acronym
    part-of-speech ::= noun |
adjective
    {entry}+

entry ::= sense
        orthography-variants
        grammar-features
        semantics
        usage
        aliases
        translation
        example
        annotation.
```

The FORM-TYPE zone specifies the type of the entry: the TYPE values can be "inflected" (e.g for nouns which exist in plural only), "lemma" (the traditional way of spelling a headword in a dictionary, "phrase" (in case a proper name consists of several words), "abbreviation" (for any abbreviated proper names) and "acronym" (if a proper name is an acronym). The PART-OF-SPEECH zone values in the onomastica can only be "noun" or "adjective". In the entry, the SENSE zone contains the citation form of a name with a sense index concatenated to it. e.g. *New-Yorker-1,* that is, the first sense of the word *New-Yorker.* The ORTHOGRAPHY-VARIANTS zone lists orthographic variants of the headword, e.g., the British vs. the American spelling as in *Centre*, and *Center*, respectively. The GRAMMAR-FEATURES zone contains information about the morphological and syntactic category constraints. For example, for French, this zone will contain a gender feature with the values "masculine" and "feminine." For German, the value "neuter" will be added.

The SEMANTICS zone attaches a semantic category to a name or its components. We fully realize that it is impossible to come up with a universally acceptable set of features. In the case of proper names, there is, however, hope for a relatively simple projection between different sets of features. For this project, we have preliminarily defined 45 semantic categories, organized in an hierarchy (Figure 1). These categories are defined as concepts in the Ontology (Mahesh and Nirenburg. 1995) which are sub-concepts of Name. One of these categories, "Name-link," is special. It refers to a connecting word or phrase inside a name but not properly a part of it (e.g., *van der, of, de la, de).* According to the TEI documentation, "It is often a matter of arbitrary choice whether or not such components are regarded as part of the surname."

The five top-level semantic categories are: **Animate, Place, Organization, Occasion** (event, e.g., *Christmas* or a period, e.g., *The Middle Ages)* and **Artifact** (e.g., *Walkman).* Below we define some other categories which might warrant an explanation:

- **Residence:** names of persons living in a particular city, village, country, continent, etc. *(Londoner, American, African).*
- **Generation:** name components used to indicate generational information *(Junior);*
- Role: name components which indicate that the referent has a particular role or position in society, such as an official title or rank. Sub-types for **Role** are: **Nobility** *(Lord, Baron),* **Honorific** *(Dr., Mme.),* **Office** *(President, Governor),* **Military** *(Colonel),* **Epithet,** a traditional descriptive phrase or nickname *(The Hammer, The Great),* **Religious** *(Pope);*
- **Geographical:** names of geographical entities: **River, Valley, Mountain, Lake, Sea, Ocean, Astronomical.** May contain a common noun identifying some geographical feature contained within a geographic name, such as *valley, mount,* etc. (the *Mississippi River);*
- **Geopolitical:** name of a geopolitical entity: **Bloc** *(Southeast Asia),* **Country, Region** *(Yorkshire),* **Settlement** *(New York)* or **SettlePart** *(Champ-Ely sees);*
- **Government:** names of governmental organizations *(The Senate);*
- **Political:** names of political organizations *(The Republican Party);*
- **Industry:** names of industrial organizations, corporations *(IBM);*
- **Service**: names of accountancy partnerships, restaurants, etc.*(McDonald's);*
- **Commercial:** names of supermarkets, department stores, pharmacies, etc.*(Dillard's);*
- **Educational:** name of school, college, university *(NMSU);*
- **Media:** TV or Radio network or program; newspaper, magazine (the *Washington Post).*

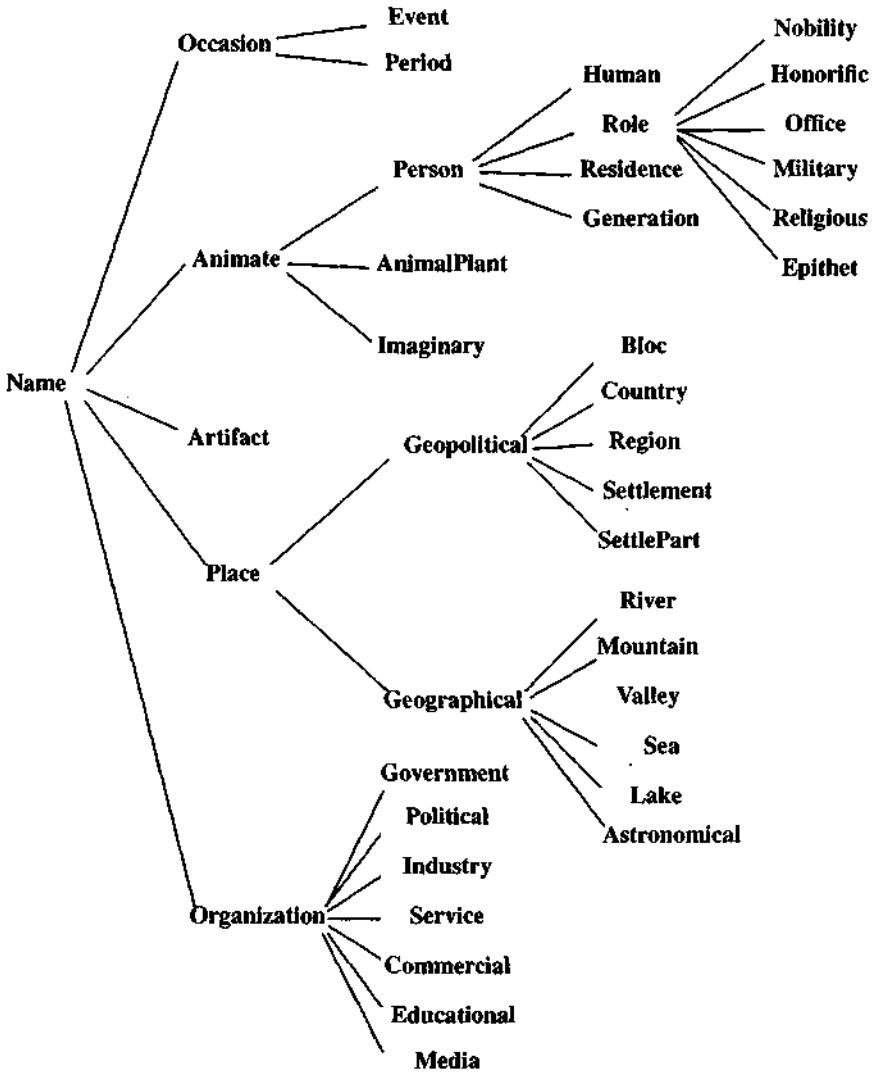

Figure 1: The hierarchy of semantic types.

The information contained in the SEMANTICS zone will be used by information retrieval, extraction and summarization engines because it helps to separate proper names into rubrics which could be used in formulating search patterns. Semantic categories and their proper assignment are also important for translation. For example, the Russian proper name *Moskva* should be translated as "Moscow" when it refers to the city and (transliterated) as "Moskva" when it refers to the river. In this example, correct translation would not be facilitated if the higher-level semantic category **Place** were assigned to the word *Moskva* in the Russian monolingual onomasticon Indeed, while both *Moskva* (city) and *Moskva* (river) are, in fact, places, in order to successfully resolve this ambiguity, one needs to assign the category **Settlement,** or at least **Geopolitical!)** to the former and the category **River** (or at least **Geographical!)** to the latter. In some languages proper names are not capitalized the way they are capitalized in English, and in some cases capitalization depends on whether the word is a part of a proper name. For example, the Russian phrase *mayor Pronin* requires capitalization of both components when translated into English *(Major Pronin).* The category **Honorific** is assigned to the word *mayor* followed by a **Last** name *Pronin*, which determines that in English the word *Major* must be capitalized. If the word *mayor* is not followed by a personal name it should be rendered without capitalization.

The USAGE zone defines restrictions on the usage of some word. The sub-zones of the USAGE zone include **Time** (**Contemporary** or **Archaic**), **Geography** (location where the name is used), **Domain** (the sublanguage, if any), **Style** (the level of formality, e.g. **Formal, Colloquial**), **Connotation** (e.g., **Pejorative**). This information can be useful for stylistic attribution of texts. The ALIASES zone contains a link to another entry which refers to a different name of the same entity. For example, for *St.Petersburg* the value of this zone is *Leningrad.* The information contained in the ALIASES zone helps coreference resolution. The EXAMPLE zone lists usage examples from a corpus. The ANNOTATION zone contains the modification audit trail for an entry.

The TRANSLATION zone lists equivalents for every target language. In case of an onomasticon other than English, this zone contains only English equivalents. In case of English the TRANSLATION zone will contain a list of all other languages which reference the actual translation in a corresponding onomasticon. In translation, proper names are often left unchanged (if the same alphabets are used in both the source and target languages), or transliterated (if the alphabets are different). Of course, in a large number of cases where transliteration does not work, translations into English must be used (e.g., the German *Wien*, the Russian *Vena*, the Ukrainian *Viden'*, the French *Vienne* are translated as the English *Vienna).*

3. The Methodology of Acquisition

In accordance with the kind of the resource (a multilingual knowledge base for proper names), the acquisition work is divided into several phases:

1. Identification and systematic listing of semantic categories of proper names. The first task is systematization of the inventory of tags for Spanish, French, German, Chinese, Japanese, Korean, Persian, Arabic, Serbo-Croatian, Russian, Turkish in terms of fillers of the onomasticon entry zones.
2. Extracting proper names from corpora. This stage includes development of rules for extracting lists of proper name candidates from corpora in the languages involved. This can be done for every particular language depending upon its regularities fixed in standard grammars and other available resources. For example, for many languages one can exploit the fact that proper names are capitalized; for the English language the tool can be trained on the part-of-speech tagged WSJ corpus. The development of the rules is performed manually, parallel for all languages in question followed by manual check of the proper candidate list.
3. Development of onomasticon acquisition interface (see below).
4. Compilation of a list of senses for each proper name.To perform this task we use the list of proper names derived at the previous stage and raw corpora. The task, thus, consists in the analysis of the corpus and manual listing of the senses detected in it. The information is then entered into the onomastica entries while thoroughly analyzing the list of senses compiled before.
5. Compilation of lists of fillers of the entry zones for each proper name. This task can be done in parallel with the previous stage. The methodology is the same.

4. The Acquisition Interface

To acquire and maintain the onomastica, a state of the art acquisition interface and an administration tool is being developed. Onomasticon entries are stored in the ObjectStore database as Java objects. The database and tools support multiple users who can access individual entries: concurrence is defined at the entry level and supports one writer and multiple readers. However, the administrator could obtain locks for an entire dictionary or even for the entire database. For each onomasticon there are methods to:

- Retrieve an entry given a key (primary or secondary);
- Ask if a key already exists;
- Add an entry with a new key;
- Remove an entry;
- Iterate through all entries and apply the method to each entry.

When an entry is obtained, any modification of that entry will become persistent upon successful completion of the transaction. For each of the onomasticon, it will be possible to produce (both manually and automatically):

- The list of entries which are not checked;
- The list of entries which are checked but contain only the primary key (the citation form of a superentry);
- The list of entries created or modified by a given author with optionally the specification of a date;
- The list of entries (for onomastica of languages other than English) which do not have an English translation;
- The list of entries for the English onomasticon with the specification of a list of target languages.
- A list which can be the intersection of any of the above.

These lists can be stored persistently in the database. These methods are available in the administration tool.

The onomasticon database can be viewed as a table where lines are proper names and columns are languages: at any point during the development, this table has many empty cells: the goal of the development is to fill these cells. Thus, the administration graphical user interface includes functions to list these empty cells in order to prepare some acquisition task.

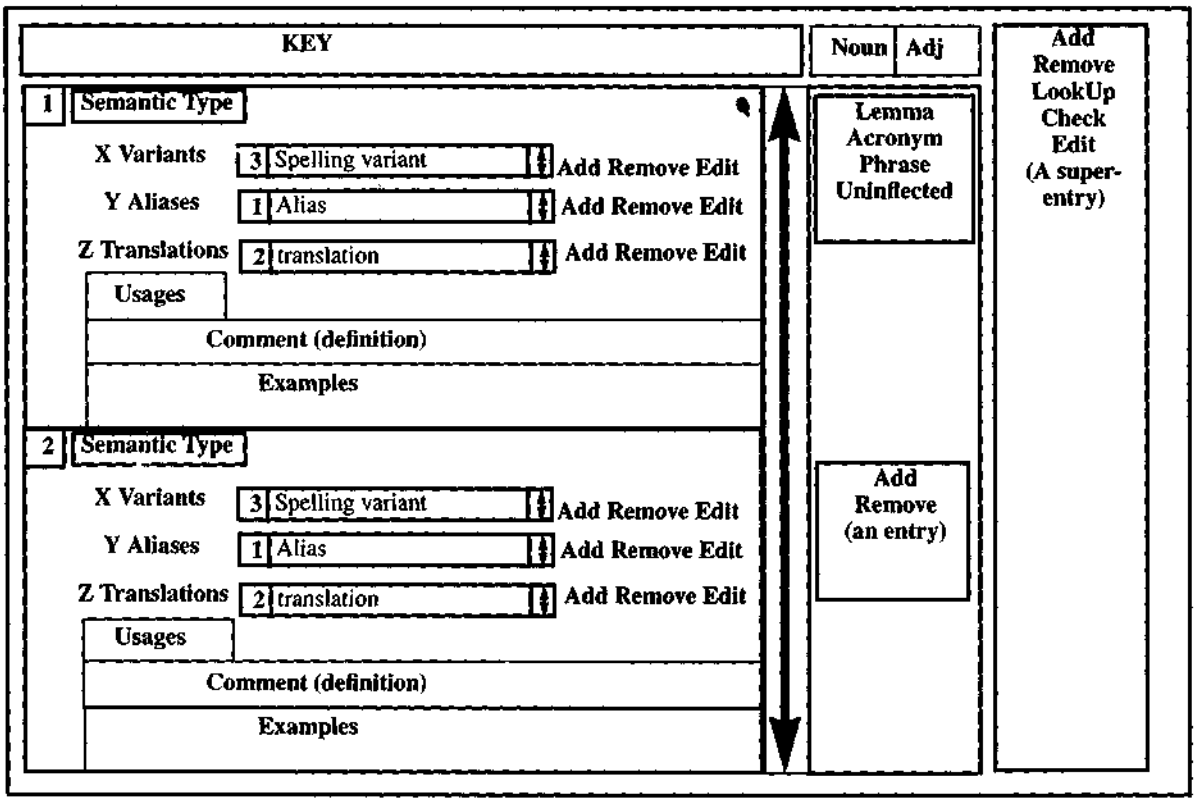

Figure 2: The acquisition of interface.

When starting the interface, the user is already presented with the full interface (Figure 2), all fields being empty. The user has to choose a particular source onomasticon on which to work using a file menu. Once the source onomasticon identified, the acquirer is presented with a list of entries, each numbered, with the zones either empty or filled with default values. For example, the default value for the TYPE zone is "lemma" (see above), for the TRANSLATION zone default is transliteration. To edit a value the pop-up menus of the values for every zone are provided, the acquirer needs only to click on the one chosen. When a value of a zone is selected in the source language, the target language entry takes the same value as the default. Finally, both the source and target panel have a lookup widget in which the acquirer can type in a string to look up in the dictionary.

5. Examples of superentries

5.1.English

CITATION-FORM: ***New-Yorker***
FORM-TYPE: phrase
PART-OF-SPEECH noun
SENSE: ***Nev-Yorker-1***
ORTH-VARIANTS: {none}
GRAMMAR-FEAT: masculine, feminine
SEMANTICS: residence;the semantic path: Name -> Animate ->Person
USAGE: contemporary
ALIASES: {none}
TRANSLATION: { Spanish, French, German, Chinese, Japanese, Korean, Persian,Arabic, Serbo-Croatian, Russian,Turkish}; these are the references to the entries in other onomastica containing *NewYorker-1* in their TRANSLATION zone;
EXAMPLE: *New Yorkers are brave people.*
ANNOTATION: Svetlana, Dec 3,1997: created;

SENSE: ***New-Yorker-2***
VARIANTS: {none}
GRAMMAR-FEAT: {singular}
SEMANTICS: media; the semantic path: Name -> Organization -> Media
USAGE: contemporary
ALIASES: {none}
TRANSLATION: {Spanish, French, German, Chinese, Japanese, Korean, Persian,Arabic, Serbo-Croatian, Russian, Turkish}; these are the references to the entries in other onomastica containing *New Yorker-2 in* their TRANSLATION zone;
EXAMPLE: ***New Yorker*** *is* a very *popular magazine*
ANNOTATION: Svetlana, Dec 3,1997: created;

5.2. Russian

```
CITATION-FORM: Moskva
   FORM-TYPE: lemma
   PART-OF-SPEECH noun
   SENSE: Moskva -1
   ORTH-VARIANTS: {none}
   GRAMMAR-FEAT: feminine
   SEMANTICS: settlement; the semantic
   path: Name -> Place -> Geopolitical
   USAGE: contemporary
   ALIASES: {none}
   TRANSLATION: Moscow; only the
   English translation is given,
   references to translations into
   other languages are given in the
   entry Moscow of the English
   onomasticon;
   EXAMPLE: Moskva - stolitsa Rossii.
   ANNOTATION: Svetlana, Dec 3,1997:
   created;
   SENSE: Moskva-2
   ORTH-VARIANTS: {none}
   GRAMMAR-FEAT: {feminine}
   SEMANTICS: river; the semantic
   path: Name -> Place -> Geographical
   USAGE: contemporary
   ALIASES: {none}
   TRANSLATION: Moskva; only the
   English translation is given,
   references to translations into
   other languages are given in the
   entry Moskva of the English
   onomasticon;
   EXAMPLE: Moskva techet netoroplivo.
   ANNOTATION: Svetlana, Dec 3, 1997:
   created;
```

6. Conclusions

Once the onomastica become widely available, many new applications and uses will appear. In addition to the obvious uses in multilingual text processing applications, we also suggest using this resource for the following:

- Developing automatic recognizers for proper names as part of ongoing research on NLP systems; this could be done using statistical methods or machine learning of recognition rules; this task will be aided if the program were able to recognize special suffixes which are not used in words other than proper names;
- Automatic assignment of semantic categories to proper names; this could be done by generating patterns using the text occurring around the proper nouns; it can add information to the POS tagging as well as assist in the determination of context, disambiguation of neighboring words, anaphora resolution and the creation of internal representations of sentences and text.
- Providing additional knowledge to spell check the proper names;
- Developing modules for computer-aided instruction systems, for example, to teach students how to spell proper names correctly;

The work on the onomasticon is tightly coupled with the work on systems for recognizing proper names in text. Several research groups have studied recognition of proper names (e.g., Borgman & Siegfried, 1992; Wakao et al., 1996; Strzalkowski & Wang, 1996; Mani & MacMillan, 1996). However problems inherent in multilingual NLP require further enhancements to the recognizers, for example, due to widespread syntactic, semantic and positional ambiguity between proper and common names. For example, a) syntactically, the Russian toponym *Mirnyj* is ambiguous with the common adjective meaning *"peaceful;"* b) semantically, *"Apple"* may refer to the company or to a fruit; c) positionally, as mentioned above, words are translated (at least, capitalized) differently depending on whether they are a part of a multi-word proper name (e.g., in the Russian *Ural'skij gosudarstvennyj universitet* only the first word is genuinely a proper name, but it will be translated into English as *"The Ural State University,"* with the last two words capitalized). The onomasticon we are developing will be included as a component in the machine translation environments: Expedition (Nirenburg, 1998), Corelli, and Shiraz and the MINDS summarizer (http://crl.nmsu.edu/Research/Projects), all under development at NMSU CRL.

References

Borgman, C.L. & Siegfried, S.L. (1992). Getty's Synonym and Its Cousins: A Survey of Applications of Personal Name-Matching Algorithms. *Journal of the American Society for Information Science (JASIS)*, 43(7).

Cowie, J. (1995). Description of the CRL/NMSU Systems Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Morgan Kaufmann.

Mahesh, K and Nirenburg, S. (1995). A situated ontology for practical NLP. In the *Proceedings of IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing.* Montreal, August 19-21.

Mani, I. & MacMillan, T.R. (1996). Identifying Unknown Proper Names in Newswire Text. In B.Boguraev, and J.Pustejovsky (eds) *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge MA.

Nirenburg, S. (1998). Project Boas: "Linguist in the Box" as a Multi-Purpose Language Resource. *Proceedings of the First International Conference on Language Resources and Evaluation.* Granada. May 28-30.

Strzalkowski, T. & Wang, J. (1996). A Self-Learning Universal Concept Spotter. In *Proceedings of*

Coling-96: The 16th International Conference on Computational Linguistics. Copenhagen.

Wakao, T., Gaizauskas, R. & Wilks, Y. (1996). Evaluation of an Algorithm for the Recognition and Classifications of Proper Names. In *Proceedings of Coling-96: The 16th International Conference on Computational Linguistics.* Copenhagen.

Zajac, R. (1998). The Habanera Lexical Database Management System. *Proceedings of the First International Conference on Language Resources and Evaluation.* Granada, May 28-30.