Disambiguation of Proper Names Using Finite-State Local Grammars

Elisabete Ranchhod[1,] and Samuel Eleutério[2]

[1]University of Lisbon, FLUL [2]Technical University of Lisbon, IST
[1]Alameda da Universidade, [2]Av. Rovisco Pais
[1]e_ranchhod@fl.ul.pt, [2]sme@ist.utl.pt

Abstract. Like common noun phrases, proper names contain ambiguous conjoined phrases that make their delimitation and classification difficult in text. This paper presents a finite-state approach to the disambiguation of Portuguese candidate proper name strings containing the coordinating conjunction *e* (and). In such name strings, the conjunction can denote a relation between two independent names, but it can also be part of a multiword proper name. The coordination of multiword independent names may involve ellipsis of some lexical constituents, which causes additional difficulties to proper name identification and classification.

Keywords: ambiguous named entities, ambiguous proper names, information extraction, conjunction, coordination, ellipsis, finite-state grammar, local grammar.

1 Introduction

This paper presents an experiment aiming to resolve ambiguity caused by the coordinating conjunction *e* (and) in Portuguese named entity recognition, and describes the heuristics used to disambiguate candidate proper name strings containing that conjunction. We are interested in naming expressions referring to entities as *persons* (given and family names, titles, etc.), *organizations* (corporations, public and governmental institutions, etc.) and *locations* (regions, countries, cities, mountains, rivers, etc.).

As well as in other Romance languages ([1] for French; [2] for Spanish), in Portuguese the formal criterion more operational for proper names, that distinguishes them from common nouns, is the initial capital letter. In fact, in Portuguese texts, the use of initial capitalized words in the interior of a sentence, i.e. in an unambiguous position, typically indicates the presence of a proper name (or of part of a multiword proper name). Although such criterion presents some difficulties, mainly due to case inconsistency[1], in this paper we assume that Portuguese names correspond to a word or a string of words initialized with a capital letter (e.g. *Lisboa* 'Lisbon'; *Hospital de Santa Maria* 'Saint Mary's Hospital')[2].

The challenging problem with candidate named entity strings is the correct delimitation and subsequent classification of entities [2]. In fact, a string of consecutive initial capitalized words – containing potentially functional words, such as determiners and prepositions, or the coordinating conjunction *e* –, can represent either a single name or a series of embedded or independent names:

(1) a assinatura do *Acordo Geral sobre Tarifas e Comércio*.
(the signing of the General Agreement on Tariffs and Trade)

(2) a vantagem de *Barack Obama sobre Hillary Clinton e John Edwards*.
(the advantage of Barack Obama over Hillary Clinton and John Edwards)

[1] For instance, *ministro* (minister) can be capitalized or not in the same text: *o ministro; o Ministro*.

[2] Whenever possible and appropriate, we give an approximate translation in English for our Portuguese examples. We omit the translation for person names, and for those names whose spelling coincides largely with that of the original language (e.g. *Bonnie e Clyde* 'Bonnie and Clyde').

Despite their similar structure, example (1) corresponds to a single named entity (the naming of a particular agreement) that contains a preposition (*sobre* 'on') and a conjunction (*e* 'and'), while example (2) illustrates a discourse structure containing three independent person names connected by a preposition and a conjunction. This means that, as with common noun phrases, proper names exhibit structural ambiguity in prepositional phrase attachment and in conjunction scope.

In this paper we are concerned with the disambiguation of proper name strings containing the coordinating conjunction *e*. The ambiguity caused by this conjunction is not negligible. Using simple algorithms, we could estimate that, in a journalistic corpus with 20,463 candidate names, the conjunction occurs 1038 times, which indicates that around 5% of the strings are ambiguous. This proportion is analogous to estimations reported for English before [3], where in a sample of 545 candidate named entity strings, 31 conjunctions were found.

Determining the correct analysis of ambiguous strings with coordinating conjunction is important for Portuguese Named Entity Recognition and Classification, and obviously for all applications that rely on named entity extraction.

2 Related Work and Motivation

Despite the intensive work on named entities in the last fifteen years (for a comprehensive survey, see [4]), research on the problems that coordinating conjunctions pose to named entity recognition and classification is still fairly limited.

For English, the ambiguity caused by conjunctions has been noticed since one of the first research papers on the recognition and extraction of company names [5]. Most recently, [3] have addressed the problem of disambiguating candidate name entity strings with conjunctions using a "machine-learned classifier" and "limited knowledge sources: […] gazetteers that contain the most frequent proper nouns that appear in [the] corpus", and "the so-called 'name-internal' properties". Our project is closely related to that work, but our approach relies more deeply on linguistic and contextual knowledge. We constructed proper noun grammars that use "name internal evidence" but are also "context sensitive" [6]. As we will see later (section 4.), contextualization is an important external evidence for entity delimitation and classification.

For Portuguese, [7] observe that the coordinating conjunction *e* may cause structural ambiguity, gave some examples of that ambiguity, and mention that its resolution would require deep syntactic parsing of the text. In the HAREM evaluation contest [8], the proportion of ambiguity and mistagging caused by the erroneous analysis of name coordination have not been evaluated. We participated in the HAREM evaluation, using a system that scored high, but was not capable of handling conjunction in an effective way. The present paper describes the work developed since then on the topic.

For the purpose of this experiment we have considered four semantic classes of proper names: PESSOA (person), names referring to people; LUGAR (location), geographical names referring to countries, cities, mountains, etc., and ORGANIZACAO (organization), names denoting companies, governmental institutions, etc. These are the three main types of names, collectively known as “enamex” since the MUC-6 evaluation (1995). The fourth category, DIVERSO, is a residual, semantically heterogeneous class (“miscellaneous” in the CONLL conferences). It includes naming expressions for events, artifacts, book and movie titles, etc.

3 Linguistic Description

Our concern is the correct delimitation and classification of Portuguese proper names in strings that contain the coordinating conjunction *e*[3]. We have extracted from the corpus all the proper name strings containing at least one and at most two conjunctions. The linguistic analysis of such data let us distinguish two different values of the conjunction: (i) the conjunction is a constituent of the name, i.e. the named entity contains "an internal conjunction" [3], or (ii) the conjunction denotes a relation between two independent names, i.e. the conjunction is name-external. As with common noun phrases [9], when the conjunction links two multi-word independent names some name constituents can be ellipted from one of the names. The ellipses (of part of a name) causes additional difficulty to name identification and classification.

In next sections we describe and illustrate the main linguistic types of coordinated name structures.

3.1 Internal Conjunction

We consider that the conjunction is name-internal if it is an inherent and distinctive element of the proper name. In Portuguese, the coordinating conjunction *e* can be an internal constituent of the four pre-defined semantic types of entities. A few examples of proper names containing conjunctions follow:

- PESSOA: family names (Maria *Brito e Cunha*), artistic groups and bands (*Chutos e Pontapés*; *Despe & Siga*),
- LUGAR: countries (*São Tomé e Príncipe*), streets containing person and geographical names (Rua *Melo e Costa*),
- ORGANIZACAO: institutions (*Câmara de Comércio e Indústria* 'Chamber of Commerce and Industry'), companies (*Águas do Douro e Paiva*),
- DIVERSO: mentions in texts to events, museums, books, operas, movies etc. (*Crime e Castigo* 'Crime and Punishment', *Tristão e Isolda* 'Tristan and Isolda', *Bonnie e Clyde* 'Bonnie and Clyde').

In all these instances, proper names are multiword nouns, constituted of at least three words, of which one is the lower case coordinating conjunction.

3.2 External Conjunction

A totally different and most frequent situation is that where the coordinating conjunction is not part of the name, but it denotes a relation between two independent names. This is the case illustrated by the example:

(3) Futre, Rui Águas e Rui Barros
(person names)

The conjunction appears at the end of an enumeration, linking the last two person names.

Ellipsis. Coordination often involves ellipsis, which is a means of avoiding repetition [9]. For example, the repetition of *Carolina* is avoided in:

(4) Carolina do Norte e do Sul
(North and South Caroline)

[3] The conjunction *e* can be represented by its variant spelling &, but this is rather infrequent in Portuguese. For comparison, in our corpus, there are 20 forms &, against 1018 forms *e*. In all these occurrences, & is a name-internal conjunction.

The ellipsis of *Carolina* before *do Sul* (i.e. in the second conjoined noun phrase) leaves the second name lexically incomplete. Following [9], we will call this reduction *anaphoric ellipsis*, since it implies the recuperation of information (a word, in this instance) mentioned before in the discourse structure.

Ellipsis can involve the omission of a word that will be mentioned later in the discourse. For this reason, we will call that reduction *cataphoric ellipsis*. In the following example, the family name *Rocha*, common to *Andrée* and *Clara*, is omitted after *Andrée*, only appearing later, next to *Clara*:

(5) a mulher e a filha de Torga, *Andrée e Clara Rocha*
(the wife and the daughter of Torga, *Andrée and Clara Rocha*)

These few examples illustrate that the coordination of two proper names often involves the removing of lexical items from one of them: in anaphoric ellipsis, the second name is incomplete; in cataphoric ellipsis, on the contrary, it is the first name that is affected by lexical reduction. In both types of ellipsis the missing words can be exactly recovered (*Carolina do Norte e Carolina do Sul*, example (4); *Andrée Rocha e Clara Rocha*, example (5)).

But in coordinated proper names the ellipted items need not be identical in all respects:

(6) Câmaras Municipais de Braga e do Porto
(Municipalities of Braga and of Oporto)

Taking into consideration the real world, the interpretation of (6) is obviously that only two municipalities are mentioned in the discourse, i.e. *Câmara Municipal de Braga* and *Câmara Municipal do Porto* (Municipality of Braga and Municipality of Oporto). What is missing in the second conjoined name of (6) is *Câmara Municipal* (and not *Câmaras Municipais*). So, in situations like this one, the correct treatment of ellipsis requires the reconstruction of two noun phrases and the identification of two singular names: *Câmara Municipal de Braga e Câmara Municipal do Porto* (Municipality of Braga and Municipality of Oporto).

The grammars that we have designed can handle ellipsis, but they are not totally satisfactory. This is one of the topics that needs further improvement.

4. Data and Resources

The textual data were extracted from a general-purpose journalistic corpus: *CETEMPúblico*. This is a Portuguese untagged public corpus, consisting of excerpts of the Portuguese daily newspaper *Público* that contains about 180 million words (for technical information about the corpus, see [10]).

For the purpose of the work described here, a subcorpus (corpus from now on) with 594,709 tokens, of which 260,071 are words integrated into 11,600 sentences, was drawn randomly from that large corpus.

Using a case-sensitive general tagger, based on a large-scale lexicon, developed previously [11], we identified in that corpus 20,463 candidate named entities, i.e. sequences of at least two consecutive initial capitalized words, containing a coordinating conjunction[4]. We obtained 1018 instances of the conjunction '*e*' and 20 instances of its variant '*&*'.

We also permit strings to include lower case functional words, articles: *o, a, os, as* (the) and prepositions: *de* (of), *para* (for), *sobre* (on), as well as some punctuation marks (apostrophe, hyphen, comma, quotes). Articles, prepositions, apostrophes and hyphens can be found inside Portuguese multiword names, in particular person and geographical names:

(7) África do Sul (South Africa, lit. 'Africa of the South')

[4] In the present study, we have restricted ourselves to candidate named entity strings that contain a single conjunction.

(8) João d'Ávila (person name)

(9) Ernesto de Melo e Castro (person name)

(10) Trás-os-Montes (Portuguese region)

In the example (7) the preposition *de* (of) is contracted with an article (*do* = *de*+*o*). The same preposition is truncated, and an apostrophe replaces the reduced vowel, in example (8). The aristocratic, or aristocratic-like, family name of example (9), *de Melo e Castro*, incorporates the preposition *de* and the conjunction *e*. In (10), *Trás-os-Montes*, as well some other geographical (e.g. the city *Dar-es-Salam*) and organization names, contains an internal hyphen linking the lexical constituents. A few complex names (i.e. multiword names with nested names) can contain commas:

(11) Seminário de História Medieval, Moderna e Contemporânea
(Seminar in Medieval, Modern and Contemporary History)

In general, commas are found in independent name apposition, and, since the conjunction represents the end of an enumeration, it can be used as an external evidence for conjunction analysis. Brackets and quotes include named entities but they are not found inside names.

We did not consider as candidate name entities capitalized words appearing after a punctuation mark requiring capitalization, since most of them correspond to single (i.e. not conjoin) uppercase initial words, the majority of which are articles and other functional words.

For parsing and tagging proper names, in addition to the general tagger, we used the following lexical, syntactic and semantic resources: a gazetteer comprising around 16,000 lexical entries in total, a list of 337 designators and a list of 380 trigger words indicating or accompanying persons, organizations, locations, etc.; a set of finite-state grammars (70), which use the lexical knowledge represented in the gazetteers, designator and trigger word lists, and describe relevant local syntactic information.

The experiments were conducted using Unitex, a modular open source toolkit based on finite-state technology [13].

In next paragraphs we describe in more details these data and resources.

Gazetteers. Lexical entries are single and multiword person, location, organization and other proper names. Some names with internal conjunction are included in these word lists (e.g. the Portuguese surname *Melo e Castro* (Cf. example 10)). Lists of names used contain:

- *person*: about 6,500 person names;
- *location*: about 5,500 major country, province, state, city and town names;
- *organization*: about 3,000 company and governmental and public institution names;
- *diverse*: about 1000 proper names mentioning gardens, bridges, months, books, newspapers, etc.

The word lists have been extracted semi-automatically from texts and word lists published on the Web, and enriched manually with linguistic features afterwards [11]. Linguistic attributes include PoS, usually Noun (N), morphological (when appropriate) and semantic classification. For example, *N+Geo+Gep* means that the entry is a noun (N), a geographical name (Geo) of subtype geo-political (Gep).

Designator and trigger word lists. We distinguished between designators and trigger words. Designators are lower case words found in the left or right context of a pre-defined pattern of initial capitalized words that permit to include them in a class, even subclass, of proper names. For example, *província de* (province of) and *rio* (river) may consistently allow geographical names to be determined and included in the semantic class LUGAR (location): província de *Cabinda*; rio *Ganges*. Nouns indicating human occupations (e.g. *arquitecto-arquitecta* ('architect', masculine and feminine), *professor-professora* (professor) often precede person names (professor *Moniz Pereira*). Designators are particularly useful to recognize and classify foreign proper names not included in the gazetteers, but mentioned in journalistic texts (e.g. professor *Donald Kettl*; actriz (actress) *Helen Mirren*; armazéns (store) *Marks and Spencer*).

Trigger words, in turn, are upper cased words that are included in multiword names and permit, as designators do, the class of that names to be discovered. For example, human titles (e.g. Sr., Eng., Prof., Ministro) are regularly upper cased words and abbreviations accompanying person names; *Companhia de Seguros* (Assurance Company), *Grupo Segurador* (Assurance Group) and *Instituto* (Institute) are examples of trigger words included in organization names (companies and institutions, respectively).

We took advantage of morphological features of Portuguese (in particular, singular, plural) and have listed separately singular and plural designator and trigger words. A designator in the plural before a string that contains two conjoined proper names is in general an external evidence for two independent names: *tenores José Carreras e Plácido Domingo* (tenors *José Carreras and Plácido Domingo*); *ilhas Terceira e São Miguel* (islands *Terceira* and *São Miguel*), *rios Ardila e Guadiana* (rivers *Ardila* and *Guadiana*).

Designators and trigger words were manually collected by applying the dictionary of proper names to texts, and looking at their context in those texts. Using Unitex facilities, designators and trigger words were compiled into a finite-state recognizer to build and tag proper names.

Finite-state local grammars. Designators and trigger words are very helpful for finding names and for classifying them semantically, but they are insufficient to handle accurately the conjunction's ambiguity.

Based on linguistic analysis (see sections 3.1, 3.2., 3.2.1), we have written grammar rules to handle the ambiguity. Such rules describe restricted patterns and constraints specific to each class of coordinated proper names. There are 70 rules in total, of which 23 for person, 15 for organization, 9 for location, and 23 for the semantically heterogeneous group. The set of these context-dependent rules can be viewed as a *local grammar* [13] for proper names. Local grammars capture typical patterns associated to designators and trigger words, but they can also represent accurately dependencies between words, and general syntactic relations (e.g. subject-verb relations), without applying to more powerful syntactic formalisms. They also make use of part of speech tags as well as of linguistic information included in the gazetteers.

Local grammars can be efficiently compiled into finite state transducers [13], and then be applied to texts to parse proper names. We compiled our grammars into FST using Unitex tools.

Fig. 1 represents a finite-state transducer that implements a local rule (or a local grammar [14]) that captures two independent coordinated person names.

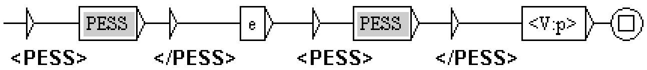

Fig. 1: A graphical representation of a finite-state transducer that implements the rule *P1 e P2.*

The rule *P1 e P2* recognizes two coordinated person names that are the subject of a plural verb form (<V:p>[5]). The rule assumes that, in such conditions, the conjunction is name-external. The rule is represented in a graph that references a person name sub-grammar, *PESS,* i.e. the main graph calls the sub-graph *PESS*. This sub-graph is partially represented in Fig. 2.

[5] Notice that the corpus has been previously tagged by a case-sensitive general tagger based on a comprehensive lexicon.

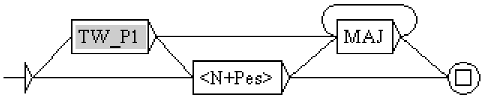

Fig. 2: A finite automaton representing the sub-grammar *PESS.*

The sub-grammar *PESS* recognizes as a person name any noun encoded in the gazetteers as <*N+Pes*>, which may be preceded by an adequate trigger word (*TW_P1*) and followed by any string of uppercase words (*MAJ*); it also predicts that any string of initial capitalized words preceded by a person trigger word is a person name.

In Fig. 1, the transducer outputs represent the sequences of tags that will be inserted into the text. The results of running the rule *P1 e P2* over the corpus are illustrated by the examples:

(12) <PESS>Pinto Balsemão</PESS> e <PESS>D. José Policarpo</PESS> foram felicitados

(13) <PESS>João Paulo II</PESS> e <PESS>Fidel Castro</PESS> trataram, com a formalidade

(14) <PESS>Cavaco Silva</PESS> e <PESS>António Guterres</PESS> conseguiram, cada qual a

The collection of the seventy local grammars has been combined and compiled into a single finite-sate transducer grammar, which was applied in one pass of the parser to proper name processing.

5 Results and Evaluation

After the processing of the corpus, using Unitex and the linguistic knowledge described in 4., the system produced a tagged version of the original text. In this annotated corpus all the identified proper names, appearing in coordinated strings, were marked up with SGML tags that specify their semantic class. For reminding, we have constituted four semantic classes, tagged as: <PESS> for person names, <ORG> for organizations, <LUG> for locations, and <DIV> for other diverse semantic types of naming expressions.

The annotation processing also took into account the conjunction type present in the string (name-internal, name-external, elliptical construction). The tagged corpus was automatically scored against the manual tagging of the same text performed by a linguist, who identified 1023 proper names involving the conjunction (*e* and &). The overall results of that evaluation are shown in Table 1.

Table 1: Overall Precision and Recall Scores

	Name Internal Conjunction	Name External Conjunction			Total
		With no Ellipsis	Anaphoric Ellipsis	Cataphoric Ellipsis	
Manual	205	803	11	4	1023
Tags	180	720	9	1	910
Correct	172	607	9	1	789
Precision	0.96	0.84	1.00	1.00	0.86
Recall	0.84	0.76	0.82	0.25	0.77
F-measure	0.89	0.80	0.90	0.40	0.82

Table 1 shows that, in our journalistic corpus, the more significant set is constituted of independent proper names linked by the conjunction (80%). Of these only a small proportion (1,5%) is affected by ellipsis. The contribution of the system to the analysis of the nature of the conjunction is also illustrated. The system scored satisfactorily, except for cataphoric ellipsis. However, the small number of instances of such constructions makes the results statistically insignificant. On the other hand, ellipsis detection may be a non trivial problem even for a human expert. Yet, these results indicate that more sophisticated resources and more extensive data are needed for analyzing elliptic constructions in an adequate manner. The best results are for name-internal conjunction.

The human annotator also included proper names into one of the four pre-defined semantic classes. Table 2 shows the human analysis, and evaluates the results not only for the system as a whole, but for each semantic class of names.

Table 2: Overall Scores of Semantic Analysis

Semantic Class	Manual Tagging	System annotation				
		Tags	Correct	P	R	F
PES	727	635	589	0.93	0.81	0.86
ORG	443	468	350	0.75	0.79	0.77
LUG	512	411	382	0.93	0.75	0.83
DIV	164	127	108	0.85	0.67	0.74
Total	1846	1641	1429	0.87	0.77	0.82

As a complementary view to conjunction analysis, we have examined in more details the data involving name-internal conjunction. Table 3 shows the achieved scores for each semantic class of proper names.

Table 3: Name-internal Conjunction Scores

Semantic Class	Manual Tagging	System annotation				
		Tags	Correct	P	R	F
PES	45	39	39	1.00	0.87	0.93
ORG	99	94	88	0.94	0.89	0.91
LUG	16	9	8	0.89	0.50	0.64
DIV	45	38	37	0.97	0.82	0.89
Total	205	180	172	0.96	0.84	0.89

The F-measure is better than its value in Table 2. The fact that the conjunction variant *&* (20 occurrences) is always internal to the name may contribute to this result.

5 Conclusion

In this work we have concentrated on the problem of the correct delimitation and classification of Portuguese proper names in strings that contain the coordinating conjunction *e* (and). Such strings are ambiguous, since the conjunction is either name-internal, i.e. it is part of the proper name, or name-external, i.e. it denotes a relation between two names.

We have shown that the various types of ambiguity can be handled satisfactorily using a small semantic lexicon and finite-state local grammars. Such local grammars do not just capture typical patterns associated to proper names, but they can also describe accurately dependencies between words, and local syntactic relations.

We have evaluated quantitatively our results against a manual tagging performed by a linguist. Unsurprisingly, the results confirm that linguistic knowledge based rules obtain better precision scores in comparison with the recall values.

In order to improve the work reported here, we believe it is necessary to further develop the syntactic component, by incorporating new and more precise local grammars.

References

1. Maurel, D.: Les mots inconnus sont-ils des noms propres? In: Purnelle, Gérald. C. Fairon, A. Dister (eds.) Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles. vol. 2., pp. 776-784, Presses Universitaires de Louvain, Louvain (2004)
2. Galicia-Haro, S., Gelbukh, A.: Complex named entities in Spanish texts: Structures and properties. In: Sekine, S., Ranchhod, E. (eds) Named Entities: Recognition, classification and use, special issue of Lingvisticae Investigationes, 30:1, pp. 69-94, John Benjamins, Amsterdam (2007)
3. Mazur, P., Dale, R.: Handling Conjunctions in Named Entities. In: Sekine, S., Ranchhod, E. (eds) Named Entities: Recognition, classification and use, special issue of Lingvisticae Investigationes, 30:1, pp. 49-68, John Benjamins, Amsterdam (2007)
4. Nadeau, D. and Sekine, S.: A survey of named entity recognition and classification. In: Sekine, S., Ranchhod, E. (eds) Named Entities: Recognition, classification and use, special issue of Lingvisticae Investigationes, 30:1, pp. 3-26, John Benjamins, Amsterdam (2007)
5. Rau, L.: Extracting Company Names from Text. In: Proceedings of the Seventh Conference on Artificial Intelligence Applications of IEEE (1991)
6. McDonald, D.: Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In: Proceedings of SIGLEX Workshop on Acquisition of Lexical Knowledge from Text, pp. 32-43 (1993)
7. Mota, C., Santos, D., Ranchhod, E.: Avaliação de reconhecimento de entidades mencionadas: princípio de AREM. In: Santos, Diana (Ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa, pp. 161-176, IST-Press, Lisboa (2007)
8. Santos, D., Seco, N., Cardoso, N.,Vilela, R.: HAREM: an Advanced NER Evaluation Contest for Portuguese. In: Proceedings of LREC'2006, Genoa, Italy (2006)
9. Quirk, R., Greenbaum, S., Leech, S., Svartvik, J.: A Grammar of Contemporary English. Longman Group, Ltd. (1980)
10. Santos, D., Rocha, P.: Evaluating CETEMPúblico, a free resource for Portuguese. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pp. 442-449, Toulouse (2001)
11. Ranchhod, E., Mota, C., Carvalho, P,: Portuguese Large-scale Language Resources for NLP Applications. In: Proceedings of the IV Conference on Language Resources and Evaluation, LREC, pp. 1755-1759, Lisboa (2004)
12. Paumier, S.: Unitex 1.2. User Manual. http://www-igm.univ-mlv.fr/~unitex/
13. Gross, M.: The Construction of Local Grammars. In: Roche, E., Schabes, Y. (eds.) Finite-State Language Processing, pp. 329-354, The MIT Press, Cambridge, Massachusetts (1997)
14. Mohri, M.: Local Grammar Algorithms. In: Arppe, C., Heinämäki, L., Miestamo, P. (eds.) A Finnish Computer Linguist: Kimmo Koskenniemi. Festschrift on the 60th Birthday, pp. 87-96, CSLI Publications (2005)