Using the Incremental Finite-State Architecture to create a Spanish Shallow Parser

Núria Gala Pavia
Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France
Nuria.Gala-pavia@xrce.xerox.com

Abstract

This paper describes a Spanish Shallow Parser built using the Incremental Finite-State Parsing Architecture (IFSP). This approach for Shallow Parsing permits a constructivist syntactic analysis: each transducer uses as input[1] the result of the analysis given by the previous transducer for obtaining a more accurate syntactic analysis. The output is a bracketed and annotated text where main segments and syntactic functions are identified. The different transducers which make up the shallow parser are built using regular expressions which describe the syntactic characteristics of the language to parse, in this case Spanish.

1 Introduction

This paper presents a Spanish shallow parser for unrestricted Spanish text that uses the Incremental Finite-State Parsing (IFSP) architecture ([Aït-Mokhtar and Chanod, 1997]). The aim of the incremental finite-state parser is to provide syntactic annotations that later will be used for extracting dependencies.

Shallow parsers perform partial syntactic analysis over unrestricted text. Research on shallow parsing has seen an important development over the last ten years. Two directions can be distinguished: a constructivist approach, based on constraints progressively added during the parsing, ([Abney, 1991], [Joshi, 1996], [Grefenstette, 1996]) as well as a reductionist approach, in which restrictions are applied to eliminate potential analysis already provided ([Karlsson et al., 1995], [Chanod and Tapanainen, 1996]).

The applications for shallow parsing are useful for large scale texts. These applications include knowledge extraction, information retrieval as in the FASTUS system ([Appelt et al., 1993]), word sense disambiguation ([Dini et al., 1999]), translation memory and multilingual comprehension aids.

In Spanish, very few work has been done on parsing. One example is that of [Atserias et al., 1998] (based on charts) in which an incremental methodology is followed to obtain syntactic annotated corpora, using a context-free grammar instead of finite-state rules.

2 The IFSP approach

The IFSP approach for shallow parsing permits a constructivist syntactic analysis performed by different transducers making up a cascade. This approach permits to increase the efficiency and the robustness of the parser and to reach a broad coverage of linguistic phenomena.

Each transducer is specialized on detecting a specific chunk or syntactic function involving different levels of complexity. The order of the transducers on the cascade depend on the language being parsed, each language having particular syntactic features to be analyzed independently from other languages.

[1] The input is unrestricted text (technical documentation, newspaper articles, web pages, etc.) tagged and preprocessed.

The identification of a chunk or a syntactic function is marked using syntactic annotations, represented by reserved symbols, which are used by the following transducers on the cascade. This strategy permits to use intermediate annotations which help at a certain stage of the analysis but can be later removed for the final result.

The transducers identifying the different kind of clusters are compiled from regular expressions [Karttunen et al., 1997][2]. These regular expressions have a simple syntax but can create very complex transducers covering complex syntactic phenomena. This is one of the advantages of the finite-state approach for shallow parsing.

3 The Spanish Parser

As mentioned above, the IFSP approach leads to a partial syntactic analysis which is done following various steps. The Spanish shallow parser described in this section shows the application of this methodology.

There is first a preliminary stage of preprocessing mainly consisting of refining some tags given by the part-of-speech tagger. This stage is followed by the shallow parsing analysis, consisting on chunking (primary segmentation) and analyzing main syntactic functions (subject and object). Finally, the shallow parser extracts syntactic dependencies, especially subject/verb and object/verb relations ([Aït-Mokhtar and Chanod, 1997]).

The following sections describe the preprocessing as well as the three main modules of the Spanish Shallow Parser: two of them performing the syntactic analysis (primary segmentation and marking of syntactic functions) and a third one extracting dependencies.

3.1 Preprocessing

The preprocessing prepares the input text by performing two main tasks: assignment of a part-of-speech tag and refining some of these tags. The aim of the preprocessing is to obtain a more accurate input for the first transducer of the parser.

3.1.1 POS Tagging

The input of the shallow parser is a tagged text[3]. The Spanish tagger uses 55 morphosyntactic tags and 15 more tags are added by the preprocessing. Most of the tags are part-of-speech (POS) tags, some of them carrying out more precise information: about number in nouns and determiners (+DETSG, +DETPL) or about sub-groups within a group (+PRONREL, +PRONSUBJ etc.). In some cases, a specific tag can be assigned to very specific words such as *se* (+SE), *ser* and the different forms of its paradigm (+AUX) etc.

Once the POS tags are assigned, some transformations are performed in order to format the input text.

3.1.2 Refining tags

The refinement of certain tags is the last transformation performed on the input text before it becomes completely preprocessed to be parsed by the first transducer on the cascade.

The revisions of certain tags (in general grammatical words) are meant to provide a more accurate tagged input to the parsing process. The more the input is accurate, the more precise is the segmentation and the marking of syntactic relations.

One example of this is the preposition *"según"* which comes from the tagger with the +PREP tag and it is refined with +PREP_SEG in order to easily distinguish, later on the analysis, the syntactic relation corresponding to an inverted subject (in, for instance, *"El protocolo fue firmado por los diferentes representantes, según afirmó un portavoz del gobierno"*).

3.2 Primary segmentation

The purpose of marking segments is to delimitate with accuracy linguistic sequences that determine the syntactic function of a word. The principle is to recognize segments of words syntactically linked to each other or to a main word (head).

Unlike the traditional linguistic definition of phrases, a cluster (or chunk) is here defined as a sequence of words in which the last element is

[2] For a detailed presentation of the syntax and semantics of the Xerox FS calculus see: http://www.xrce.xerox.com/research/mltt/fst/home.html

[3] We use the Xerox Spanish linguistic tools which performs tokenisation, lexical lookup and part-of-speech disambiguation, based on statistical methods.

the head (hence chunks are the core sub-part of a traditionnal phrase).

Sub-clause segments can contain nominal phrases if they precede the verb. However, they do not include nominal chunks following the verb, because the linguistic analysis at this stage does not permit to identify the correct attachment of a nominal phrase on the right of a verb. This segmentation is very cautious to handle the significant structural ambiguity of the elements on the right of the head.

3.2.1 Strategy

For Spanish text, the general strategy consists on first identify verb clusters (VCs), because they present less structural ambiguity than other segments such as nominal chunks. After verb clusters, adjectival phrases (APs) are marked taking into account coordinations of adjectives.

On the next step, noun phrases (NPs) are identified. Adjectival segments preceding a noun are integrated on the noun phrase; other APs (for instance on the right of a noun or on the right of a verb) are not modified. All the annotations provided at this stage and some other additional linguistic information permit to identify prepositional phrases (PPs).

The next stage consists on identifying sub-clause boundaries. The annotations provided by the previous transducers help on the analysis of sub-clauses. This last module of the primary segmentation is quite complex due to the many syntactic possibilities marking the beginning of a principal or subordinate clause in Spanish.

3.2.2 VC segmentation

The verb cluster segmentation adds three kind of syntactic annotations depending on the type of verb group: [IV IV] for infinitive verb chunks, [VG VG] for present participle chunks and [v v] for finite verb segments.

The finite verb clusters contain necessarily a finite verb. They can also include clitics, adverbs as well as coordinations (*fueron convocados, no se lo creerán, etc.*) preceding the verb. The non-finite-verbs can contain prepositions (*por haber firmado*, *antes de salir corriendo* etc.).

Verb constructions with modals are split into two clusters ([v *no pudo* v] [IV *llegar* IV], [v *se lo ha querido* v] [IV *decir* IV]) because, as mentioned above, the last element of the segment is always a finite verb, the non-finite verb being considered an argument of the finite verb.

3.2.3 AP and NP segmentation

Adjectival phrases are recognized taking into account the +ADJ POS tag of the words of the tagged and preprocessed input text. However, at this stage it is also possible to recover an error coming from the tagger, that is, adjectives wrongly tagged as past participles. This wrong past participles will be then marked inside an adjectival cluster.

The rule to identify APs takes also into account eventual suites of coordinated adjectives because this does not imply structural ambiguity:

```
define GroupAdj [
 (PreAdj) ADJ (TCM [(PreAdj) ADJ TCM]+)
 (TCoord (PreAdj) [ADJ|PAP])
];
```

This regular expression defines an adjectival segment as a chunk containing an adjective, eventually preceded by a preadjective (comparatives as well as certain adverbs). The adjective can eventually be followed by other adjectives preceded by coma or coordination.

To identify noun phrases (NPs) different elements are taken into account: single pronoun (*[NP Tú NP] convences*), proper noun (*[NP Juan NP] convence*), noun with determinant (*[NP el protocolo NP] fue firmado*), noun without determinant (*somos [NP ciudadanos NP] de a pie*), a single determinant (*[NP el NP] de la izquierda no habla*), etc.

In sequences containing a preposition followed by several proper names it has been decided that the first proper name will be the single element of a nominal chunk, the rest of the proper names being included into a different NP.

This decision has been taken after evaluating the number of occurrences of this structure on unrestricted Spanish text. The results of the heuristic applied show that in 84% of the cases the first proper name has no links with the following proper names on a suite, whereas the contrary appears to be true in 16% of the examples. This choice provides an accurate analysis for a sentence such as *[NP "El líder NP] de [NP Rusia NP] [NP Boris Eltsin NP]"* but is source of error in *"[NP El comportamiento NP] de [NP Pedro NP] [NP Vargas Iturriaga NP]"*.

3.2.4 PP segmentation

The identification of PPs basically consists on adding [PP and PP] marks in sequences containing a preposition followed by a NP or an AP.

The parser makes the difference between prepositions contracted with determiners or simple prepositions, because the prepositional attachment of the segments appears to be slightly different. In the first case a single PP is marked (*"El encuentro trató [PP del interesante e importante proyecto PP]"*) whereas in the second case a PP has to be followed by a NP (*"Trató [PP de interesante e importante PP] [NP el proyecto NP]"*).

3.2.5 SC segmentation

The last step of the primary segmentation permits to identify sub-clause boundaries. A subclause is here defined as a segment going from the beginning of a clause to a finite verb.

To identify its boundaries, it is important to first identify reliable and less reliable beginnings of subclause, later define the levels of embedding and finally match the most reliable temporary annotations of beginning of sub-clause with the end of sub-clause tags.

The principle to apply is to give a temporary mark ([S1) to the very reliable beginnings of clause (relative pronouns, conjunctions, verb coordinations in particular contexts, specific cases like the preposition *según* followed by a verb, the preposition *desde* followed by the verb *hacer* etc.). Another annotation ([S2) is given to less reliable beginnings of clause (coordinations in other contexts). An intermediate tag (S>) is also used to mark possible ends of clause (replacing the previous v] mark).

At this intermediate stage the output for *"El problema tiene una dimensión mayor y trasciende a lo que ocurre hoy."* is:

```
[NP El^el+DETSG problema^problema+NOUNSG
NP] [v tiene^tener+VERBFIN v] S> [NP una
^un+DETQUANTSG dimensin^dimension+NOUNSG
NP] [AP mayor^mayor+ADJSG AP] y^y+COORD
[S2 [S1 [v trasciende^trascender+VERBFIN
v] S> [PP a^a+PREP lo^el+DETSG PP] [S1
que^que+QUE [v ocurre^ocurrir+VERBFIN v]
S> hoy^hoy+ADV .^.+SENT
```

As this example shows, at this stage the initial text contains the annotations delimiting the main clusters as well as some temporary tags to identify the beginnings and endings of clause.

To mark the definitive sentence boundaries we apply the strategy consisting on:

1. First match each S1 with an end of subclause (this principle is applied twice because we consider at most two embedding levels);

2. The very beginnings of sub-clause are then marked if at the same level there is a temporary end of clause;

3. Later, S> are transformed into SC] if there is a beginning of sub-clause annotation at the same level;

4. Finally S2 are matched with remaining S>.

These operations being done, the temporary tags are all removed.

In Spanish, one relevant source of problem at this stage is due to the ambiguous form *que*. This form has multiple functions some of them implying a beginning of sub-clause and some of them not. A detailed study of the contexts in which this form can appear is necessary, in order to specify in which cases *que* stands for a mark of beginning of sub-clause.

The following contexts have been observed:

- preceded by a segment that does not contain a comparative adverb and followed by a verb cluster:
  "Hay situaciones **[SC que** *[NP un presidente NP] [PP de partido PP]* **:v debe SC]** *saber afrontar."*

- preceded by a segment containing a comparative adverb (*más, menos*) but followed by a verb cluster:
  "Las opiniones **más** *apasionadas proceden de las comunidades serbias y albanesas ,* **[SC que** *:v han protagonizado* **SC]** *varias manifestaciones en el país."*

- immediately preceded by a preposition and followed by a verb cluster:
  "Los portavoces insistieron **[SC [PP en que** *PP] :v no habrá* **SC]** *una intervención aliada por tierra."*

The contexts where *que* is not a beginning of sub-clause concern:

- constructions with comparative adverbs followed by nominal phrases:
  “Tiene **[NP más** *poder NP]* **que** [NP el presidente NP].”

- constructions with non finite verbs:
  “El acuerdo tendrá **que [IV redactarse IV]** *de nuevo.”*

All these constraints have to be defined on the regular expressions to avoid wrong segmentation. Errors at this stage have important consequences on the following steps of the parsing analysis.

3.3 Marking syntactic functions

The temporary annotations being removed, the input text has at this stage all the syntactic annotations identifying chunks and sub-clause boundaries. NPs have been marked as well with nominal marks (/N) to better prepare the following stage of the analysis (these nominal marks will be replaced by subject or object marks if a certain number of conditions are satisfied).

To give an example, at this stage the result of the analysis for the sentence *“La posición del gobierno francés ha sido interpretada como una manera de eludir el problema.”* is:

```
[SC [NP La^el+DETSG posicion^posicion
+NOUNSG NP]/N [PP del^de=el+PREPDET
Gobierno^gobierno+NOUNSG PP] [AP
frances^frances+ADJSG AP] :v ha^haber
+HAB sido^ser+PAPAUX interpretada^
interpretar+PAPSG SC] como^como+COMO
[NP una^un+DETQUANTSG manera^manera
+NOUNSG NP]/N [IV de^de+PREP_DE eludir
^eludir+VERBINF IV]  [NP el^el+DETSG
problema^problema+NOUNSG NP]/N .^.+SENT
```

The next step is to use the information provided here to mark the main syntactic functions, subjects and objects.

3.3.1 Subjects

The general strategy consists on marking the NPs potential subjects and then removing those that appear to be incorrect. Several constraints are described by the regular expressions, to take into account the variety of possibilities concerning subjects in Spanish.

Here is an example of regular expression cleaning wrong subjects candidates:

```
[TSUBJ] -> [TNF] ||
    [ EndNP TNF BeginWord TCoord
    BeginNP ~$EndNP EndNP] _
```

This regular expression removes a potential subject if it is coordinated with a preceding NP not marked as subject (but with a nominal annotation).

Inverted subjects are next identified. Spanish language presents a characteristic that makes difficult the parsing at this stage: the lack of surface subjects is allowed. So the rules concerning inverted subjects have to be very restrictive in order to avoid overgeneration of errors (i.e. identifying NP objects as inverted subjects). Here are the expected analysis for the following sentences:

```
[SC :v Muere SC] [NP un joven NP]/<SUBJ
[SC :v Conecte SC] [NP el cable NP]/OBJ
```

From a syntactic point of view there is no difference between the two examples. One could say that the verbs have a different tense (present for *“Muere”* and imperative for *“Conecte”*, in this example). The tagger giving the tag +VERBFIN for both forms, it becomes difficult to make any difference between the two sequences.

According to these difficulties, the IFSP spanish parser identifies inverted subjects in the following cases:

- a verb followed by a subject pronoun (+PRONSUBJ) (*“... si tiene usted alguna sugerencia”.*)

- a verb preceeded by the preposition *según* and followed by an NP (*“... según explicó el secretario..”*)

- at the end of a sentence, a verb preceeded by a coma and followed by a NP (*“..., dijo un representante.”*)

This first analysis is refined by another transducer on the cascade which cleans wrong inverted subjects in the following cases:

- a NP coordinated with a PP (*“... ya que pordría poner en peligro a otros conductores o peatones.”*)

- a NP following another inverted subject without coordination (probably its apposition) (*“... explicó Tom Wocks máximo responsable.”*)

- after a verb cluster containing an auxiliary (*"... es un problema."*)

3.3.2 Objects

The same strategy is applied to identify objects: first marking under constraints and later removing false objects. Here is an example of a regular expression for identifying objects:

```
define MarkOBJ [

 [ [TNF] -> TOBJ ||
        [ Beginvfin ~$[PassiveVC]
 [EndSC|Endv] (~$[BeginSC|
   EndSC|Beginv|TSPunct]
 \[EndNP|TSPunct|TCOMO] )
         BeginNP [~$[EndNP|TPRONSUBJ] &
 $[TNum|TNoun|TDet|TPRONREL|
 TPRON|TPROP]] EndNP] _ ]
];
```

This rule transforms a nominal mark at the end of NPs by an OBJ mark if at the same level of embedding there is a finite verb (not in the passive form) followed by an NP not preceeded by punctuation or by *como*. This NP cannot contain a subject pronoun and can be of different types.

Other regular expressions remove wrong objects if they are NPs not coordinated with another NP already marked as OBJ.

3.4 Dependency extraction

The following is a sample of the input text at the end of the strictly shallow parsing analysis, for the sentence: *"El problema tiene una dimensión mayor y trasciende a lo que ocurre en el día de hoy."*

```
[SC [NP El^el+DETSG problema^problema
+NOUNSG NP]/SUBJ :v tiene^tener+VERBFIN
SC] [NP una^un+DETQUANTSG dimension^
dimension+NOUNSG NP]/OBJ [AP mayor^mayor
+ADJSG AP] y^y+COORD [SC :v trasciende^
trascender+VERBFIN SC] [PP a^a+PREP
lo^el+DETSG PP] [SC que^que+QUE :v
ocurre^ocurrir+VERBFIN SC] [PP en^
en+PREP el^el+DETSG dia^dia+NOUNSG PP]
de^de+PREP_DE hoy^hoy+ADV .^.+SENT
```

The results of the shallow parser after the primary segmentation and the marking of main syntactic functions permit to extract syntactic dependencies.

A dependency is a syntactic relation within the heads of the chunks of the sentence (in some cases prepositions can be part of the dependency). The dependencies are extracted within the following formats:

- DEPENDENCY(head1, head2)
- DEPENDENCY(head1, preposition, head2)

The word "dependency" stands on the examples for the name of the syntactic relation: DOBJ (direct object), ATTR (attribute of *ser*), ADVADJ (adverb modifying an adjective) etc.

Below are different examples of some of the dependencies the shallow parser extracts:

- SUBJ
  (Subject of an active verb)
  "Asegúrese de que haya buen contacto eléctrico y que los bornes de la batería estén firmes ."
  SUBJ(borne,estar)
- INVSUBJ
  (Inverted subject of an active verb)
  "La más multitudinaria de todas las marchas se desarrolló el sábado por el centro de Roma y a ella se sumaron unas 100.000 personas."
  INVSUBJ(sumar,persona)
- DOBJ
  (Direct object of a verb different from the auxiliary *ser*)
  "Una amplia mayoría acepta también la versión oficial ."
  OBJ(aceptar,versión)
- BEOBJ
  (Noun object of the auxiliary *ser*)
  "Según me dicen , es un gran avance.."
  BEOBJ(ser,avance)
- VMODOBJ
  (Prepositionnal modifier of an active verb)
  " Pero los estadounidenses siguen sin entender por qué están sus intereses nacionales en juego ."
  VMODOBJ(estar,en,juego)
- NNPREP
  (Noun modifying another noun with a preposition)
  ''Quise creer que lo que pasó en la ONU entre los Gobiernos británico e iraní era el final

de la historia.”
NNPREP(final,de,historia)
*NNPREP(ONU,entre,gobierno)

We are currently working on writing rules to extract around 30 types of dependencies describing noun/verb, verb/noun, verb/adverb, verb/adjective, noun/adjective, noun/noun, adverb/adjective and adverb/adverb relations.

4 Performances

The parser is implemented as a sequence of finite state networks. There are currently 25 networks which require about 1,5 MBytes of disk space. The speed of processing is about 115 words per second on a SPARCstation-10, including preprocessing.

A preliminary evaluation has been conducted on subject and object recognition using a technical manual text (4328 words, 292 sentences) and a newspaper text from *El País* (4517 words, 251 sentences). Precision and recall rates are given in tables 1 and 2.

Dependency	Precision	Recall
SUBJ	80,69 %	71,78 %
DOBJ	96,14 %	83,61 %

Table 1. Precision and recall rates for technical text.

Dependency	Precision	Recall
SUBJ	81,67 %	75,38 %
DOBJ	70,53 %	59,82 %

Table 2. Precision and recall rates for newspaper text.

Low results on recognizing subjects are mainly due to tagging errors, NPs containing time expressions and ambiguous coordinations (for instance, in *“Las pruebas de carretera o dinamómetro pueden ser necesarias”*, *“pruebas”* and *“dinamómetro”* are extracted as subjects).

On objects, the most important source of error corresponds to inverted subjects following verbs (*“periodista”* is the extracted object for *“Ya había llegado un periodista de Alemania.”*). Recall rate is quite low on newspaper articles because real objects containing “animate” nouns (in Spanish preceded by the preposition *“a”*) are not identified (as in *“Curiosamente, no se veía a ningún estudiante.”*).

5 Conclusions

The Spanish shallow parser presented in this article is a cascade of finite state transducers which apply over unrestricted spanish text. The parser follows the main lines of the incremental finite state parser developped by S. Aït-Mokhtar and J.P. Chanod.

Some specific linguistic features for Spanish have been taken into account in order to achieve broad coverage and accurate results. The lack of surface subjects in some cases or the variety of syntactical information carried by the word “que” are examples of complex specific phenomena the Spanish shallow parser deals with.

Future work will focus on:

- A wide evaluation of the linguistic accuracy of the Spanish parser in terms of precision and recall, as well as a detailed error analysis;
- An expansion of the coverage of the shallow parser techniques introducing lexical-semantic information (PhD work).

6 Parsing Samples

Below are some samples from the complete shallow parser analysis over unrestricted spanish text[4]:

“Las relaciones sociales son muy informales, en el sentido de que las personas se visitan sin previo aviso; ”

```
[SC [NP Las relaciones NP]/SUBJ [AP
sociales AP] :v son SC] [AP muy
informales AP], [PP en el sentido PP]
[SC [PP de que PP] [NP las personas
NP]/SUBJ :v se visitan SC] [PP sin
previo aviso PP] ;
```

SUBJ(relación,ser)
SUBJREFLEX(persona,visitar)
ATTR(relación,informal)
VMODOBJ(visitar,sin,aviso)
*VMODOBJ(ser,en,sentido)
PADJ(relación,social)
ADJ(previo,aviso)

[4] We have deliberately marked with “ * ” the wrong dependencies extracted due to overgeneration.

ADVADJ(muy,informal)

"Los componentes deben limpiarse cuidadosamente antes de la inspección previa a su montaje ."

```
[SC  [NP Los componentes NP]/SUBJ :v
deben SC] [IV limpiarse IV] cuidadosamente
[PP antes_de la inspeccion PP] [AP previa
 AP]  [PP a su montaje PP].
```

SUBJ(componente,deber)
SUBJREFLEX(componente,limpiar—se)
*NNPREP(inspección,a,montaje)
*VMODOBJ(limpiar—se,a,montaje)
VMODOBJ(limpiar—se,antes$_d$*e, inspección*)
PADJ(inspección,previo)
PADV(limpiar—se,cuidadosamente)

Acknowledgements

I would like to thank Salah Aït-Mokhtar and Gregory Grefenstette for the very valuable advices on finite-state technology and shallow parsing.

References

[Abney, 1991] Abney, S. (1991). Parsing by chunks. In Berwick, R., Abney, S., and Tenny, C., editors, *Principle-Based Parsing*. Academic Publishers.

[Aït-Mokhtar and Chanod, 1997] Aït-Mokhtar, S. and Chanod, J. P. (1997). Incremental Finite-State Parsing. In *Proceedings of ANLP-97*, Washington.

[Aït-Mokhtar and Chanod, 1997] Aït-Mokhtar, S. and Chanod, J. P. (1997). Subject and Object Dependency Extraction Using Finite-State Transducers. Technical report, Xerox Research Centre Europe (XRCE), Grenoble.

[Appelt et al., 1993] Appelt, D., J., H., J., B., Israel, D., and Tyson, M. (1993). 'FASTUS': a finite-state processor for information extraction from real-world text. In *Proceedings of IJCAI-93*, Chambry.

[Atserias et al., 1998] Atserias, J., Castellon, I., and Civit, M. (1998). Syntactic Parsing of Unrestricted Spanish Text. In *Proceedings of Fisrt International Conference on Languages Ressources and Evaluation*, Grenade, Spain.

[Chanod and Tapanainen, 1996] Chanod, J. P. and Tapanainen, P. (1996). A Nondeterministic Tokeniser for Finite-State Parsing. In *ECAI'96 workshop on Extended finite state models of language*, Budapest.

[Dini et al., 1999] Dini, L., Di Tomaso, V., and Segond, F. (1999). Ginger II, an example-driven word sense disambiguator. *Computers and Humanities, special issue.*

[Grefenstette, 1996]
Grefenstette, G. (1996). Light parsing as finite state filtering. In *Workshop on extended finite state models of language, ECAI'96*, Budapest, Hungary.

[Joshi, 1996] Joshi, A. (1996). A parser from antiquity: An early application of finite-state transducers to natural language parsing. In *Proceedings ECAI'96, workshop on extended finite state models of language*, Budapest.

[Karlsson et al., 1995] Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995). *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin and New York.

[Karttunen et al., 1997] Karttunen, L., Chanod, J. P., Grefenstette, G., and Schiller, A. (1997). Regular Expressions for Language Engineering. *Natural Language Engineering*, 2(4):305–328.