An Architecture for Document Routing in Spanish: Two Language Components, Pre-Processor and Parser

Guillermo Rojo, M^a. Concepción Álvarez, Pilar Alvariño, Adelaida Gil, María Paula Santalla, Susana Sotelo

Dept. of Spanish Language, University of Santiago de Compostela Burgo das Nacións, s/n. E-15771 Santiago de Compostela, Spain {fegrojo, femcal, fepili, iagilma, fempsr, fesdocio}@usc.es

Abstract

This paper describes the language components of a system for Document Routing in Spanish. The system identifies relevant terms for classification within involved documents by means of natural language processing techniques. These techniques are based on the isolation and normalization of syntactic unities considered relevant for the classification, especially noun phrases, but also other constituents built around verbs, adverbs, pronouns or adjectives. After a general introduction about the research project, the second Section relates our approach to the problem with other previous and current approaches, the third one describes corpora used for evaluating the system. The linguistic analysis architecture, including pre-processing and two different levels of syntactic analysis, is described in following fourth and fifth Sections, while the last one is dedicated to a comparative analysis of results obtained from the processing of corpora introduced in third Section. Certain future developments of the system are also included in this Section.

1. Introduction

The work described in this exposition partially reports on the DoRo¹ research project, which set the main requirements for its concrete development. The Doro project aims at improving traditional results of Information Retrieval applications, specifically Document Routing. Document Routing is the automatic routing of incoming text documents based on content analysis of text by a computer and knowledge of the characteristics of the candidate destinies. The innovative idea in the case of DoRo is to achieve this improvement of results by performing what may be considered, with respect to more traditional approaches, a deep linguistic analysis, non interpretative, of involved texts. This analysis should be based on linguistic resources (formal grammars and lexicons) developed in order to be able to (i) precisely identify keywords and phrases that characterize the input document and are relevant for the classification, (ii) normalize them to avoid sparsity of data caused by linguistic variation.

The role of the University of Santiago de Compostela as one of the members of the consortium constituted for the development of the project consisted of the production of the linguistic resources needed for Spanish, together with a first scientific testing of the whole system. The purpose of this article is the description of the linguistic components of the system as it works for Spanish, this is done by means of the description and analysis of the test performed at USC, which showed a strong dependency of results on the typology of texts involved.

After this introduction, the structure of the article is the following: the second Section compares various approaches to the problem, the third Section describes the two corpora used for testing, the following fourth, fifth and sixth Sections describe various phases of the entire system for the classification. The fourth Section is dedicated to the pre-processing phase, which constitutes the first phase of the processing of documents, preprocessing prepares documents for being processed by parsers described in the fifth Section. Parsing of the documents is described in the fifth Section, it constitutes the second phase of the process of classification, it is performed in two steps and delivers documents as collections of, possibly, nested terms. Classification itself, the assignment of documents to categories, is performed by the LCS^2 , Linguistic Classification System, an statistical module which unnests terms delivered by previous phase, assigns frequencies to these terms and compares them with previously stored class profiles of documents. Sixth Section shows results of various tests performed on the corpora described in third Section, it also includes an evaluation of these results.

2. Approaches to the problem

The problem faced by Information Retrieval and other more specific applications in the field, such as Document Routing, is, in first place, an statistical problem based on the estimate of the weigth of frequencies of *terms* found in documents to be retrieved. In this article we will not be concerned by the statistical side of the problem, but by the identification of *terms* used for the calculations: What is to be considered as a term and how terms are to be found.

¹ The DoRo project (1997-1999) was partly financed within the framework of the 6th Call of the Information Technologies Research and Technology Development (RTD) ESPRIT programme, managed by the European Commission Directorate General III (DGIII) and included in the Fourth Framework Programme (1994-1998) of European Union.

² See footnote 14.

From this point of view, conventional approaches to the problem rely on the identification of letter sequences —a term is a single word, or, with a slight refinement, a term is a pair of contiguous words (this is known as the SMART system, Buckley et al., 1996)—. These conventional approaches make also use of limited linguistic information. Certain heuristics are used to eliminate frequently occurring words that carry no important information for classification (*stop words*) and to group the morphological variants of a word (rough *stemming*).

Since then, several improvements have been proposed, which include more accurate stemming, incorporation of lexicon resources, heuristics on the weighting of proper names, *relevance feedback*, various purely statistical refinements, etc.

However, these improvements do not change the nature of the basic approach: it remains a keyword-based, enhanced with several useful heuristics which do not qualitatively affect the way terms are identified.

A qualitative change in the way the problem is faced takes place a bit later: the new idea is to replace *keywords* with more expressive units of texts, identified by means of the use of Natural Language Processing techniques.

A first approach of this type makes a limited use of such techniques. By means of various linguistics heuritics, it basically identifies noun phrases, including in certain cases, non continuous groups of heads and modifiers (Evans et al., 1996).

Finally, a second approach of this type is more ambitious with respect to the exploitation of NLP techniques. Not only noun phrases are considered, but also other predicate-argument binary (not necessarily continuous) structures are taken into account. Most of these approaches (Strzalkowski et al., 1997) rely on a specifically IR-oriented linguistic analysis, which may even previously eliminate stop words, performing after a rough "parsing" for the identification of certain patterns of linguistic relations. With respect to these approaches, ours syntactically analyses the complete sequences contained in documents, without eliminating parts of them. This makes the analysis more independent of its concrete application in Document Routing, leaving open, for instance, the possibility of integrating the grammar in a grammar for full-sentence analysis.

3. Description of Corpora

We take into account two text corpora:

Texts provided by $IECISA^3$, specific domain documents.

Texts taken from the Spanish newspaper El Mundo.

As for the texts in the IECISA corpus, they are classified into four text-groups: Contracts, Curricula, Customer Letters and Press News texts. They have specific characteristics regarding the four departments they belong to inside the company:

Contracts Department. Proposals and contracts between IECISA and other companies.

Customer Support. Customer's letters rating the customer service and the precise equipment the company works with in a positive or negative way.

Human resources. Curricula, covering letters and also application letters sent by people applying for a job.

Press Room. Press news about issues of interest for IECISA daily-life: markets, evolution of economy, other companies situation, general economical issues and so on.

The documents taken from *El Mundo* share the same characteristics that the ones in IECISA's Press Room Department, as far as the format is concerned. Classification of the texts in this corpus is as follows: culture, sports, economy, international news and national news.

Following tables show a comparative study between the two text corpora, regarding different criteria. Numbers in the following tables point out the fact that texts delivered by IECISA are much more homogeneous within various groups of interest that texts from *El Mundo*. The nature of texts, in this sense, is clearly different in both corpora: texts grouped under "Culture" in *El Mundo*, for instance, show a wide variety of issues: books and drama reviews, cinema, education, etc.

Sections	Tokens	Types	Different					
		• •	token every					
Documents from IECISA								
Press Room	56,292	8,342	6.75					
Customer	13,605	1,538	8.85					
Support								
Human	49,976	2,961	16.88					
Resources								
Contracts	107,610	8,346	12.76					
Total corpus	227,483	15,389	14.78					
Documents from EL MUNDO								
Culture	23,782	6,027	3.95					
Sports	21,243	5,246	4.05					
Economy	21,919	4,396	4.99					
International	22,604	5,447	4.15					
news								
National	39,845	7,067	5.64					
news								
Total corpus	129,393	17,808	7.27					

Table 1: Distribution of types and tokens

The IECISA corpus is larger than *El Mundo* -227.483 tokens as opposed to 129.393 ones. On the contrary, if we have a look on the total of types, we see that the number of types in IECISA is lower than in the newspaper *El Mundo* -17.808 types as opposed to 15.389. This apparent contradiction disappears if we take into account

³ IECISA stands for *Informática El Corte Inglés, Sociedad Anónima*, the company involved in the DoRo project as the spanish user for the final DoRo product.

	Types		Tokens	Av.				
	F	%	F	%				
IECISA	IECISA							
4 Depts.	458	2.98	130,713	57.46	285.4			
3 Depts.	890	5.78	33,711	14.82	37.9			
2 Depts.	2,644	17.18	29,769	13.08	11.3			
1 Dept.	11,3975	74.06	33,290	14.63	2.9			
Totals	15,3896	100.00	227,483	99.99				
EL MUNE	00							
5 Depts.	686	3.85	78,994	61.04	115.15			
4 Depts.	756	4.24	9,511	7.35	12.58			
3 Depts.	1,284	7.21	10,174	7.86	7.92			
2 Depts.	2,795	15.69	11,055	8.54	3.96			
1 Dept.	12,287	69.00	19,659	15.19	1.60			
Totals	17,808	99.99	129,393	99.98				

the main factor referred above: the clearly different nature of both corpora⁴.

Table 2: Frequency average and number of types according to their presence in one or more departments

The statistic results in this table are parallel to the ones found in the prior table. For a larger number of tokens, the number of types repeated in various departments, three or four, is lower in IECISA, which means that texts belonging to different departments are more different. On the other side, the number of types repeated in only two or, especially, one department —the ones which are not *hapax legomena*— is also lower in IECISA, which means that these types are more repeated, more characteristic, in the relevant departments in this case.

4. Pre-processor

Since AGFL⁷ parsers require each unity of analysis to be formatted on one line of input, we need to work with some form of automated pre-processing. For that reason we have built up a generator of text pre-processors: GenPrep. Each of the pre-processors generated from GenPrep is composed of a set of automata performing the main following two tasks: sentence segmentation and expansion of abbreviations.

1	Cualqui	era	de	las	par	tes	podrá	reso	lver	el	
2	mismo,	com	unic	ando	po po	r es	scrito	a la	otr	a	
3	parte,	su	inte	encić	ón e	n ta	l sent	ido	con	una	
4	antelac	ión	mír	ıima	de	tres	meses	s. En	ı el	caso	

⁴ The general fact that the ratio between tokens and types decreases as the corpus size increases cannot explain by itself the fact that the number of types in the IECISA corpus is lower not only in proportional terms, as we would have expected, but also in absolute terms.

⁶ 9,400 are *hapax legomena*.

5 de que la resolución del contrato se 6 efectuase antes de la finalización del período 7 contratado y a instancias de I.E.C.I.S.A., 8 esta sociedad vendrá obligada a devolver 9 a GRUPO ASNKF EQUIFAX el importe proporcional 10 al período no consumido. Si la cancelación 11 del contrato fuera a instancias de GRUPO ASNEF 12 EQUIFAX, I.E.C.I.S.A., no vendrá obligado a 13 devolver cantidad alguna. 14 Cualquier cambio en las condiciones de este 15 contrato exige la firma de uno nuevo que 16 anule al presente.

Figure 1. Example⁸ from co498_505.txt source document.

Sentence Segmentation. The main functionality of the pre-processor is to segment running text into unities of analysis for the AGFL parser. GenPrep allows the user to determine, by means of regular expressions, where the preprocessor generated is going to segment input, that is, in our case, what we want to consider as the end of a sentence. The same holds for decapitalization of capital letters at the beginning of the sentence. The pre-processor generated for the current application considers end of sentence the usual ones: period, colon, semi-colon, exclamation and interrogation marks. Initial capitals are decapitalized.

Abbreviation file. GenPrep allows also the specification of a list of abbreviations and their corresponding expanded forms. Prior and final context may be specified in order to avoid possible ambiguities and wrong expansions. A large and especially dedicated list of abbreviations was developed for the application described.

When a contextual disambiguation was not possible (i.e., regarding prior and final context), a statistical study was carried out. The more frequent possibility of expansion was included in the abbreviation file taking into account the domain of the application.

 cualquiera de las partes podrá resolver el mismo, comunicando por escrito a la otra parte, su intención en tal sentido con una antelación mínima de tres meses.
 en el caso de que la resolución del contrato se efectuase antes de la finalización del período contratado y a instancias de Informática

⁸ The Example 1 can be approximately translated as follows:

"Any of the parts will be able to resolve the contract itself, by telling in writing to the other part involved her intention to this effect with a minimum of three month beforehand. In case that the contract resolution take place before the negotiated period finishes and at the request of I.E.C.I.S.A., this society will be obliged to refund the GRUPO ASNFK EQUIFAX the value in proportion to the left period. If the cancellation of the contract were at the request of GRUPO ASNEF EQUIFAX, I.E.C.I.S.A. would not be obliged to refund any kind of value. Any change in the contract conditions demands to sign a new contract that cancels the current one".

⁵ 6,654 are *hapax legomena*, so useless for classification.

⁷ See footnote 9.

```
El Corte Inglés, Sociedad Anónima, esta sociedad
vendrá obligada a devolver a GRUPO ASNKF EQUIFAX
el importe proporcional al período no consumido.
3 si la cancelación del contrato fuera a
instancias de GRUPO ASNEF EQUIFAX, Informática
El Corte Inglés, Sociedad Anónima, no vendrá
obligado a devolver cantidad alguna.
4 cualquier cambio en las condiciones de este
contrato exige la firma de uno nuevo que anule
al presente.
```

Figure 2. Result of pre-processing for Example 1

For the evaluation, it was necessary to carry out two types of pre-processing of texts: fully automatic preprocessing (the one described before), and semi-automatic pre-processing. Semi-automatic pre-processing manually improves the results of fully automatic pre-processing, removing certain mistakes from the texts (problems mainly caused by headings, dates, addresses, charts, tables or signatures) and, therefore, minimizing the stir that these errors could cause in the final analysis.

5. Identification of Terms

Using the AGFL⁹ system, this task is performed in two steps: two different grammars written in the AGFL formalism are used, the output of the parser generated from first one (*AVALON_PhL.gra*) is the input for the parser generated from the second one (*terms.gra*). The work done by each one of them, as well as their characteristics, are described along this Section.

5.1. First step of parsing: AVALON_PhL.gra

5.1.1. Robust Phrase Identification

*AVALON_PhL.gra*¹⁰ (which stands for *AVALON Phrase Level*) identifies phrases contained in linguistic sequences delivered by the pre-processing phase as unities of analysis, sentences, for the parser.

AVALON_PhL.gra gives account for syntax covering the contextualization of verbs, nouns, pronouns, adverbs and adjectives in linguistic sequences, that is, AVALON_PhL.gra describes linguistic mechanisms used to extend from lexical categories to phrases playing syntactic functions at the level of the clause.

• The constitution of Verbal Phrases covers morphological addition of auxiliaries, semiauxiliaries and clitic pronouns to main verbs (that is, the grammar gives account for the constitution of, for instance, *ha habido que decírselo* from lexical item *decir*, to obtain the syntagm playing the syntactic function PREDICATE in *A ella ha habido que decírselo primero*, "It was necessary to to tell it to her before").

• The constitution of Noun, Pronoun, Adverb and Adjective Phrases covers syntactic addition of determiners and modifiers to nouns, pronouns, adverbs and adjectives (that is, the grammar gives account for the constitution of, for instance, *la espectacular explosión del camión de gasolina*, through the recursive expansion from lexical item *explosión* by means of the addition of two modifiers, *espectacular* and *del camión de gasolina*, and one determiner, *la*, constituting the syntactic function SUBJECT in *La espectacular explosión del camión de gasolina fue retransmitida en directo por todas las cadenas de televisión*, "The spectacular explosion of the petrol truck was broadcasted live by all TV channels")

• A syntactic phrase is defined as each one of the binary unities resulting from the single application of one mechanism of expansion onto a lexical category or phrase (that is, *este conjunto* in *todo este conjunto de colores* is a phrase), but phrases isolated as terms for the classification are only the ones playing a syntactic function within the clause (that is, *todo este conjunto de colores*, which is the SUBJECT in *Todo este conjunto de colores* aturde al espectador, "All this collection of colours stuns the viewer", is a phrase and will be isolated as a term). This means that, for our grammar, linguistically motivated recursive identification of phrases is relevant for the identification of terms.

• On the other side, the grammar identifies contiguous phrases contained in sequences of analysis delivered by the pre-processing phase. These phrases are all placed at the same level without being assigned syntactic functions (that is, for *La espectacular explosión del camión de gasolina fue retransmitida en directo por todas las cadenas de televisión* the grammar delivers a sentence constituted by *la espectacular explosión del camión de gasolina, fue retransmitida, en directo* and *por todas las cadenas de televisión*).

• Such a grammar is very robust by itself. In spite of this, the grammar is also reinforced by a level of word analysis (a kind of simple tagging) for the case that an analysis into phrases is not found for a sequence.

• Finally, the grammar is optimised, that is, ambiguities that cannot be solved on the base of phrase level analysis have been removed (they are underspecified or the more frequent one has been selected as the only possible one).

Example 1: the analysis offered by *AVALON_PhL* for the first sequence in Example 1 (see previous Section), is

⁹ The AGFL system, which is a collection of software systems for NLP, mainly for parser generation, has been developed, and is still developed and maintained, by professor C.H.A. Koster and his collaborators in the Department of Software Engineering of the University of Nijmegen, which was, in fact, one of the partners involved in the DoRo project (URL: http://www.kun.nl/agfl).

¹⁰ The correspondent parser has been generated with 1.7.52 version of the AGFL parser generator.

partially¹¹ showed here. This output format is only internally handled by the first step of parsing. As Section 5.1.2 will explain, within the routing system, the *AVALON_PhL* parser is executed to produce an alternative output. The one showed here, nevertheless, is the more illustrative one with respect to the grammar itself. Furthermore, this is very helpful for the understanding of the operations carried out by transduction (see Section 5.1.2):

```
UTTERANCE
  Pronoun Phrase(...)
    Pronoun Phrase NUCLEUS(...)
      Pronoun(...)
         "cualquiera"
    Prepositional Phrase MODIFIER(...)
      Prepositional Noun Phrase(...)
         Prepositional LINKER(...)
           Preposition(...)
             "de"
         LINKED Noun Phrase(...)
           Noun Phrase(...)
             DETERMINER (....)
               Determiner(...)
                  "las"
             Noun Phrase NOMINAL I( ... )
               Noun(...)
                  "partes"
  Periphrastic Verbal Phrase(...)
    Periphrasis(...)
      Verb(...)
         "podrá"
    Non Personal Verbal Phrase(...)
      Verb(...)
         "resolver"
  Nominalization(...)
    DETERMINER(...)
      Determiner(...)
         "el"
    Nominalization NOMINAL I( ... )
      Determiner(...)
         "mismo
  Punctuation Mark
    Clause(...)
    PREDICATE(...)
      Non Personal Verbal Phrase(...)
         Verb(...)
           "comunicando"
  Adverb(...)
  | "por escrito"
  Prepositional Noun Phrase(...)
    Prepositional LINKER(...)
      Preposition(...)
         "a"
    LINKED Noun Phrase(...)
      Noun Phrase(...)
        DETERMINER (....)
          Determiner(...)
              "la'
         Noun Phrase NOMINAL II(...)
           Noun Phrase I(...)
```

```
<sup>11</sup> Attributes values should appear between parentheses, they are not showed here to give a more direct sight on the analysis. Vertical lines between constituents at the same level are not part of the original output, they are introduced here for the same reason.
```

```
DETERMINER ( ... )
             Determiner(...)
                "otra"
           Noun Phrase NOMINAL I(...)
             Noun(...)
                "parte"
Punctuation Mark
| ", "
Noun Phrase(...)
  DETERMINER(...)
    Determiner(...)
       "su"
  Noun Phrase NOMINAL I(...)
    Noun(...)
       "intención"
Prepositional Noun Phrase(...)
  Prepositional LINKER(...)
    Preposition(...)
       "en"
  LINKED Noun Phrase(...)
    Noun Phrase(...)
      DETERMINER ( ... )
        Determiner(...)
           "tal"
      Noun Phrase NOMINAL I(...)
         Noun(...)
           "sentido"
Prepositional Noun Phrase(...)
  Prepositional LINKER(...)
    Preposition(...)
       "con"
  LINKED Noun Phrase(...)
    Noun Phrase(...)
      DETERMINER(...)
        Determiner(...)
           "una"
      Noun Phrase NOMINAL I(...)
         Substantive Phrase(...)
           Substantive Phrase NUCLEUS( ... )
             Substantive Phrase(...)
                Substantive Phrase NUCLEUS( ... )
                  Noun(...)
                     "antelación"
                Adjective Phrase MODIFIER( ... )
                  Adjective(...)
                     "mínima"
           Prepositional Phrase MODIFIER(...)
             Prepositional Noun Phrase( ... )
                Prepositional LINKER(...)
                  Preposition(...)
                     "de"
                LINKED Noun Phrase(...)
                  Noun Phrase(...)
                    DETERMINER ( ... )
                       Determiner(...)
                         "tres"
                    Noun Phrase NOMINAL I(...)
                       Noun(...)
                         "meses"
Punctuation Mark
  . .
```

Figure 3: Analysis of Example 1, *AVALON_PhL* parser, transduction option: off

5.1.2. Transduction

The AGFL formalism allows the specification of a transduction for each rule contained in the grammar. By means of transductions, default output of the parser (the

one showed by the example in previous Section 5.1.1, Figure 3) may be modified: elements may be removed, added or reorganized. Executing the parser with the adequate option the output obtained is the one defined by transductions specified in the rules.

All AVALON_PhL.gra rules are provided with transduction, this satisfies the first stage in the so-called process of syntactic normalization, which tries to overcome surface or trivial differences —from the point of view of classification— between phrases, obtaining the same terms from them. Within the Document Routing system described here, the AVALON_PhL parser is obviously executed to produce the output specified by transductions in the rules. Transductions in AVALON_PhL.gra are responsible for the following facts:

• Each identified phrase is delivered as a unity of analysis for the second phase of analysis (that is, each identified phrase is placed on a different line).

• Phrases are tagged according to their types. SGML tagging is used to simplify structures obtained in the analysis.

• Irrelevant elements from the point of view of classification are removed from the output (auxiliaries, semiauxiliaries, clitic pronouns, determiners, all pronouns except personal pronouns, prepositions, conjunctions, modifying adverbs).

• Relevant elements are reorganized ensuring that they will be always delivered to the second phase of analysis in the same order: a head possibly followed by one modifier.

Example 1: The output for first sequence in Example 1 produced by the *AVALON_PhL* parser (first phase of syntactic analysis of the parsing module included in the Document Routing environment) is the following¹²:

```
<PronP> , <N> partes </N> </PronP>
1
2 <MAIN VERB> resolver </MAIN VERB>
3
4
5 <MAIN VERB> comunicando </MAIN VERB>
 <Av> por escrito </Av>
6
7
  <N> parte </N>
8
9 <N> intención </N>
10 <N> sentido </N>
11 <NP> <NP> <N> antelación </N> , <AJ> mínima
</AJ> </NP> , <N> meses </N> </NP>
12
```

Figure 4: Analysis of Example 1, *AVALON_PhL* parser, transduction option: on (1st step of the parsing module)

5.2. Second step of parsing: terms.gra

Terms.gra is also written in the AGFL formalism, it satisfies the second stage in the process of syntactic normalization. Rules in *terms.gra* analyse the output generated by *AVALON_PhL* parser. All rules in *terms.gra* are also provided with transduction and they produce the final output of the parsing module of the whole system for Document Routing. From the point of view of *terms.gra*, the syntactic trees produced by *AVALON_PhL.gra* are delivered as binary frames delimited by SGML tags and composed of a Head and a Modifier, or unary frames that consist only of a Head. *Terms.gra* and generate the input format required by the unnester integrated in the classifier (the LCS, see Section 6). This means the following:

• SGML tags are removed and structures are enclosed between square brackets that contain either a single Head or a Head and a Modifier separated by a comma. Both the Head and the Modifier can be, in turn, binary structures composed by a Head and a Modifier.

• Square brackets are added only in certain cases. Various rules determine, according to the type of phrase involved, whether the Head is going to be bracketed by itself or not. This is important for the unnester (integrated in the LCS, see Section 6) to identify these heads also as single terms by themselves and assign frequencies.

• Empty modifiers delivered by the previous phase of analysis are removed in this phase.

Example 1: The output for first sequence in Example 1 produced by the *terms* parser (second, and final, phase of syntactic analysis of the parsing module included in the Document Routing environment) is the following¹³:

```
1 [ ,[partes ]]
2 [resolver ]
5 [comunicando ]
6 [porescrito ]
7 [parte ]
9 [intención ]
10 [sentido ]
11 [[[antelación ], mínima], [meses ]]
```

Figure 5: Analysis of Example 1, terms parser, transduction option: on (2nd, and final, step of parsing module)

6. Classification results, output from LCS

The *Linguistic Classification System*¹⁴ is the software module which finally decides into which category a

¹⁴ The LCS, Linguistic Classification System, has been

¹² Line numbers are not part of the AGFL output, they have been included to clarify it. Each line contains the transduction of each constituent of the first sequence of Example 1, according to analysis in Figure 3. Some of them are just unconsidered for further processing (*el mismo*, a Pronoun Phrase without modification, and punctuation marks).

¹³ Remind that each phrase result was placed by transduction included in the AVALON_PhL parser on a different line, so it constitutes a different unity of analysis for the *terms* parser. To facilitate the understanding of the output, we maintained here the line numbers, which are not part of the real AGFL output, identifying each phrase result in Figure 4.

document is to be classified. After processing by the parsing module, the document has been *indexed*, that is, reduced to a collection of —syntactically identified—relevant terms, possibly nested. The first task of the LCS is the unnesting of nested terms and the assignment of frequencies to resulting unnested terms.

Example 1: Terms and frequecies extracted from first sequence in Example 1 are the following:

1	["partes"]
1	["resolver"]
1	["comunicando"]
1	["por escrito"]
1	["parte"]
1	["intención"]
1	["sentido"]
1	["antelación"]
1	["antelación", "mínima"]
1	["antelación", "meses"]
1	["meses"]

The whole collection of terms of this type extracted from an input document is then compared, using some *similarity measure*, with class profiles, previously stored collections of terms associated with various possible categories of classification, and the document is classified.

Class profiles are constructed automatically by a *learning system*, which is trained on sample documents that have been categorized by human indexers. The system also provides a *testing system* for the selection and tuning of learning algorithms (see Derksen, 1998 for details). We show here results of three test experiments performed on the corpora described in Section 3. Apart from corpora, we handle two variables also introduced in Section 4: fully automatic pre-processing and semi-automatic pre-processing. We use in all cases 80% of the documents for training and 20% of the documents for testing, the total number of documents is 836 for the IECISA corpus, domain texts, and 218 for the *El Mundo* corpus.

Precision/Recall/F:							
cl	90.32%	87.	50%	0.89			
CO	88.24%	88.2	24%	0.88			
CV	98.81%	100.0	00%	0.99			
dp	88.89%	88.8	89%	0.89			
Corre	ct/Errone	ous/Er	ror-r	ate:			
cl	28	3	9.	68%			
CO	15	2	11.	76%			
CV	83	1	1.	19%			
dp	32	4	11.	11%			
Total	158	10	5.	95%			

Figure 6¹⁵: Domain texts, fully automatic pre-processing

Precision/Recall/F:								
cl	90.32%	1(30.00%	0.95				
CO	100.00%	8	35.00%	0.92				
cv	98.81%	9	97.65%	0.98				
dp	91.67%	9	94.29%	0.93				
_								
Corre	ect/Erronec	ous/l	Error-ra	ate:				
cl	28	3	9.688	Ś				
CO	17	0	0.00	Ś				
cv	83	1	1.198	5				
dp	33	3	8.338	5				
Total	. 161	7	4.178	Ś				

Figure 7: Domain texts, semi-automatically pre-processed

Precision/Recall/F:								
cul	62.50%	71	.43%	0.67				
dep	22.22%	100	.00%	0.36				
eco	30.77%	50	.00%	0.38				
int	57.14%	40	.00%	0.47				
nac	57.14%	23	.53%	0.33				
Correct	t/Errone	ous/E	rror-	rate:				
cul	5	3	37.	50%				
dep	2	7	77.	78%				
eco	4	9	69.	23%				
int	4	3	42.	86%				
nac	4	3	42.	86%				
Total	19	25	56.	82%				

Figure 8¹⁶: Newspaper texts, semi-automatically preprocessed

First of all, we want to point out that the differences found in results must be, to a certain extent, attributed to the tuning of resources, which were developed to give account especially for the domain of the application. Even if the grammar and the lexicon were conceived as general resources, certain characteristics are more conditioned by the texts found in the domain. This is especially remarkable for proper names, which have been collected and included in the lexicon only in the case of the domain texts. On the other side, the corpus of the domain is bigger, so training is better.

Nevertheless, we think that the different results cannot be explained just on the base of the tuning of resources. It is, in fact, the nature of texts that we think can explain such a difference. The ratio types/tokens, in Table 1, shows a much higher lexical variation for newspaper texts. This is a definitive factor and numbers in Table 2 have to be evaluated taking into account that they have more or less weight as terms depending on the total number of tokens (that is, types in one department for domain texts are not only more, but they are also much more relevant than types in one department for newspaper texts, because the number of tokens in the first example almost doubles the second). The percentage of proper

developed, and is still developed and maintained, by professor C.H.A. Koster and his collaborators in the Department of Software Engineering of the University of Nijmegen.

¹⁵ In Figures 6 and 7 keys for the identification of categories are the following : *cl*, Customer Support, *co*, Contracts Department, *cv*, Human Resources, *dp*, Press Room.

¹⁶ In Figure 8 keys for the identification of categories are the following : *cul*, Culture, *dep*, Sports, *eco*, Economy, *int*, International news, *nac*, National news.

name types, on the other side, in domain texts is higher (26.9%) than in newspaper texts (only 21.78%). Even though the difference between frequencies is not excessive, we must take into account two features: on the one hand, the lack of balance between number of types and tokens in corpora (as we have described before), and, on the other hand, the way in that we have extracted information about proper names from corpora, manually for the domain texts but automatically for the newspaper corpus (by matching regular expressions that reject ambiguities such as upper case after period).

The improvement of results for newspaper texts, to meet results of domain texts, may be achieved by means of the following strategies, which constitute the obvious future developments of the proposed system:

-General resources: obviously, grammars and lexicons should be extended to cope with not only newspaper texts, but various types of them, with the same efficiency they can ensure for domain texts. Together with them, additional specific domain resources must be developed for concrete applications, ensuring that certain particularities of concrete texts will be correctly interpreted by the system.

-The pre-processing phase may be considerably improved by the implementation of mechanisms for the identification of headings, addresses, signatures and other fixed expressions, as well as of tables, charts and other figures not strictly linguistically motivated. More interaction between segmentation and tokenization, on the one side, and the lexicon, on the other one, would be also desirable.

-Last but not least, integration of morphological and semantic normalization, that is, lemmatization and reduction of synonyms, hypernyms and hyponyms to a conventional representative should cause a strong improvement of results.

For the time being, we expect to have showed an operative methodology for syntactic normalization in the wide area of Information Retrieval, whose results, in absolute terms, may be considerably improved by normalization at morphological and semantic levels. We wanted also to show that certain strategies for Information Retrieval which may constitute a solution for certain kind of texts, need to be enriched to become effective in other contexts, which may considerably influence results.

7. References

- Álvarez, C., P. Alvariño, A. Gil, T. Romero, M^a.P. Santalla, and S. Sotelo, 1998. AVALON, una gramática formal basada en corpus. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 23: 132-139.
- Arampatzis, A.T., T. Tsoris, C.H.A. Koster, and Th.P. van der Weide, 1998. *Phrase-based Information Retrieval*. Technical Report CSI-R9809, University of Nijmegen.
- Arampatzis, A.T., Th.P. van der Weide, C.H.A. Koster, and P. van Bommel, 1999. *Text Filtering using Linguistically-motivated Indexing Terms*. Technical

Report CSI-R9901, University of Nijmegen.

- C. Buckley, A. Singhal, and M. Mitra, 1996. New retrieval approaches using smart: Trec4. In D.K. Harman (ed.), *The Fourth Text Retrieval Conference* (*TREC-4*), U.S. Department of Commerce.
- Derksen, C., 1995. *Manual for the GEN Parser Generator Version* 1.5, Computing Science Department, University of Nijmegen.
- Derksen, C., 1998. *The Profiling System.User Manual*, Computing Science Department, University of Nijmegen.
- D. A. Evans, and C. Zhai, 1996. Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. *Proceedings of the 34th Annual Meeting of Association for Computational Linguistics, Santa Cruz, California, June 24-28.*
- Koster, C.H.A., 1991. Affix grammars for natural languages. In H. Albas and B. Melichar (eds)., Attribute Grammars, Applications and Systems, International Summer School SAGA, Prague, June, 1991, Lecture Notes in Computer Science 545, Heidelberg: Springer-Verlag.
- Koster, C.H.A., and C.P.A. Tiberius, 1997. Affix grammars for full-text information retrieval. In A. Ralli et al. (eds.), *Working Papers in Natural Language Processing*, Athens: Ekdosis Diavlos.
- T. Strzalkowski, F. Lin, J. Pérez-Carballo, and J. Wang, 1997. Natural Language Information Retrieval: TREC-6 Report. *Proceedings of the TREC-6 Conference*.