The Spanish Resource Grammar: pre-processing strategy and lexical acquisition

Montserrat Marimon, Núria Bel, Sergio Espeja, Natalia Seghezzi
IULA - Universitat Pompeu Fabra
Pl. de la Mercè, 10-12
08002-Barcelona
{montserrat.marimon,nuria.bel,sergio.espeja,natalia.seghezzi}@upf.edu

Abstract

This paper describes work on the development of an open-source HPSG grammar for Spanish implemented within the LKB system. Following a brief description of the main features of the grammar, we present our approach for pre-processing and ongoing research on automatic lexical acquisition.[1]

1 Introduction

In this paper we describe the development of the Spanish Resource Grammar (SRG), an open-source[2] medium-coverage grammar for Spanish. The grammar is grounded in the theoretical framework of HPSG (*Head-driven Phrase Structure Grammar*; Pollard and Sag, 1994) and uses *Minimal Recursion Semantics* (MRS) for the semantic representation (Copestake et al, 2006). The SRG is implemented within the *Linguistic Knowledge Building* (LKB) system (Copestake, 2002), based on the basic components of the grammar Matrix, an open–source starter-kit for the development of HPSG grammars developed as part of the LinGO consortium's multilingual grammar engineering (Bender et al., 2002).

The SRG is part of the DELPH-IN open-source repository of linguistic resources and tools for writing (the LKB system), testing (The [incr tsbd()]; Oepen and Carroll, 2000) and efficiently processing HPSG grammars (the PET system; Callmeier, 2000). Further linguistic resources that are available in the DELPH-IN repository include broad-coverage grammars for English, German and Japanese as well as smaller grammars for French, Korean, Modern Greek, Norwegian and Portuguese .[3]

The SRG has a full coverage of closed word classes and it contains about 50,000 lexical entries for open classes (roughtly: 6,600 verbs, 28,000 nouns, 11,200 adjectives and 4,000 adverbs). These lexical entries are organized into a type hierachy of about 400 leaf types (defined by a type hierarchy of around 5,500 types). The grammar also has 40 lexical rules to perform valence changing operations on lexical items and 84 structure rules to combine words and phrases into larger constituents and to compositionally build up the semantic representation.

We have been developing the SRG since January 2005. The range of linguistic phenomena that the grammar handles includes almost all types of subcategorization structures, valence alternations, subordinate clauses, raising and control, determination, null-subjects and impersonal constructions, compound tenses, modification, passive constructions, comparatives and superlatives, cliticization, relative and interrogative clauses and sentential adjuncts, among others.

Together with the linguistic resources (grammar and lexicon) we provide a set of controlled hand-constructed test suites. The construction of the test suites plays a major role in the development of the SRG, since test suites provide a fine-grained diag-

[1] This research was supported by the Spanish *Ministerio de Educación y Ciencia*: Project AAILE (HUM2004-05111-C02-01), *Ramon y Cajal*, *Juan de la Cierva* programmes and PTA-CTE/1370/2003 with *Fondo Social Europeo*.

[2] The Spanish Resource Grammar may be downloaded from: http://www.upf.edu/pdi/iula/montserrat.marimon/.

[3] See http://www.delph-in.net/.

Proceedings of the 5th Workshop on Important Unresolved Matters, pages 105–111,
Ann Arbor, June 2005.

nosis of grammar performance and they allow us to compare the SRG with other DELPH-IN grammars. In building the test suites we aimed at (a) testing specific phenomena in isolation or in controlled interaction, (b) providing test cases which show systematic and exhaustive variations over each phenomenon, including infrequent phenomena and variations, (c) avoiding irrelevant variation (i.e. different instances of the same lexical type), (d) avoiding ambiguity, and (e) including negative or ungrammatical data. We have about 500 test cases which are distributed by linguistic phenomena (we have 17 files). Each test case includes a short linguistic annotation describing the phenomenon and the number of expected results when more than one analysis cannot be avoided (e.g. testing optionality).

Test suites are not the only source of data we have used for testing the SRG. Hand-constructed sentences were complemented by real corpus cases from: (a) the Spanish questions from the Question Answering track at CLEF (CLEF-2003, CLEF-2004, CLEF-2005 and CLEF-2006), and (b) the *general* sub-corpus of the *Corpus Tècnic de l'IULA* (IULA's Technical Corpus; Cabré and Bach, 2004); this sub-corpus includes newspaper articles and it has been set up for contrastive studies. CLEF cases include short queries showing little interaction of phenomena and an average of 9.2 words; newspaper articles show a high level of syntactic complexity and interaction of phenomena, sentences are a bit longer, ranging up to 35 words. We are currently shifting to much more varied corpus data of the *Corpus Tècnic de l'IULA,* which includes specialized corpus of written text in the areas of computer science, environment, law, medicine and economics, collected from several sources, such as legal texts, textbooks, research reports, user manuals, … In these texts sentence length may range up to 70 words.

The rest of the paper describes the pre-processing strategy we have adopted and on our on-going research on lexical acquisition.

2 Pre-processing in the SRG

Following previous experiments within the *Advanced Linguistic Engineering Platform* (ALEP) platform (Marimon, 2002), we have integrated a shallow processing tool, the FreeLing tool, as a pre-processing module of the grammar.

The FreeLing tool is an open-source[4] language analysis tool suite (Atserias et al., 2006) perfoming the following functionalities (though disambiguation, named entity classification and the last three functionalities have not been integrated):

- Text tokenization (including MWU and contraction splitting).
- Sentence splitting.
- Morpho-syntactic analysis and disambiguation.
- Named entity detection and classification.
- Date/number/currency/ratios/physical magnitude (speed, weight, temperature, density, etc.) recognition.
- Chart-based shallow parsing.
- WordNet-based sense annotation.
- Dependency parsing.

FreeLing also includes a guesser to deal with words which are not found in the lexicon by computing the probability of each possible PoS tag given the longest observed termination string for that word. Smoothing using probabilities of shorter termination strings is also performed. Details can be found in Brants (2000) and Samuelson (1993).

Our system integrates the FreeLing tool by means of the LKB *Simple PreProcessor Protocol* (SPPP; http://wiki.delph-in.net/moin/LkbSppp), which assumes that a preprocessor runs as an external process to the LKB system, and uses the LKB inflectional rule component to convert the PoS tags delivered by the FreeLing tool into partial descriptions of feature structures.

2.1 The integration of PoS tags

The integration of the morpho-syntactic analysis in the LKB system using the SPPP protocol means defining inflectional rules that propagate the morpho-syntactic information associated to full-forms, in the form of PoS tags, to the morpho-syntactic features of the lexical items. (1) shows the rule propagating the tag AQMS (adjective qualitative masculine singular) delivered by FreeLing. Note

[4] The FreeLing tool may be downloaded from http://www.garraf.epsevg.upc.es/freeling/.

that we use the tag as the rule identifier (i.e. the name of the inflectional rule in the LKB).

```
(1) aqms :=
    %suffix ()
    [SYNSEM.LOCAL[CAT adj,
                  AGR.PNG[PN 3sg,
                          GEN masc]]]
```

In Spanish, when the verb is in non-finite form, such as infinitive or gerund, or it is in the imperative, clitics[5] take the form of enclitics. That is, they are attached to the verb forming a unique word, e.g. *hacerlo* (*hacer+lo*; to do it), *gustarle* (*gustar+le*; to like to him). FreeLing does not split verbs and pronouns, but uses complex tags that append the tags of each word. Thus, the form *hacerlo* gets the tag *VMN+PP3MSA* (verb main infinitive + personal pronoun 3rd masculine singular accusative). In order to deal with these complex tags, the SRG includes a series of rules that build up the same type of linguistic structure as that one built up with the structure rules attaching affixes to the left of verbal heads. Since the application of these rules is based on the tag delivered by FreeLing, they are included in the set of inflectional rules and they are applied after the set of rules dealing with complement cliticization.

Apart from avoiding the implementation of inflectional rules for such a highly inflected language, the integration of the morpho-syntactic analysis tags will allow us to implement default lexical entries (i.e. lexical entry templates that are activated when the system cannot find a particular lexical entry to apply) on the basis of the category encoded to the lexical tag delivered by FreeLing, for virtually unlimited lexical coverage. [6]

2.2 The integration of multiword expressions

All multiword expressions in FreeLing are stored in a file. The format of the file is one multiword per line, having three fields each: form, lemma and PoS.[7] (2) shows two examples of multiword fixed expressions; i.e. the ones that are fully lexicalized and never show morpho-syntactic variation, *a través de* (through) and *a buenas horas* (finally).

```
(2) a_través_de a_través_de SPS00
   a_buenas_horas a_buenas_horas RG
```

The multiword form field may admit lemmas in angle brackets, meaning that any form with that lemma will be a valid component for the multiword. Tags are specified directly or as a reference to the tag of some of the multiword components. (3) builds a multiword with both singular and plural forms (*apartado(s) de correos* (P.O Box)). The tag of the multiform is that of its first form ($1) which starts with NC and takes the values for number depending on whether the form is singular or plural.

```
(3) <apartado>_de_correos apar-
tado_de _correos \$1:NC
```

Both fixed expressions and semi-fixed expressions are integrated by means of the inflectional rules that we have described in the previous subsection and they are treated in the grammar as word complex with a single part of speech.

2.3 The integration of messy details and named entities

FreeLing identifies, classifies and, when appropriate, normalizes special text constructions that may be considered peripheral to the lexicon, such as dates, numbers, currencies, ratios, physical magnitudes, etc. FreeLing also identifies and classifies named entities (i.e. proper names); however, we do not activate the classification functionality, since high performance of that functionality is only achieved with PoS disambiguated contexts.

To integrate these messy details and named entities into the grammar, we require special inflectional rules and lexical entry templates for each text construction tag delivered by FreeLing. Some of these tags are: W for dates, Z for numbers, Zm for currencies, ... In order to define one single entry for each text construct, we identify the tag and the STEM feature. (4) shows the lexical entry for dates.[8]

[5] Actually, Spanish weak pronouns are considered *pronominal affixes* rather than *pronominal clitics*.

[6] The use of underspecified default lexical entries in a highly lexicalized grammar, however, may increase ambiguity and overgeneration (Marimon and Bel, 2004).

[7] FreeLing only handles continuous multiword expressions.

[8] Each lexical entry in the SRG consists of a unique identifier, a lexical type, an orthography and a semantic relation.

```
(4)
date := date_le &
[STEM <"w">,
SYNSEM.LKEY.KEYREL.PRED time_n_rel]
```

The integration of these messy details allows us to release the analysis process from certain tasks that may be reliably dealt with by shallow external components.

3 Automatic Lexical Acquisition

We have investigated Machine Learning (ML) methods applied to the acquisition of the information contained in the lexicon of the SRG.

ML applied to lexical acquisition is a very active area of work linked to deep linguistic analysis due to the central role that lexical information has in lexicalized grammars and the costs of hand-crafting them. Korhonen (2002), Carroll and Fang (2004), Baldwin (2005), Blunsom and Baldwin (2006), and Zhang and Kordoni (2006) are just a few examples of reported research work on deep lexical acquisition.

The most successful systems of lexical acquisition are based on the linguistic idea that the contexts where words occur are associated to particular lexical types. Although the methods are different, most of the systems work upon the syntactic information on words as collected from a corpus, and they develop different techniques to decide whether this information is relevant for type assignment or it is noise, especially when there are just a few examples. In the LKB grammatical framework, lexical types are defined as a combination of grammatical features. For our research, we have looked at these morpho-syntactically motivated features that can help in discriminating the different types that we will ultimately use to classify words. Thus, words are assigned a number of grammatical features, the ones that define the lexical types.

Table 1 and Table 2 show the syntactic features that we use to characterize 6 types of adjectives and 7 types of nouns in Spanish, respectively.[9] As can be observed, adjectives are cross-classified according to their syntactic position within the NP, i.e. (*preN*(ominal)) vs *postN*(ominal), the possibility of co-occurring in predicative constructions (*pred*) and being modified by degree adverbs (*G*), and their subcategorization frame (*pcomp*); whereas lexical types for nouns are basically defined on the basis of the *mass/countable* distinction and valence information. Thus, an adjective like *bonito* (nice), belonging to the type *a_qual_intr*, may be found both in pre-nominal and post-nominal position or in predicative constructions, it may also be modified by degree adverbs, this type of adjectives, however, does not take complements. Nouns belonging to the type *n_intr_count*, like *muchacha* (girl), are countable intransitive nouns.

TYPE/SF	preN	postN	pred	G	pcomp
a_adv_int	yes	no	no	no	no
a_adv_event	yes	yes	no	no	no
a_rel_nonpred	no	yes	no	no	no
a_rel_pred	no	yes	yes	no	no
a_qual_intr	yes	yes	yes	yes	no
a_qual_trans	yes	yes	yes	yes	yes

Table 1. Some adjectival types of the SRG

TYPE/SF	mass	count	intr	trans	pcomp
n_intr_mass	yes	no	yes	no	no
n_intr_count	no	yes	yes	no	no
n_intr_cnt-mss	yes	yes	yes	no	no
n_trans_mass	yes	no	no	yes	no
n_trans_count	no	yes	no	yes	no
n_ppde_pcomp_count	no	yes	no	yes	yes
n_ppde_pcomp_mss	yes	no	no	yes	yes

Table 2. Some nominal types of the SRG

We have investigated two methods to automatically acquire such linguistic information for Spanish nouns and adjectives: a Bayesian model and a decision tree. The aim of working with these two methods was to compare their performance taking into account that while the decision tree gets the information from previously annotated data, the Bayesian method learns it from the linguistic typology as defined by the grammar. These methods are described in the following subsections.

3.1 A Bayesian model for lexical acquisition

We have used a Bayesian model of inductive learning for assigning grammatical features to words occurring in a corpus. Given a hypothesis space (the linguistic features of words according to its lexical type) and one or more occurrences of the

[9] The SRG has 35 types for nouns and 44 types for adjectives.

word to classify, the learner evaluates all hypotheses for word features and values by computing their posterior probabilities, proportional to the product of prior probabilities and likelihood.

In order to obtain the likelihood, grammatical features are related to the expected contexts where their instances might appear. The linguistic typology provides likelihood information that is the learner's expectation about which contexts are likely to be observed given a particular hypothesis of a word type. This likelihood is used as a substitute of the computations made by observing directly the data, which is what a supervised machine learning method does. As said, our aim was to compare these two strategies.

The decision on a particular word is determined by averaging the predictions of all hypothesis weighted by their posterior probabilities. More technically, for each syntactic feature $\{sf_1, sf_2, ..., sf_n\}$ of the set SF (Syntactic Features) represented in the lexical typology, we define the goal of our system to be the assignment of a value, {*no, yes*}, that maximizes the result of a function $f: \sigma \rightarrow SF$, where σ is the collection of its occurrences ($\sigma = \{v_1, v_2, ..., v_z\}$), each being a *n*-dimensional vector. The decision on value assignment is achieved by considering every occurrence as a cumulative evidence in favour or against of having each syntactic feature. Thus, our function $Z'(SF, \sigma)$, shown in (5), will assess how much relevant information is got from all the vectors. A further function, shown in (8), will decide on the maximal value in order to assign $sf_{i,x}$.

$$(5)\quad Z'(sf_{i,x}, \sigma) = \sum_{j}^{z} P(sf_{i,x} \mid v_j)$$

To assess $P(sf_{i,x}|v_j)$, we use (6), which is the application of Bayes Rule for solving the estimation of the probability of a vector conditioned to a particular feature and value.

$$(6)\quad P(sf_{i,x} \mid v_j) = \frac{P(v_j \mid sf_{i,x}) P(sf_{i,x})}{\sum_k P(v_j \mid sf_{i,k}) P(sf_{i,k})}$$

For solving (6), the prior $P(sf_{i,x})$ is computed on the basis of a lexical typology too, assuming that what is more frequent in the typology will correspondingly be more frequent in the data. For computing the likelihood $P(v_j|sf_{i,x})$, as each vector is made of *m* components, that is, the linguistic cues $v_z = \{lc_1, lc_2, ..., lc_m\}$, we proceed as in (7) on the basis of $P(lc_l|sf_{i,x})$; i.e. the likelihood of finding the word in a particular context given a particular syntactic feature.

$$(7)\quad P(v_j \mid sf_{i,x}) = \prod_{l=1}^{m} P(lc_l \mid sf_{i,x})$$

Finally Z, as in (8), is the function that assigns the syntactic features to σ.[10]

$$(8)\quad Z = \left\{ \begin{array}{l} Z'(sf_{i,x=yes} \mid \sigma) > Z'(sf_{i,x=no} \mid \sigma) \rightarrow yes \\ Z'(sf_{i,x=no} \mid \sigma) > Z'(sf_{i,x=yes} \mid \sigma) \rightarrow no \end{array} \right\}$$

For computing the likelihood, we count on the conditional probabilities of the correlations between features as defined in the typology. We use these correlations to infer the expectation of observing the linguistic cues associated to particular syntactic features, and to make it to be conditional to a particular feature and value. However, linguistic cues and syntactic features are in two different dimensions; syntactic features are properties of lexical items, while linguistic cues show the characteristics of actual occurrences. As we assume that each syntactic feature must have at least one corresponding linguistic cue, we must tune the probability to acknowledge the factors that affect linguistic cues. For such a tuning, we have considered the following two issues: (i) to include in the assessments the known uncertainty of the linguistic cues that can be present in the occurrence or not; and (ii) to create a dummy variable to deal with the fact that, while syntactic features in the typology are independent from one another, evidences in text are not so.

We have also observed that the information that can be gathered by looking at all word occurrences as a complex unit have a conclusive value. Take for instance the case of prepositions. The observation of a given prepositions in different occurrences of the same word is a conclusive evidence for considering it a bound preposition. In order to take this into account, we have devised a function that acts as a dynamic weighting module. The function $app\_lc(sf_i, \sigma)$ returns the number of contexts where the cue is found. In the case that in a

[10] In the theoretical case of having the same probability for *yes* and for *no*, *Z* is undefined.

particular signature there is no context with such a *lc*, it returns '1'. Thus, *app_lc* is used to reinforce this conclusive evidence in (5), which is now (9).

(9)

$$Z'(sf_{i,x=yes},\sigma) = \left(\sum_{j}^{z} P(sf_{i,x=yes} \mid v_j)\right) * app\_lc(sf_i,\sigma)$$

$$Z'(sf_{i,x=no},\sigma) = \sum_{j}^{z} P(sf_{i,x=no} \mid v_j)$$

3.2 A Decision tree

Linguistic motivated features have also been evaluated using a C4.5 Decision Tree (DT) classifier (Quinlan, 1993) in the Weka implementation (Witten and Frank, 2005). These features correspond to the expected contexts for the different nominal and adjectival lexical types.

We have trained the DT with all the vectors of the word occurrences that we had in the different gold-standards, using their encoding for the supervised experiment in a 10-fold cross-validation testing (Bel et al. 2007).

3.3 Evaluation and Results

For the evaluation, we have applied both methods to the lexical acquisition of nouns and adjectives.

We have worked with a PoS tagged corpus of 1,091,314 words. Datasets of 496 adjectives and 289 nouns were selected among the ones that had occurrences in the corpus. Some manual selection had to be done in order to have all possible types represented but still it roughly corresponds to the distribution of features in the existing lexicon.

We evaluated by comparing with Gold-standard files; i.e. the manually encoded lexicon of the SRG. The usual accuracy measures as *type precision* (percentage of feature values correctly assigned to all values assigned) and *type recall* (percentage of correct feature values found in the dictionary) have been used. F1 is the usual score combining precision and recall.

Table 3 shows the results in terms of F1 score for the different methods and PoS for feature assignment. From these data, we concluded that the probabilistic information inferred from the lexical typology defined in our grammar is a good source of knowledge for lexical acquisition.

PoS	**noun**	**adj**
Z	0.88	0.87
DT	0.89	0.9

Table 3. F1 for different methods and PoS.

Table 4 shows more details of the results comparing between DT and Z for Spanish adjectives.

	SF = no		**SF = yes**	
	Z	**DT**	**Z**	**DT**
prep_a	0.98	0.97	0.72	0.44
prep_en	0.98	0.99	0.27	0
prep_con	0.99	0.99	0.60	0
prep_para	0.98	0.99	0.51	0.53
prep_de	0.88	0.97	0.34	0.42
postN	0	0	0.99	0.99
preN	0.75	0.83	0.44	0.80
Pred	0.50	0.41	0.59	0.82
G	0.85	0.80	0.75	0.72
Sent	0.97	0.97	0.55	0.44

Table 4. F1 for Spanish adjectival features.

Finally, Table 5 shows the results for 50 Spanish nouns with only one occurrence in the corpus. These results show that grammatical features can be used for lexical acquisition of low frequency lexical items, providing a good hypothesis for ensuring grammar robustness and adding no overgeneration to parsing results.

	DT			***Z***		
	prec.	***rec.***	***F***	***prec.***	***rec.***	***F***
MASS	0.50	0.16	0.25	0.66	0.25	0.36
COUNT	0.97	1.00	0.98	1.00	0.96	0.98
TRANS	0.75	0.46	0.57	0.68	0.73	0.71
INTRANS	0.85	0.95	0.89	0.89	0.76	0.82
PCOMP	0	0	0	0.14	0.20	0.16

Table 5. Results of 50 unseen nouns with a single occurrence.

4 Future Work

We have presented work on the development of an HPSG grammar for Spanish; in particular, we have described our approach for pre-processing and ongoing research on automatic lexical acquisition. Besides extending the coverage of the SRG and continuing research on lexical acquisition, the specific aims of our future work on the SRG are:

- Treebank development.

- To extend the shallow/deep architecture and integrate the structures generated by partial parsing, to provide robust techniques for infrequent structural constructions. The coverage of these linguistic structures by means of structure rules would increase both processing time and ambiguity.

- To use ML methods for disambiguation; i.e. for ranking possible parsings according to relevant linguistic features, thus enabling the setting of a threshold to select the n-best analyses.

- The development of error mining techniques (van Noord, 2004) to identify erroneous and incomplete information in the linguistic resources which cause the grammar to fail.

References

J. Atserias, B. Casas, E. Comelles, M. González, L. Padró and M. Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. 5t*h International Conference on Language Resources and Evaluation*. Genoa, Italy.

T. Baldwin. 2005. Bootstrapping Deep Lexical Resources: Resources for Courses, *ACL-SIGLEX 2005. Workshop on Deep Lexical Acquisitio*n. Ann Arbor, Michigan.

N. Bel, S. Espeja, M. Marimon. 2007. Automatic Acquisition of Grammatical Types for Nouns. *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Rochester, NY, USA,

E.M. Bender, D. Flickinger and S. Oepen. 2002. The grammar Matrix. An open-source starter-kit for the rapid development of cress-linguistically consistent broad-coverage precision grammar. *Workshop on Grammar Engineering and Evaluation, 19th International Conference on Computational Linguistics*. Taipei, Taiwan.

P. Blunsom and T. Baldwin. 2006. Multilingual Deep Lexical Acquisition for HPSGs via Supertagging. *Conference on Empirical Methods in Natural Language Processing.* Sydney, Australia.

T. Brants. 2000. TnT: A statistical part-of-speech tagger. *6th Conference on Applied Natural Language Processing*. Seattle, USA.

T. Cabré and C. Bach, 2004. El corpus tècnic de l'IULA: corpus textual especializado plurilingüe. *Panacea, V*. 16, pages 173-176.

U. Callmeier. 2000. Pet – a platform for experimentation with efficient HPSG processing. *Journal of Natural Language Engineering 6(1): Special Issue on Efficient Processing with HPSG: Methods, System, Evaluation*, pages 99-108.

A. Copestake, D. Flickinger, C. Pollard and I.A. Sag. 2006. Minimal Recursion Semantics: An Introduction. *Research on Language and Computation* 3.4:281-332.

A. Copestake. 2002. I*mplementing Typed Features Structure Grammars*. CSLI Publications.

A. Korhonen. 2002. 'Subcategorization acquisition'. As Technical Report UCAM-CL-TR-530, University of Cambridge, UK.

M. Marimon. 2002. Integrating Shallow Linguistic Processing into a Unification-based Spanish Grammar. *9th International Conference on Computational Linguistics*. Taipei, Taiwan.

M. Marimon and N. Bel. 2004. Lexical Entry Templates for Robust Deep Parsing. *4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

S. Oepen and J. Carroll. 2000. Performance Profiling for Parser Engineering. *Journal of Natural Language Engineering 6(1): Special Issue on Efficient Processing with HPSG: Methods, System, Evaluation*, pages 81-97.

C.J. Pollard and I.A. Sag. 1994. *Head-driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

R.J. Quinlan 1993. C4.5: Programs for Machine Learning. Series in Machine Learning. Morgan Kaufman, San Mateo, CA.

C. Samuelson. 1993. Morphological tagging based entirely on Bayesian inference. *9th Nordic Conference on Computational Linguistics*. Stockholm, Sweden.

I.H. Witten and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

G. van Noord. 2004. Error mining for wide-coverage grammar engineering. *42th Annual Meeting of the ACL*. Barcelona, Spain.

Y. Zhang and V. Kordoni. 2006. Automated deep lexical acquisition for robust open text processing. 5t*h International Conference on Language Resources and Evaluation*. Genoa, Italy.