Use of Internet for Augmenting Coverage in a Lexical Acquisition System from Raw Corpora: application to Russian

Antoni Oliver
IN3
U. Oberta de Catalunya
aoliverg@uoc.edu

Irene Castellón
GRIAL Group
Lingüística General - UB
castel@lingua.fil.ub.es

Lluís Màrquez
TALP Research Center
LSI, UPC
lluism@lsi.upc.es

Abstract

This paper presents a methodology for the automatic acquisition of lexical resources from raw corpora. This methodology has proved to be efficient for those languages that, like Russian, present a rich and mainly concatenative morphology. This method can be applied for the creation of new resources, as well as in the enrichment of existing ones. We also present an extension of the system that uses automatic querying to Internet to acquire those entries for which there is not enough information in our corpus. The new basic acquisition methodology achieves similar results compared to the previous methods, but the use of Internet queries allows to increase recall levels with only a slight decrease in precision, obtaining significantly better overall results.

1 Introduction

The implementation of different NLP applications requires a lot of lexical information. Some languages does not have extensive resources, and for those languages that have resources, they are not always easily available. The compilation of lexical resources usually requires a lot of human effort. In this paper, we present a method for the automatic acquisition of lexical and morpho–syntactic information from raw corpora. This methodology can be useful for the creation of new resources, as well as in the enrichment of existing ones.

1.1 Previous work

Our first experimental systems for lexical acquisition were developed for Croatian (Oliver 01) and (Oliver *et al.* 02). The methodology was based on the co-occurrence of different forms of the paradigm in the corpus and the presence of the lemma was necessary to validate the acquisition. The main problem we found was the ambiguity of the rules for the acquisition task. This ambiguity caused the confusion of one form of the paradigm with the lemma of another paradigm and led to low precision results: 34.65% of precision and 85.25% of recall (F_1=49.27 for Russian). To solve this problem, we developed an algorithm to analyse the rules and classify them into ambiguous and non-ambiguous (Oliver *et al.* 03). Applying rule analysis, and acquiring with non ambiguous rules in two steps, we obtained quite good results of precision (93.49%) but a dramatic drop in recall (38.52%, F_1=54.55). To increase recall, we applied the same methodology to several partitions of the whole corpus into relatively small subsets, splitting the corpus in an alphabetical way, using one, two, and three initial letters. Best results were obtained with the alphabetical process with three initial letters (91.93% of precision and 67.19% of recall, F_1=77.64).

The new methodology presented in this paper offers several advantages. First, it does not need the presence of the lemma in the corpus to validate an acquisition, that is, is able to give the morpho–syntactic information an the related lemma of a word form even if the lemma is not present. Besides, the system offers information of all the possible interpretations of the forms that have not been acquired. This information will be very useful to try to disambiguate between competing alternatives using a bigger corpus or, as in our experiments, using Internet as a very big corpus.

1.2 Goal of the acquisition system

The main goal of the acquisition system is to automatically acquire a list, as complete as possible, of word forms with their associated lemma and morpho–syntactic information. The morpho–syntactic information is expressed by tags following the Multext recommendations (Véronis & Khouri 95; Erjavec 01). For instance, *word form*: мостом; *lemma*: мост; *part-of-speech*: NCMSI[1]; all expressed as мостом:мост:NCMSI. As an example, in tables 1 and 2 we can see the complete declension, explained with this notation, of the masculine noun мост (*bridge*) and the feminine noun карта (*map*).

[1] Singular Masculine Common Noun in Instrumental

мост:мост:NCMSN,NCMSAI
моста:мост:NCMSG,NCMSAA
мосту:мост:NCMSD
мостом:мост:NCMSI
мосте:мост:NCMSP
мосты:мост:NCMPN,NCMPAI
мостов:мост:NCMPG,NCMPAA
мостам:мост:NCMPD
мостами:мост:NCMPI
мостах:мост:NCMPP

Table 1: Declension of the A-type masculine noun мост (*bridge*)

карта:карта:NCFSN
карты:карта:NCFSG,NCFPN,NCFPAI
карте:карта:NCFSD,NCFSP
карту:карта:NCFSA
картой:карта:NCFSI
карт:карта:NCFPG,NCFPAA
картам:карта:NCFPD
картами:карта:NCFPI
картах:карта:NCFPP

Table 2: Declension of the A-type feminine noun карта (*map*)

2 Components of the acquisition system

The acquisition system needs some linguistic knowledge contained basically in the raw corpus and in the set of morphological rules. However, since the system deals only with inflectional regular forms, we also need word lists of irregular words and of words belonging to non-inflectional categories and closed categories.

2.1 Word-lists of non–inflectional categories and closed categories

The words belonging to non–inflectional and closed categories are excluded from the acquisition process. We have manually constructed lists of words of such categories. Adverbs are provisionally included in this list, but they cannot be considered as a closed class. We plan to include, in the near future, derivative rules for the most productive processes of adverb formation from other categories.

2.2 Irregular word list

Irregular words are also excluded from the acquisition process. These words are declared in a list that includes all forms with the associated lemma and morpho–syntactic information. This list is currently being developed by hand. Our criterion is to write all forms of the irregular words included in the 5,000 most frequent Russian words according to (Sharov 01).

The concept of irregularity is related to the non assignment of a lemma to a paradigm. For this reason, regularity and irregularity are related to the number of paradigms present in our model, and from the computational point of view, it is equivalent to the number of paradigms implemented in our system. Extreme cases are the Russian morphological analyser of Mikheev (Mikheev & Liubushkina 95) and the Croatian analyser of Tadić (Tadić 94). These analysers consider no irregular words, that is, all words must belong to a given paradigm, to the extend that some paradigms are applicable to only one word. Following this approach, the number of paradigms gets high values, and although this is not a problem for word analysis or generation, the high number of paradigms could be a problem for the automatic acquisition of lexical information.

2.3 Morphological rules

Morphological rules are implemented following a morphological stripping formalism (Alshawi 92). These rules are converted into Perl regular expressions at running time. The rules are of the form `FE:LE:Desc`, where `FE` stands for the form ending, `LE` for the lemma ending, and `Desc` for morphological description. For example, the generic rule

```
ом::NCMSI
```

may express the entry мостом:мост:NCMSI. As it can be observed, this rule represents a null lemma ending. By using Perl regular expressions we can describe the lemma ending with more precision. An example can be seen in the following rule:

```
([^аяеэиыоуюькгхжшщчцй])ом:\1:NCMSI
```

where '^' means the complementary set of symbols written between square brackets and '\1' is a variable representing the symbol matched by the regular expression between brackets. Regular expressions allow to express other complex morphological phenomena, such as vowel alternation. For example, the rule:

```
ль([^аяеэиыоуюькгхжшщчцй])а:ле\1:NCMSG
```

may express an entry as `льва:лев:NCMSG`.

The rules have been hand developed following the most productive models of the (Zaliznjak 77) dictionary. We have developed 565 rules for nouns, 219 rules for adjectives, and 12,038 rules for verbs. The high number of rules corresponding to verbs is due to the fact that some forms, as participles, are declinable. The rules corresponding to declined forms are derived automatically from the rule that expresses the base form. Not all the rules will be used in the acquisition process. Those rules expressing alternative endings that are equal to other endings in the paradigm are left out. For example, the rule

```
([^аяеэиыоуюькгхжшщчцй])а:\1:NCMPN
```

that expresses the alternative plural nominative in "a" for the masculine nouns, is left out because the `FE` "a" is equal to that of the genitive singular. Aspectual information of verbs has also been eliminated.

At running time, rules are transformed into Perl substitutions, for example the rule

```
([^кгхжшщчц])у:\1а:NCFSA
```

is transformed into the substitution

```
s/([^кгхжшщчц])у$/\1а/
```

This expression means "replace the final `у` by an `а`, if the preceding letter does not belong to the list `{к,г,х,ж,ш,щ,ч,ц}`". This substitution allows the formation of the lemma of a feminine noun from its singular accusative.

2.4 Corpus

A 16,000,000 word corpus has been compiled from newspapers and literary texts[2]. The corpus has been automatically segmented into sentences and no other kind of linguistic information has been added.

3 Acquisition methodology

The acquisition methodology can be divided in seven steps. The first three steps (rule expansion, filtering and reorganisation) are related to the adaptation of the rule notation for the acquisition process. The fourth step is related to the treatment of the word forms of the corpus. The fifth step is the acquisition process and the two last steps consist of the creation of the files of unsolved entries and the querying to Internet.

3.1 Rule expansion

Rules expressing morphological contexts are expanded to all their possibilities. For example:

```
([^аяеэиыоуюькгхжшщчцй])ом:\1:NCMSI
```

is expanded to the following rules:

```
ъом:ъ:NCMSI  фом:ф:NCMSI  том:т:NCMSI
сом:с:NCMSI  ром:р:NCMSI  пом:п:NCMSI
ном:н:NCMSI  мом:м:NCMSI  лом:л:NCMSI
зом:з:NCMSI  дом:д:NCMSI  вом:в:NCMSI
бом:б:NCMSI
```

3.2 Rule filtering

The expansion of the rules leads to a multiplication of the number of rules. Some of these rules cannot be applied because the combinations of final letters expressed by the rules are not found in the language (or at least are not found in the corpus). After rule expansion, rules are filtered using the set of letter n–grams extracted from the word–form endings in our corpus. By this process a large number of rules that cannot be applied are eliminated, and make the rest of the process faster. For example, our set of 9,821 rules are expanded to 74,956 rules, and after filtering 22,033 rules remain.

In the filtering process of the example above, the rule `ъом:ъ:NCMSI` was eliminated because the ending `ъом` is not found in the corpus. No matter if the ending actually exists, if it does not exist in our corpus the rule is not applicable. The rule filtering process must be repeated each time that the corpus is modified.

3.3 Rule reorganisation

Once expanded and filtered, the morphological rules are reorganised. We make a list of endings that share the part of speech and flexive model, as well as the lemma ending. These groups are stored indexed by flexive group and lemma ending, as in the example:

```
NCMA:т   т,та,ту,том,те,ты,тов,там,тами,тах
NCFA:та  та,ты,те,ту,той,т,там,тами,тах
```

3.4 Splitting word forms and grouping by possible paradigms

The algorithm splits all the word forms in the corpus (in stem and ending) by all possible end-

[2] Thanks to Огонек, Правда and Библиотека Максима Мошкова for letting us use their texts

ings and they are grouped by paradigms. To explain this step, let us consider that the corpus is formed by the words: мост, моста, мостом, мостах, карта, карты, картой and картах. If we split these words by all possible endings and group them, the following result would be obtained:

NCMA:мост	мост,моста,мостом,мостах
NCFA:моста	моста,мост,мостах
NCMA:карт	карта,карты,картах
NCFA:карта	карта,карты,картой,картах

3.5 Acquisition by comparison

Once all stems and endings are grouped, the acquisition process can be started. For each word in the corpus, all possible divisions are found and the system chooses as the correct one the division that has more forms in the corpus. For example, if we take the word form мостах we observe that can be associated with the group NCMA:мост or with the group NCFA:моста, but the first option has four associated forms and the second only three, so we choose the first option. With this information the word form мостах can be associated with the lemma мост and can be retrieved the morpho–syntactic information from the morphological rules. Then a new entry can be created: мостах:мост:NCMPP. In the same way, the word form карты can be associated with the group NCMA:карт, with three associated forms, or with the group NCFA:карта that has four associated forms. This last option is taken as the correct one.

It is worth noting that is not necessary for the lemma to be present in the corpus. We can acquire an entry even if the lemma does not occur. For example, if our corpus is formed by the word forms: моста, мостом, мостах, карта, карты, картой, and карта, that is, the lemma мост is not present, the grouping of possible paradigms would be as follows:

NCMA:мост	моста,мостом,мостах
NCFA:моста	моста,мостах

If we now consider the form мостах, it can be associated with the group NCMA:мост and with the group NCFA:моста, but the first has now three associated forms, and the last only two. With all this information we can associate the word form мостах with the lemma мост, which actually does not occur in the corpus.

3.6 Creation of the file of unsolved entries

In the examples above everything worked fine because there were present in the corpus some word forms with endings belonging to one paradigm but not to the other (мостом and картой). Let us consider now that the corpus is formed by the word forms: мост, моста, мосту, мостах, карта, карты, карту, and картах. After splitting word forms and grouping by possible paradigms we get the following results:

NCMA:мост	мост,моста,мосту,мостах
NCFA:моста	моста:мост:мосту:мостах
NCMA:карт	карта:карты:карту:картах
NCFA:карта	карта:карты:карту:картах

If the system tries to acquire the lemma and the morpho–syntactic information associated with the word form мостах, two possible options are found: NCMA:мост and NCFA:моста, both with four associated word forms. Thus it is impossible to determine the correct option. The same happens if we try to acquire the associated information of the word form карты because we find the two tied options NCMA:карт and NCFA:карта with the same number of associated word forms.

In cases like the previous ones, our system generates a file of unsolved entries. In this file all the options for each unsolved entry are specified. In the examples above this file would show the following information:

word form: мосту	*word form*: карты
opt. 1: мос,NCMA,т	*opt. 1*: кар,NCMA,т
opt. 2: мос,NCFA,та	*opt. 2*: кар,NCFA,та

For each option, this file gives the associated stem, the flexive group, and the lemma ending. This file of unsolved entries is interesting because we can use it in next steps to try to solve it with a bigger corpus, or by Internet querying.

3.7 Querying to Internet

The unsolved entries are mainly due to the lack of certain forms of the paradigm in the corpus. If we increase enough the size of the corpus the system would find the needed word forms to solve the ambiguities. Existing corpora or Internet search engines can be used to verify the existence of such word forms that can help us to discriminate between the different options. To determine the forms that discriminate between options we generate all the forms corresponding to each option.

The discriminating forms will be those present in one model but not in the others. Following the example above, to discriminate the options of the word form мостах we would generate all the forms corresponding to the option мос:NCMA:т (мост, моста, мосту, мостом, мосте, мосты, мостов, мостам, мостами, мостах) and all the forms corresponding to the option мос:NCFA:та (моста, мосты, мосте, мосту, мостой, мост, мостам, мостами, мостах). The algorithm then composes the queries from the non common word forms among the different options, that is, in our example the queries are composed from the word forms мостом and мостов, corresponding to the first option, and the word form мостой corresponding to the second option. In the example, it would generate the query 'мостом||мостов' (the symbol || means OR) and the query 'мостой'. The first query would return a greater number of documents than the second, so the system would validate the first option. In a similar way, to discriminate between the options of the word form карты we generate the queries 'картом||картов' and 'картой'. In this case, the second query would return a greater number of documents, validating the кар:NCFA:та option. In the case of verbs, there are a lot of discriminating forms, so we limit the number of forms to make the query, otherwise we would generate errors in the Internet search engine.

4 Experimental evaluation

A test corpus has been built only with regular and known forms, large enough and with a distribution of lemmas and forms as real as possible. All the forms are known so the result of each experiment can be evaluated automatically. The test corpus was built as follows: a word–form list was created with all the forms of 78,519 lemmas, totalling 1,247,202 forms. Each of these forms was included only if it occurs in our 16,000,000 word corpus of Russian texts. The result is a corpus of 232,770 regular word forms corresponding to 43,543 lemmas. The resulting corpus is, in fact, a word–form list.

To solve the file of unsolved entries we use two Internet search engines: *Yahoo*[3] and the Russian engine *Yandex*[4].

Additionally, we have carried out a simulation of querying to Internet using a Russian morphological analyser, *mystem*, belonging to the Russian company Yandex[5]. We have carried out this simulation for two main reasons: First, the process of Internet querying is slow, since there is a great amount of queries to do and we were allowed to perform only a limited number of queries per day. Second, because we are interested in comparing the results using an existing tool and using Internet as a big corpus.

Due to the limit of queries per day, we can not offer the complete results for both search engines. We will present the results obtained by the basic acquisition process (without resolution of the file of unsolved entries) and by resolution of unsolved entries both with the simulation using *mystem* and with the search engine Yahoo. On the contrary, for the Yandex search engine we can only present the result for four sub-corpus, namely, for all the words beginning with letters: а, б, в, and г.

4.1 Results

This section presents the results of our experiments, which we have evaluated with the common measures for recognition: precision (P), recall (R) and the harmonic mean between precision and recall (F_1=2PR/(P+R)).

In table 3 can be seen the results obtained for the treatment of whole corpus. In the '*Acq.*' row the results of the plain acquisition method are presented. '*Sim.*' row contains the results of the acquisition process with resolution of the doubts by simulation through *mystem*. Finally, the rows labelled with different values of n contain the results of the acquisition process with resolution by Internet querying through Yahoo. The parameter n means the minimum number of documents that must be returned in order to validate a query. With this parameter the system can avoid to validate a query with a very low number of returned documents, that may correspond to misspelled words. In this way, the balance between precision and recall can be adjusted.

The results obtained with the new acquisition methodology without resolution of the doubts are much better than those obtained with the previous methodology (Oliver *et al.* 03) without alphabetical process (F_1 of 75.43 compared to 54.55). Compared with the best results of the previous

[3] `http://www.yahoo.com`
[4] `http.//yandex.ru`

[5] Thanks to Yandex for the possibility to use their morphological analyser in our experiments

	OVERALL			Nouns			Adjectives			Verbs		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Acq.	95.53	62.32	75.43	97.76	50.92	66.96	99.46	64.82	78.49	90.49	81.57	85.8
Sim.	94.47	78.26	**85.60**	97.35	67.27	**79.59**	99.07	83.72	**90.75**	87.18	93.05	**90.02**
n=1	89.62	78.80	83.86	90.65	69.30	78.55	94.99	82.10	**88.08**	83.87	93.34	88.35
n=10	92.02	77.47	**84.12**	94.58	68.17	**79.23**	96.54	79.85	87.41	85.25	92.74	**88.84**
n=100	92.78	75.42	83.20	95.30	66.10	78.06	97.61	77.31	86.28	85.94	91.37	88.57
n=1000	93.42	68.84	79.27	95.85	58.76	72.86	98.69	70.30	82.11	86.77	86.79	86.78

Table 3: Results of the acquisition process with resolution by simulation and the search engine Yahoo for different values of n for the whole corpus

	n=1			n=10			n=100			n=1000		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
а acq.	99.22	70.07	82.14									
а simul.	98.51	89.24	**93.65**									
а yahoo	97.87	90.18	**93.87**	98.62	89.40	93.78	98.75	86.89	92.44	98.84	79.03	87.83
а yandex	97.84	90.21	93.87	98.44	89.87	**93.96**	98.68	87.59	92.80	98.78	80.40	88.65
б acq.	94.87	61.84	74.87									
б simul.	94.78	78.35	**85.79**									
б yahoo	90.65	79.70	84.82	92.45	78.64	**84.99**	92.90	75.70	83.42	93.53	68.84	79.31
б yandex	90.13	79.48	84.47	91.49	79.06	**84.82**	92.45	76.49	83.72	93.24	70.19	80.09
в acq.	95.42	69.07	80.13									
в simul.	95.14	81.80	**87.97**									
в yahoo	91.39	82.35	86.63	93.53	81.01	**86.82**	94.27	79.86	86.47	94.89	75.46	84.07
в yandex	90.84	82.47	86.45	92.63	81.76	**86.86**	93.96	80.27	86.58	94.65	76.07	84.35
г acq.	94.55	59.37	72.94									
г simul.	94.22	78.60	**85.70**									
г yahoo	89.38	72.93	80.32	91.67	74.95	**82.47**	92.92	73.48	82.06	94.26	67.88	78.92
г yandex	89.19	79.44	84.03	91.06	78.74	**84.45**	92.5	76.73	83.88	93.71	70.32	80.35

Table 4: Results of the acquisition process with resolution by simulation and the search engines Yahoo and Yandex for the letters а, б, в and г

methodology, corresponding to an alphabetic process with three initial letters, the new results are slightly worse but comparable (F_1 of 75.43 compared to 77.64). The new method, however, is faster and allows the possibility to treat the file of unsolved entries. Simulating the Internet querying through *mystem* we obtain an F_1 of 85.60, which is 10 points better than the result of the acquisition, with an increase of 16 points in recall and with a slight decrease of only 1 point in precision. With the real queries to Yahoo we obtain similar results (F_1=84.12 for $n = 10$).

In table 4 we can observe partial results for the sub-corpora of word forms beginning with the letters а, б, в and г. For brevity reasons, this table presents the global results and not the results for each part-of-speech (they show a similar behaviour than those in table 3). From these results, it can be observed that, for $n = 1$, the system obtains better results of precision with the simulation, but better results of recall with the queries to Internet. We can also observe that for all values of n we obtain a great increase of F_1 compared with the acquisition process alone. In tables 3 and 4 best results are marked in boldface.

In figure 1 we can observe the evolution of the overall precision, recall, and F_1 values, by varying the n parameter from 1 to 1,000. These values are calculated for the acquisition and the resolution by the engine Yahoo for the whole corpus. In figure 2 we can observe the values of the F_1 measure (again varying n between 1 and 1,000) for nouns, adjectives, verbs, and the overall value calculated in the same way. As we can observe, best values of F_1 are obtained for verbs, because this is the category with more word forms per lemma. Consequently, the lowest values are obtained for nouns. The best overall value of F_1 (84.17) is obtained for $n = 7$. Values of n greater than 100 lead to significant decreases in performance.

In our experiments, we used Internet only to discriminate between different options for those entries for which we do not have enough information in the corpus. The process implies to verify the existence of several word forms not existing in our corpus by querying to Internet engines. Once

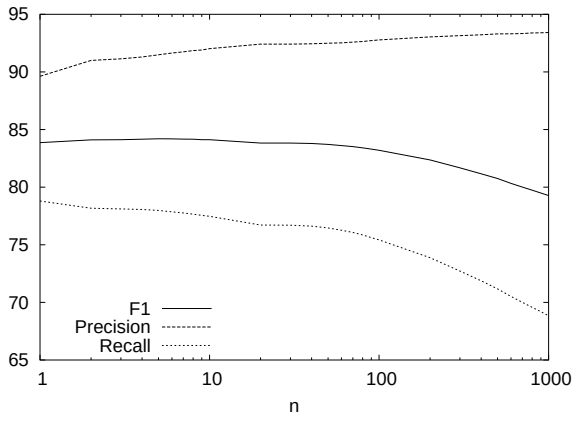

Figure 1: Precision, recall and F_1 versus n for the acquisition process with resolution by Yahoo for the whole corpus

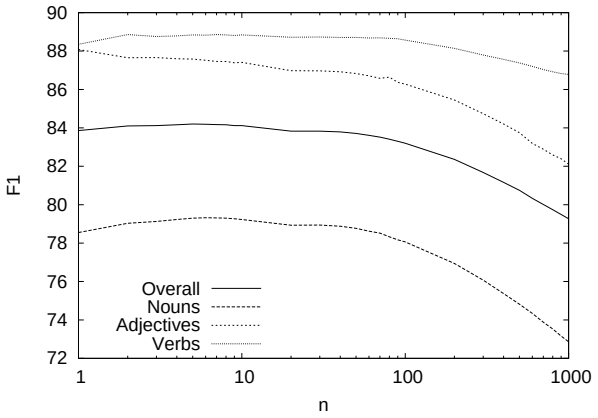

Figure 2: F_1 versus n for the acquisition process with resolution by Yahoo for the whole corpus

the existence of these word forms is verified, we can create new entries and acquire more lexical information. We have not evaluated this possibility but we plan to do this in the near future. With this extension the methodology will be able to acquire lexical information about word forms not present in the corpus.

5 Conclusions and future work

In this work, a methodology for the automatic acquisition of lexical and morpho–syntantic information from raw corpora has been presented. The methodology uses querying to Internet in order to improve the recall. This methodology has been successfully applied to a large corpus of Russian, showing that it can be very useful for the creation of new lexical resources, as well as for the improvement of existing resources. It also allows to avoid the compilation of very big corpora.

The results obtained are quite good, achieving a great improvement compared to the systems developed previously in our research (being the F_1 measure almost 10 points higher). We have tested this new methodology with Russian, a language characterised by a very rich morphology. Future work includes testing the methodology in languages such as Croatian, Spanish and Catalan.

In the experiments conducted so far, morphological rules have been written by hand based on traditional grammars as (Zaliznjak 77) for Russian and (Barić *et al.* 95) for Croatian. We plan to develop some algorithms to learn the most productive paradigms from the raw corpus, as in (Goldsmith 01). These algorithms will allow to learn also the most productive derivative processes.

Acknowledgements

This investigation is partially supported by the projects INTERLINGUA (Universitat Oberta de Catalunya and IN3 – IR266) and HERMES (TIC 2000–0335–C03–02).

6 Bibliography

References

(Alshawi 92) H. Alshawi, editor. *The Core Language Engine*. MIT Press, 1992.

(Barić *et al.* 95) E. Barić, M. Lončarić, D. Malić, S. Pavešić, M. Peti, V. Zečević, and M. Znika. *Hrvatska gramatika*. Školska Knjiga, Zagreb, 1995.

(Erjavec 01) T. Erjavec. Specifications and notation for multext-east lexicon encoding. Technical report, Multext-East/Concede.(http://nl.ijs.si/MTE/V2), 2001.

(Goldsmith 01) J. Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

(Mikheev & Liubushkina 95) A. Mikheev and L. Liubushkina. Russian morphology: An engineering approach. *Natural Language Engineering*, 1(3):235–260, 1995.

(Oliver 01) A. Oliver. Processament morfològic de les llengües eslaves: el serbo-croat. Unpublished M.Sc. thesis, Treball per a l'obtenció de la D.E.A: Universitat de Barcelona, 2001.

(Oliver *et al.* 02) A. Oliver, I. Castellón, and L. Màrquez. Adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar: aplicación al serbo-croata y ruso. *Procesamiento del Lenguaje Natural*, 29:97–104, 2002.

(Oliver *et al.* 03) A. Oliver, I. Castellón, and L. Màrquez. Automatic lexical acquisition from raw corpora: An application to russian. In *Workshop on Morphological Processing of Slavic Languages*, pages 17–24. Association for Computational Linguistics, 2003. 10th Conference of The European Chapter of the EACL.

(Sharov 01) S.A. Sharov. Chastotnyi slovar. www.artint.ru/projects/frqlist.asp, 2001.

(Tadić 94) M. Tadić. *Računalna obradba morfologije hrvatskoga književnog jezika*. Unpublished PhD thesis, Sveučilište u Zagrebu, Filozofski fakultet. Zagreb, 1994.

(Véronis & Khouri 95) J. Véronis and L. Khouri. Etiquetage grammatical multilingue: modèle. Technical report, MULTEXT Project, http://www.lpl.univ-aix.fr/projects/multext/LEX/LEX2.html, 1995.

(Zaliznjak 77) A.A. Zaliznjak. *Grammaticheskii slovar russkogo jazyka. Slovoizmenenie*. Izdatelstvo "Russkii jazyk" Moskva., 1977.