FI MU

Faculty of Informatics
Masaryk University

DESAM—Approaches to Desambiguation

by

Karel Pala
Pavel Rychlý
Pavel Smrž

FI MU Report Series FIMU-RS-97-09

 December 1997

DESAM – Approaches to Desambiguation

Karel Pala, Pavel Rychlý and Pavel Smrž
Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: {pala,pary,smrz}@fi.muni.cz

December 22, 1997

Abstract

This paper deals with Czech desambiguated corpus DESAM. It is a tagged corpus which was manually desambiguated and can be used in various applications. We discuss the structure of the corpus, tools used for its managing, linguistic applications, and also possible use of machine learning techniques relying on the desambiguated data. Possible ways of developing procedures for complete automatic desambiguation are considered.

1 Introduction

In computational linguistics, "corpus" is a collection of written (or sometimes spoken) texts. Corpora could be used in several application areas: building dictionaries, general linguistic research, natural language processing, information retrieval, machine translation etc.

In corpus exploration, a user must be able to express the query as precisely as possible in order to minimize the number of concordance items searched for. It should be possible to refer to linguistic or structural information in corpus. We use the term "tagged (annotated) corpus" for a corpus which contains not only a sequence of words but also comprises an additional information. Typically, this includes linguistic information which is associated with the particular word forms in corpus: the most common linguistic tags are *lemma* (the basic word form), *part of speech (POS)* and the respective *grammatical categories*. Another level of annotations concerns *structural information* which indentifies a metatext structure of the text in corpus. For example, we can mark (annotate) that the sequence of word forms is a part of the headline or a regular sentence in a paragraph [1].

The most reasonable way how to build large annotated corpora is an automatic tagging of the texts by computer programmes. However, natural languages display rather complex structure and therefore it is no surprise that the attempts to process them by the simple deterministic algorithms do not always yield satisfactory results. The result is that the present tagging programmes are not able to give fully reliable results and there are many ambiguities in their output.

Various strategies trying to resolve the ambiguities in the tagged corpora have been developed and applied within the field of corpus linguistics. The most frequently used are the following:

1. *probabilistic techniques* like the ones used in CLAWS tagger [2] or Cutting's tagger [3]
2. deterministic rule-based techniques using CF-like formalisms which may be enhanced with some heuristic rules
3. various combination of the former two approaches, e. g. constraint grammars approach [4]
4. attempts to apply learning techniques that would make use of previous experience in the process of desambiguating

2 The annotated Czech corpus – DESAM

The DESAM corpus has been built at Faculty of Informatics, Masaryk University, as a part of the complex project whose main purpose is to build Czech National Corpus (200 mil. word forms by the end of 1999). The DESAM corpus is:

- a **Czech** corpus: included texts are written in Czech language.
- a **general** corpus: subject field is not specifically restricted. Texts are taken from newspapers and scientific magazines (Lidové Noviny, MF Dnes, Českomoravský Profit, Vesmír, Chip).
- a **tagged** corpus: a lemma and a tag is stored for each word form in the corpus.

DESAM corpus is the first annotated and fully **desambiguated** corpus for Czech language that can be run under the corpus manager CQP (see below) and its presented version will be later included in Czech National Corpus as its annotated part.

The LEMMA programme ([5], [6]) has been used for tagging texts included in DESAM. This programme is able to perform full morphological analysis of the arbitrary raw Czech text and for each Czech word form it yields the following output:

- its lemma or lemmata, i.e. as it is usual in Czech grammar, nominative singular for nouns, adjectives, pronouns and numerals, and infinitive for verbs

- its corresponding POS symbol: presently we work with 10 basic parts of speech that are normally distinguished in Czech grammars, however, the complete list of POS tags contains about 30 items (including subclassifications). If a word form can be associated with two or more POS symbols LEMMA offers all of them
- all the grammatical categories associated with a word form, ie. in Czech this includes for nouns, adjectives, pronouns and numerals: `case`, `number` and `gender`; for verbs: `person`, `number`, `tense`, `mode`, `voice`, `aspect` and also `gender`
- optionally, also all word forms that can be generated from a given lemma.

Moreover, in the course of corpus tagging LEMMA produces all the possible combinations of lemmata and the respective tags for each input word form. A tag is conceived as a character string carrying information about POS and the respective grammatical categories using attribute-values coding convention. The total number of all the tags occurring in DESAM is about 1800, however, for Czech language as a whole the estimated number of tags is higher than 2000. In the following example the tag **k1gMnSc1** means: part of speech (**k**) = noun (**1**), gender (**g**) = male animate (**M**), number (**n**) = singular (**S**) and case (**c**) = first (**1**). Example of the LEMMA output:

```
Václav <l>Václav <c>k1gMnSc1
Havel <l>Havel <c>k1gMnSc1
přišel <l>přijít <c>k5eApMnStMmPaP,k5eApInStMmPaP
naopak <l>naopak <c>k6xMeA
s <l>s <c>k7c7
vlastním <l>vlastní <c>k2eAgMnSc67d1,k2eAgXnPc3d1,k2eAgUnSc67d1
         <l>vlastnit <c>k5eAp1nStPmIaI
volebním <l>volební <c>k2eAgMnSc67d1,k2eAgXnPc3d1,k2eAgUnSc67d1
programem <l>program <c>k1gInSc7
,
který <l>který <c>k3xQgMnSc15,k3xQgInSc145
nikomu <l>nikdo <c>k3xNnSc3
neubližuje <l>ubližovat <c>k5eNpMnStPmTaI,k5eNp3nStPmIaI
.
```

(English translation: Václav Havel, on the contrary, came with his own "election programme" which does not hurt anybody.)

The original version of LEMMA processed only individual words without considering any context.Its present version, however, is already able to process the set of basic Czech collocations (about 200 items). If we have look at the output of LEMMA we can see that more than 50% of the processed word forms have more than one possible lemma and respective tag associated with them which means that we are facing the problem of desambiguation and we have to look for a way how to solve it. The reasons for desambiguation are evident:

- if we want to perform syntactic analysis of the tagged Czech text we need to remove as many ambiguities as possible

- if we want to have a reliably tagged corpus we also need a successful desambiguation procedure.

Therefore we had to make a decision: either to start with the poorly desambiguated corpus using the output from LEMMA as it stood or to begin with the manual desambiguation. We took the latter direction and the texts in DESAM corpus were mainly desambiguated manually. A programme DES [7] which navigates and helps users in manual desambiguating has been developed for this purpose. In [7] one can find a first basic analysis of ambiguity measures for Czech. The results are summarized in Table 1.

	before analysis	**after analysis**
Total word forms	10,300	10,300
Ambiguous word forms	5,200	2,950
Total tags	33,360	16,680
Tags per ambiguous word form	5.93	2.73

Table 1: Measures of desamgiuation

3 Corpus tools

It is obvious that large corpora should be easily accessible and the users should be equipped with a friendly environment enabling them to ask as many various queries as possible. For this purpose the IMS Corpus Workbench [8] has been chosen and has been installed at one of our SUN workstations. It is a set of tools for the administration and representation of large text corpora and retrieval of the information from the corpora. One of the tools is a query processor CQP [8] which evaluates given queries and returns the result on the screen or to another output that can be used in further processing. For more comfortable interaction or presentation there is XKWIC [9], a graphical user interface running in X–Window system.

Within the IMS Corpus Workbench, a corpus is represented as a sequence of *positions*. Each position is a set of *attributes* and each attribute contains a character information. Presently we work with three attributes in DESAM corpus: **word** which represents the particular word form at this position, **lemma** representing the corresponding basic word form and **tag** associated with this possition.

As we said above we can also store some *structural tags* in the corpus. At the present moment we use two structural tags in DESAM corpus: **doc** for documents and **p** for paragraphs. Sentence boundaries have not been systematically tagged yet, however, the standard delimiters are present in the corpus and they are going to be tagged quite soon (a special programme is being developed for this purpose). Apart from this the whole corpus is also divided into many documents using the **doc** tags. In most cases, a document represents a newspaper article. Each document is divided into paragraphs using the **p** tags. Some small articles may consist of one paragraph only.

4 Linguistic results

The information about the current size of DESAM corpus is displayed in Table 2. We would like to stress that the presented tables yield new and quite interesting data about Czech language.

Documents	3 056
Paragraphs	27 763
Positions	1 247 594
Word forms (types)	1 026 733
Different word forms(tokens)	132 447
Different lemmata	34 606
Different tags	1 665
Type/token ratio	7.75
Word forms occurring once	67 059
Lemmata occurring once	11 759

Table 2: Counts of the DESAM corpus

Table 1 presents the total frequencies of words, lemmata and tags in DESAM. It can be seen that type/token ratio in DESAM (which can be considered as a good estimation for Czech in general) is 7.75 – that reflects the highly inflected nature of Czech and we regard it as one of the most interesting findings in the present research. Another important result is the measure of ambiguity in DESAM (a good approximation for Czech language as well) – we have obtained value 4,81 tags per ambiguous word form.

Table 2 also shows, roughly speaking, that 50% of different words, 34% of different lemmata are hapax legomena, ie. they occur only once in our corpus texts. This is in good agreement with some previous statistical findings, particularly with Czech Frequency Dictionary [10].

The most frequent word forms and lemmata and their respective frequencies are displayed in Table 3.

The frequencies presented in the table again fit well into the picture one can find in the existing Czech frequency dictionaries, however, the occurrence of the two verbs *být* a *mít* among 15 most frequented items is quite interesting and could be perhaps the most plausibly explained by the fact that about one third of DESAM consists of the scientific text and the rest comes from newspapers.

Using XKWIC it is possible to make more sophisticated queries. For example, we can select all contexts where word forms **s** and **se** (two forms of "with" in English) are used as prepositions – by means of the following query:

```
[word="se?"%c & pos="k7.*"]
```

Word forms	
a	22542
v	17364
se	15933
na	13671
je	9553
že	7565
s	6344
o	6059
z	5761
i	5315
do	4910
to	4172
pro	3959
ve	3645
k	3390
za	3254
by	3086
si	2844
ale	2795

Lemmata	
být	24950
a	23122
v	19791
sebe	16554
na	15005
ten	9180
že	7606
který	7121
s	6860
z	6341
o	6295
mít	5480
i	5445
on	5229
do	5217
pro	4380
ve	4054
tento	4004
k	3897

Tags	
k7c6	36865
k8xC	30519
k5eAp3nStPmIaI	28739
k7c2	22013
k1gFnSc1	21766
k1gFnSc2	21445
k9	21440
k7c4	21291
k8xS	18411
k1gInSc2	16923
k1gInSc1	16577
k1gMnSc1	15122
k1gFnSc4	14776
k6xMeAd1	14127
k3xXnSc4	13362
k6xTeA	12758
k1gInSc4	12334
k1gFnPc2	12208
k6xMeA	11968

Table 3: Most frequent word forms, lemmata and tags

Table 4 shows the graph displaying the relation between the length of the corpus text and the number of lemmata: it can be seen that the first thousand lemmata covers almost 70% of the texts in DESAM. The second graph displays the "beginning" of the displayed relation.

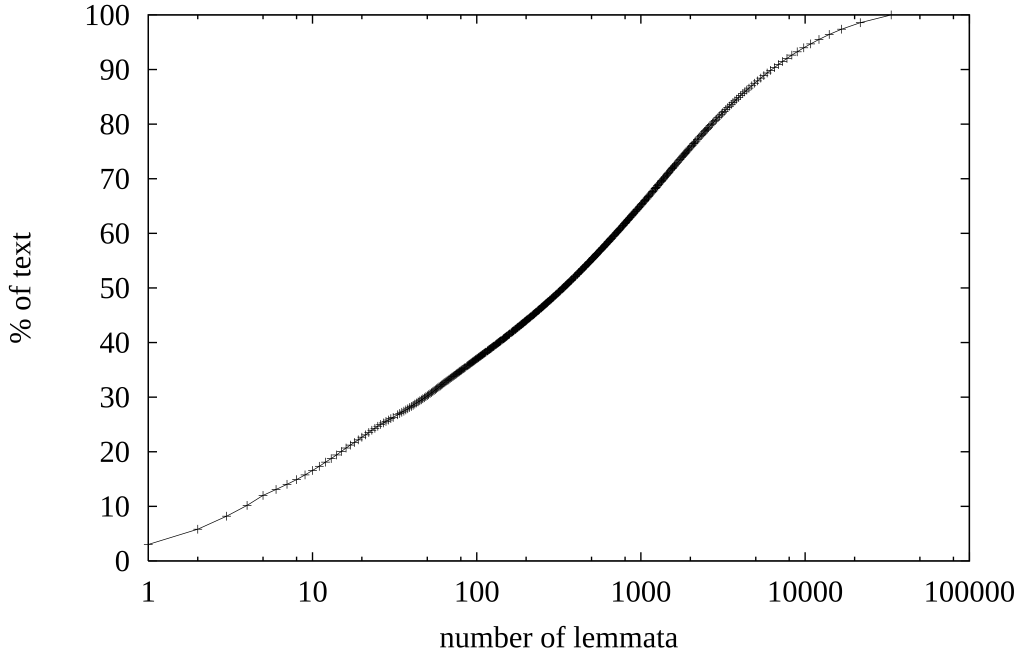
Similarly, Table 5 demonstrates the relation the length of the corpus and its coverage by the different word forms and the second graph offers again the closer look at the coverage of DESAM texts by the word forms. It is obvious that both presented graphs tell us something about the inflectional nature of Czech.

In the following example we have selected all the word forms **kolem** (meaning either *a wheel* or *around*) and computed the frequency distribution of the tags for each value of lemma (Table 6).

```
[word="kolem"%c]
```

Czech word form "kolem" could be:

- noun – lemma = "kolo" ("wheel" in English)
- preposition – lemma = "kolem" ("around" in English)
- adverb – lemma = "kolem" ("round" in English)



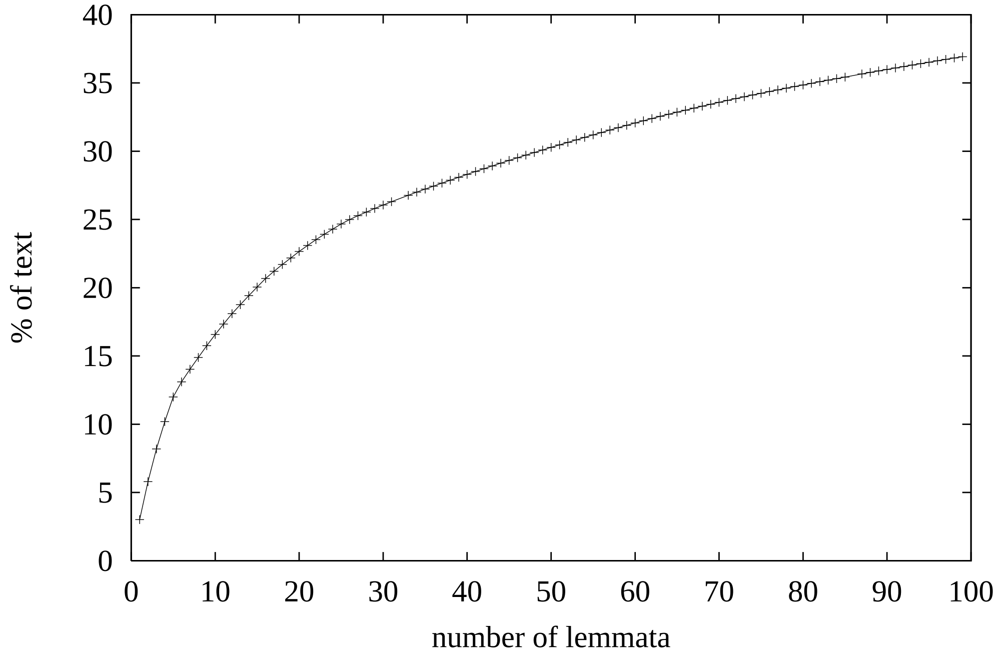
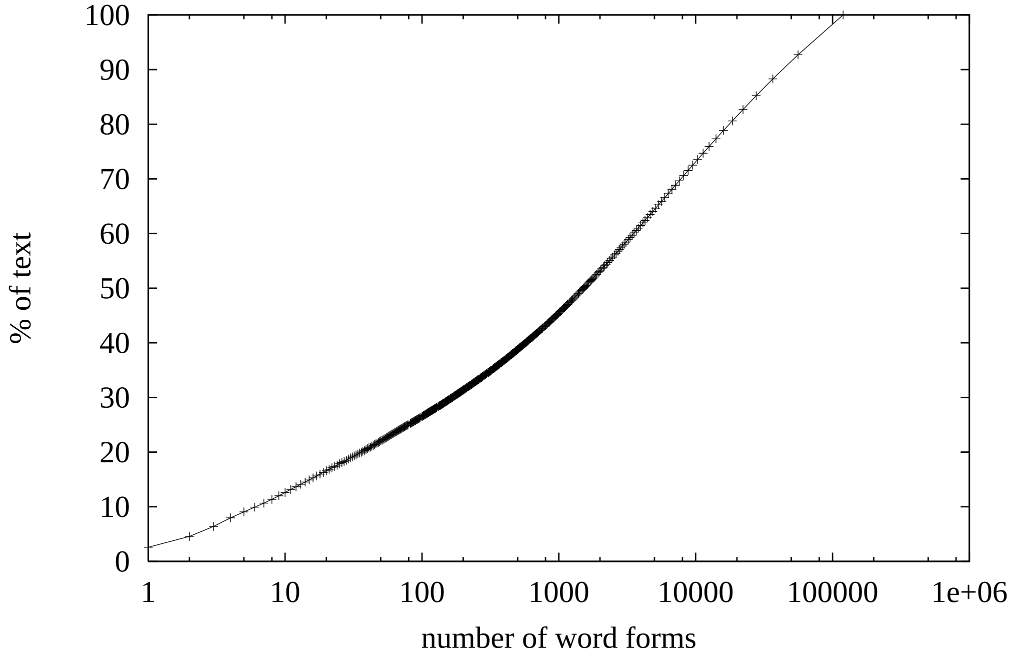


Table 4: Relation: lemmata to the length of the text



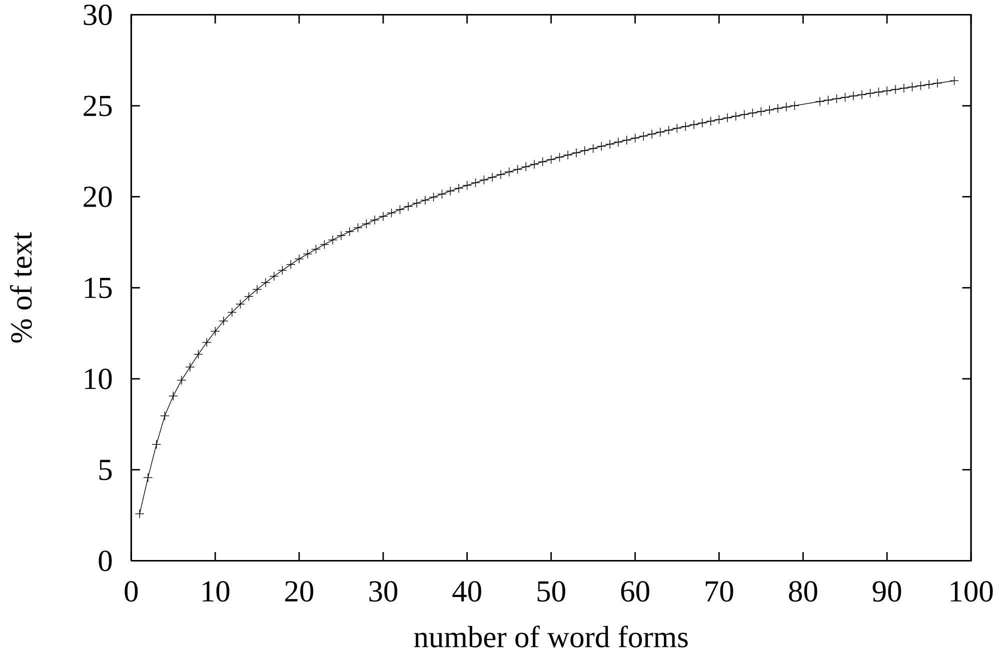


Table 5: Relation: word forms to the length of the text

kolem	k7c2	300
	k6xLeA	41
	k9xP	1
kolo	k1gNnSc7	27

Table 6: Frequences of tags for **kolem** word forms

5 Machine Learning Techniques for Automatic Desambiguation

We would like to mention some experiments we tried with machine learning methods with regard to corpus desambiguation.

5.1 Statistical methods

We have employed a simple statistical method for desambiguation. Similarly to [11], we have used the basic source channel model. The tagging procedure selects a sequence of tags T for the sentence W:

$$\Phi : W \to T. \tag{1}$$

The optimal tagging procedure maximises the product $P(W|T)P(W)$:

$$\Phi(W) = \mathrm{argmax}_T\, P(T|W) = \mathrm{argmax}_T\, P(W|T)P(T). \tag{2}$$

The basic methods of trigrams and maximum likelihood are employed to estimate the probabilities.

The results of our experiments are summarised in Table 7. The results are comparable to those published in [11] taking into consideration slightly different conditions. As stated in [11] the trigram tag prediction model needs much more data in order to get better results.

Training data	470 052 word forms
Test data	403 103 word forms
Tagging accuracy	81.64%

Table 7: Final results of probabilistic desambiguation

These results are graphically displayed in Table 8.

5.2 Connectionist methods

Recently, to avoid problems with the simple probabilistic approach described above we have started to experiment with neural networks for desambiguation. The architecture of the networks used in the experiments is similar to that

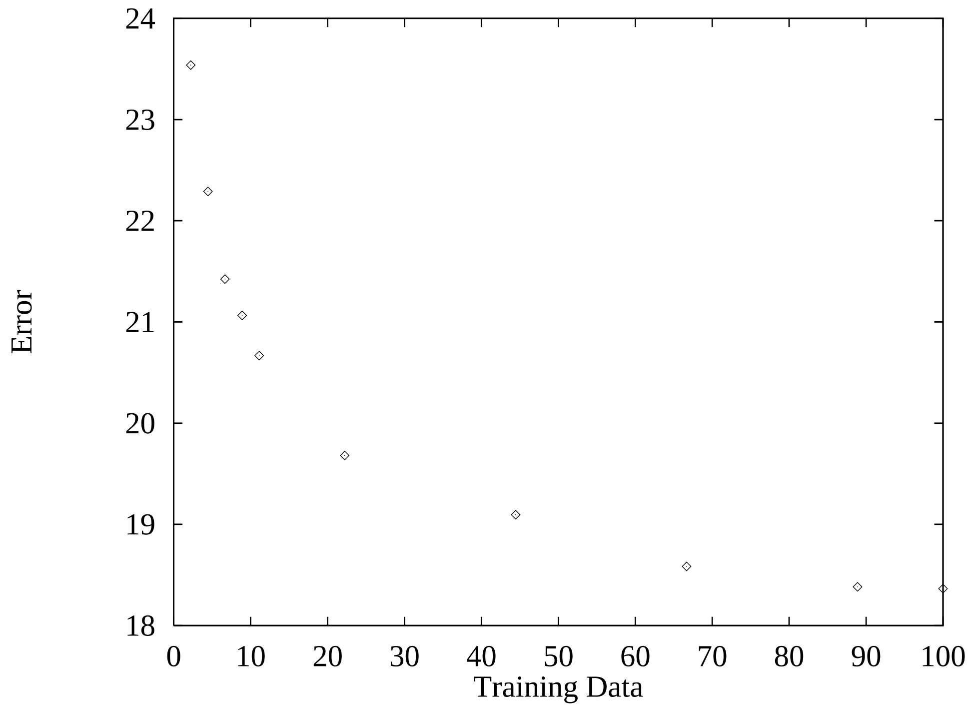


Table 8: Probabilistic desambiguation in graphical form

of NETtalk [12]. The input of the network is a series of tag sets of seven consecutive words from one of the training sentences. The central tag set in the sequence is the current one for which the output is to be produced. Three sets of possible tags on either side of this central position provide context that helps to choose the correct tag for the central word. Sentences are moved through the window so that each word with possible tags in the sentence is seen. Blanks are added before and after the sentence as needed.

One type of trained networks uses unary encoding. For each of the seven word positions in the input, the network has a set of $1665 + 1$ input units: one for each of 1097 different tags and one for blank. Thus, there are 1665 x 7 = 11655 input units. Other tested type uses compressed version of input according to the inner representation of tags in the program LEMMA. Output is coded in the same way as input, the networks have one set of neurons for the actual position only.

Training data	53,324 word forms
Test data	43,239 word forms
Tagging accuracy	75.47 %

Table 9: Final results of NN desambiguation

The neural network was able to desambiguate successfully in 75.47% of cases on a small test data. We can seriously expect that with larger data the results will improve remarkably and we hope to obtain better results than the probabilistic approach can yield. The biggest problem despite the lack of training data is the extremely long time needed for training the neural network even if the supercomputer Silicon Graphics POWER Challenge L is used. The time necessary for training takes several days.

6 Conclusions

The most important result consists in building the fully annotated and desambiguated Czech corpus DESAM at FI MU which now contains approximately 1 mil. Czech word forms (tokens). The whole process of its building took approximately 10 man-months. At the present moment DESAM runs under IMS Workbench (CQP and xkwic) and is accessible for all people who are interested in corpus applications within NLP. DESAM is already serving as a training corpus in two different ways:

- as indicated above – when using statistical approaches to desambiguation,
- for building rule-based parsing algorithms for Czech. The first results in this respect can be found in [7] and they have already been used in designing a desambiaguating programme processing Czech noun groups in raw text. It has been implemented in PROLOG – thus its name is DES.PL and it has already been used in the process of desambiguation of the second and larger part of the corpus DESAM with the fairly good results,
- now DESAM will be exploited in the second cycle: DES.PL is going to be improved so that it will contain the rules capturing verb groups, adjective and adverbial groups in Czech. In this way we should be able to build a partial parser for Czech which would be used as a semiautomatic desambiguating tool in a fashion similar to the constraint grammars [4].

References

[1] K. Pala. Desambiguating syntactic constructions from tagged corpus. In *Workshop on AI Methods in Machine Learning*, 1996.

[2] R. Garside. *The CLAWS word-tagging system, The computational analysis of English.* Longman, London, 1987.

[3] D. Cutting. A practical part-of-speech tagger. In *Proceedings of the 3rd Conference on Natural Language Processing*, Trento, Italy, March–April 1992.

[4] F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila. *Constraint Grammars*. Mouton de Gruyter, Berlin, 1995.

[5] P. Ševeček. *LEMMA – a lemmatizer for Czech*. Brno, 1996. (manuscript).

[6] K. Osolsobě. *Algorithmic description of Czech morphology*. PhD thesis, Masaryk University, Brno, 1996.

[7] V. Puža. Syntactic analysis of natural language with a view to a corpora tagging. Master's thesis, Faculty of Informatics, Masaryk University, Brno, 1997.

[8] B. M. Schulze and O. Christ. *The CQP User's Manual*.

[9] O. Christ. *The XKWIC User Manual*.

[10] J. Jelínek, J. V. Bečka, and M. Těšitelová. *Frequency Dictionary of Czech*. Academia, Praha, 1961.

[11] J. Hajič and B. Hladká. Probabilistic and rule-based tagging of an inflective language — a comparison. Technical Report 1, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, November 1996.

[12] T. J. Sejnowski and C. R. Rosenberg. Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, 1:145–168, 1987.

Publications in the FI MU Report Series are in general accessible via WWW and anonymous FTP:

`http://www.fi.muni.cz/informatics/reports/`
`ftp  ftp.fi.muni.cz (cd pub/reports)`

Copies may be also obtained by contacting:

Faculty of Informatics
Masaryk University
Botanická 68a
602 00 Brno
Czech Republic