Automatic Translation Lexicon Extraction from Czech-English Parallel Texts

Jan Cuřín, Martin Čmejrek

Abstract

We present experimental results of an automatic extraction of a Czech-English translation dictionary by means of procedures based on a probabilistic approach. We used two different bilingual corpora (computer oriented of 119,886 and fiction of 58,137 sentence pairs). For the automatic sentence alignment a statistical method was used based on sentence lengths, for the dictionary extraction a regular grammar based noun group extractor and the probabilistic model of (Brown *et al.*) were combined. The size of the resulting dictionaries is around 6,000 entries. After the significance filtering, weighted precision is 86.4% for the computer oriented Czech-English dictionary and 70.7% for the fiction part.

1 Introduction

The primary motivation for our research was to create translation lexicon of terminology of a particular discipline. Many disciplines lack relevant dictionaries or the dictionaries are obsolete because of the quick development of the discipline. The idea was that the basic part of the translation lexicon would be generated from the parallel corpus of the hitherto translated texts automatically and afterwards it could be manually edited.

The works in the field of automatic sentence alignment (Gale and Church, 1993) and automatic extraction of translation dictionary (Brown *et al.*, 1993; Wu and Xia, 1994) have exploited very large corpora. The former two used Canadian Hansards English-French Corpus, the latter one used the HKUST English-Chinese Corpus. These corpora include speeches delivered in bilingual parliaments in Canada and Hongkong. They are highly equivalent and satisfy rather strict constraints put on parallel corpora. The translations are mostly literal and tight. The situation in our country is different, we lack such a good source of large bilingual data.

Therefore, we decided to use a smaller corpus of texts taken from a particular discipline - a computer oriented corpus. This corpus consists of operating system messages from IBM AIX and of operating system guides for IBM AS/400 and VARP 4. These data are products of localization and translation of software from English into Czech. The translations are very literal and precise. In most cases sentences are translated sentence by sentence, it means that there is a one-to-one correspondence between an English sentence and a Czech sentence. But on the other hand, it is a typical feature of this kind of texts that majority of operating system messages and a big part of sentences from guides are nominal sentence, i.e. they do not have a verb.

We have also got an access to data from the Reader's Digest Výběr magazine. Every issue of this global magazine contains 30-60 % of articles that have been translated from English to the local language. We had to search in the English version to find the corresponding articles that are in the Czech version. There was also a lot of manual work to do because of the various word processor formatting. The translations in Reader's Digest are mostly very free. They include many constructions with direct speech. Some articles had to be excluded, for example such articles as "what to eat if we want to lose weight" in which the American food was substituted by Central European meals. And such substitution concerned all culture-specific facts in the process of localization. We also excluded jokes as the majority of them was translated using completely different words and strategies.

The reason why we decided to carry out the experiments also with Reader's Digest was to compare the results of methods applied to computer oriented texts on the one hand with those of journalistic texts and fiction on the other.

In section 2 we describe implementation of the statistical method for automatic paragraph and sentence alignment (Gale and Church, 1993). The important differences in the distribution of alignment categories between Canadian Hansards and Czech-English Fiction Corpora are documented.

Majority of the terms in our training data occur in the form of a noun group. The simple regular grammar based tool described in section 3 marks noun group boundaries. This is to our knowledge the first use of this method.

Section 4 describes probabilistic models 1 and 2 of (Brown *et al.*, 1993) and their implementation of the training procedure.

Output of the training procedure is filtered to produce a smaller, more useful dictionary. Section 5 is dedicated to this problem. The final evaluation and an example of the resulting dictionaries extracted from both corpora are presented.

2 Statistical Alignment of Paragraphs and Sentences in Czech-English Bilingual Corpora

For the subsequent training procedures we need to identify matching paragraphs and sentences between both languages automatically. There are two main approaches to this problem: a lexical and a statistical one, and many works use one of them or combine both (Church, Gale, 1993; Wu, Xia, 1994). Lexical approaches are based on on-line bilingual dictionaries while the statistical ones use simple probabilistic models usually based on lengths of aligned sentences. There is no on-line Czech-English dictionary available, so we have implemented the statistical model from (Gale and Church, 1993).

The problem can be formalized as follows:

Let us have Czech and English texts (typically paragraphs) T_c and T_e. The alignment is a set of pairs of parts of texts (typically $0, 1, 2 \ldots m$ sentences) $\{(L_{c,1} \rightleftharpoons L_{e,1}), \ldots, (L_{c,n} \rightleftharpoons L_{e,n})\}$ such that $L_{c,1}, \ldots, L_{c,n} = T_c$ and $L_{e,1}, \ldots, L_{e,n} = T_e$. We are looking for the best alignment $\mathcal{A}$ that maximizes the probability over all possible alignments on T_c and T_e:

$$\arg\max_{\mathcal{A}} \Pr(\mathcal{A}|T_c, T_e). \tag{1}$$

Now we have to make several approximations to obtain an effectively computable model: First, the set of possible types of matching parts of texts is limited to six categories: 1-1, 0-1, 1-0, 1-2, 2-1, 2-2. We also assume that the probabilities of individual aligned parts of texts $Pr(L_{c,i} \rightleftharpoons L_{e,i})$ are independent. And finally we assume that the probabilities of individual alignments depend only on the function δ of lengths of aligned parts $l_{\mathbf{c}}$, $l_{\mathbf{e}}$:

$$\arg\max_{\mathcal{A}} \Pr(\mathcal{A}|T_c, T_e) \approx \arg\max_{\mathcal{A}} \prod_{(L_c \rightleftharpoons L_e) \in \mathcal{A}} \Pr(L_c \rightleftharpoons L_e | \delta(l_c, l_e)). \tag{2}$$

Type of Alignment	Computer Oriented # sent.	$\Pr(L_c \bowtie L_e)$	Fiction # sent.	$\Pr(L_c \bowtie L_e)$	Canadian Hansards # sent.	$\Pr(L_c \bowtie L_e)$
1–1	109	0.90	64	0.69	1167	0.89
1–0 & 0–1	3	0.03	3	0.03	13	0.01
1–2 & 2–1	7	0.06	24	0.26	117	0.09
2–2	1	0.01	2	0.02	15	0.01
total	120	1.00	93	1.00	1312	1.00

Table 1: Sentence alignment type distribution on a hand-annotated sample.

Function δ is defined as follows:

$$\delta(l_{\mathbf{c}}, l_{\mathbf{e}}) = \frac{l_{\mathbf{e}} - l_{\mathbf{c}} e}{\sqrt{l_{\mathbf{c}} \sigma^2}}, \tag{3}$$

where $e = E(r) = E(\frac{l_{\mathbf{e}}}{l_{\mathbf{c}}})$ is the mean number of English characters generated by each Czech character, and $\sigma^2 = D^2(\frac{l_{\mathbf{e}}}{l_{\mathbf{c}}})$ is the variance. After applying Bayes' Rule and with the assumption that δ is normally distributed:

$$\begin{aligned} &\arg\max_{\mathcal{A}} \prod_{(L_c \rightleftharpoons L_e) \in \mathcal{A}} \Pr(L_c \rightleftharpoons L_e | \delta(l_c, l_e)) \approx \\ \approx &\arg\min_{\mathcal{A}} \sum_{(l_{\mathbf{c}} \rightleftharpoons l_{\mathbf{e}}) \in \mathcal{A}} - \log \frac{\Pr(\delta(l_{\mathbf{c}}, l_{\mathbf{e}}) | l_{\mathbf{c}} \rightleftharpoons l_{\mathbf{e}}) \Pr(L_{\mathbf{c}} \rightleftharpoons L_{\mathbf{e}})}{\Pr(\delta(l_{\mathbf{c}}, l_{\mathbf{e}}))} \approx \\ \approx &\arg\min_{\mathcal{A}} \sum_{(L_c \rightleftharpoons L_e) \in \mathcal{A}} - \log \left(\frac{2(1 - \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{|\sigma|} e^{\frac{-z^2}{2}} dz) \Pr(L_{\mathbf{c}} \rightleftharpoons L_{\mathbf{e}})}{\Pr(\delta(l_{\mathbf{c}}, l_{\mathbf{e}}))} \right). \end{aligned} \tag{4}$$

Parameters e, σ^2 and $\Pr(L_{\mathbf{c}} \rightleftharpoons L_{\mathbf{e}})$ are estimated from a sample of hand-aligned sentences. $\Pr(\delta(l_{\mathbf{c}}, l_{\mathbf{e}}))$ is constant and as such doesn't influence the result of minimization.

Table 1 compares parameters of three different corpora. Although Canadian Hansards and the Czech-English computer oriented corpus have very similar distribution of categories of alignment, the distribution differs substantially on the fiction corpus.

The best alignment is now easy to find if we use the dynamic programming procedure. In the first step, the algorithm aligns paragraphs in matching articles, in the second step, it aligns sentences in matching paragraphs.

Table 2 briefly summarizes results of the automatic paragraph and sentence alignment. The accuracy was 96 correctly aligned pairs on computer oriented corpora and 85 pairs on fiction corpora in a randomly selected sample of 100 pairs of sentences.

3 Groups Identification

We aim at a terminological dictionary, i.e. a dictionary containing also translations consisting of more than one word. For example, single word *typewriter* (in English) is translated by a noun group consisting in two words *psací stroj* (lit. 'writing machine') in Czech. On the other hand, the English group *construction worker* corresponds to a single Czech word *stavbař*. We model a tool which is capable of handling such cases.

	Computer Oriented	Fiction
# words (English)	1245780	959583
# words (Czech)	1089813	860757
# paragraphs (English)	88790	19567
# paragraphs (Czech)	88790	24874
# sentences (English)	120743	70872
# sentences (Czech)	121295	67856
# aligned sentences	119886	67436
types of alignment:		
1–1	117450	37039
0–1 (En/Cz)	73	5311
1–0 (En/Cz)	36	4454
1–2 (En/Cz)	1397	9501
2–1 (En/Cz)	882	7342
2–2	48	1089

Table 2: Results of automatic paragraph and sentence alignment.

The idea is to concatenate words of potential groups into one string, i.e. to consider these constructions as single "words". Identification of groups is based on a simple regular grammar. Grammar rules can be modified by the user. The grammar used in our case identifies noun groups (word sequences which consist from nouns, adjectives and some auxiliary words). Only continuous sequences of words are considered to be noun groups.

Czech is an inflective language with a lot of word forms, and in English contrastively, one word form corresponds to more POS categories. That is why we decided to proceed as follows:

- Czech groups and words are converted into their basic forms (the basic form means the first case of singular or plural for nouns, adjectives and pronouns, and the infinitive for verbs),
- English nouns and adjectives are distiguished from other POS categories,
- definite and indefinite articles are removed from English groups (there is no equivalent for the category of an article in Czech).

For an example, see Figure 1.

Tagging of Czech and English texts and conversion of Czech word forms were done by the BH tagging tools (Hajič, Hladká, 1998).

Marking of potential groups in a sentence is done separately for each language. Noun group identification algorithm works in two passes through the sentence. All possible groups in the sentence are indentified in the first pass. During the second pass the algorithm searches for such combination of groups that:

- do not overlap in one combination,
- cover the maximal number of words in the sentence,
- the number of groups in the combination is minimal.

basic form		original forms
integrovaný systém souborů	←	*integrovaný systém souborů*
	←	*integrovaného systému souborů*
	←	*integrovanému systému souborů*
	←	*integrovaný systéme souborů*
	←	*integrovaném systému souborů*
	←	*integrovaným systémem souborů*
integrated file system	←	*the integrated file system*
	←	*integrated file system*
	←	*an integrated file system*

Figure 1: Conversion of groups into their basic form.

If there is still more than one such combination, one of them is chosen randomly. Parallel sentences with marked groups are shown in Figure 2. Concatenated noun groups or nouns (one word groups) are delimited by symbols & and #.

Thanks to the fact that the learning of the dictionary is based on probabilistic methods, we have a discretion in group identification. Even if some groups are marked incorrectly, they are often eliminated by the probabilistic algorithm which handles a big amount of mostly good data.

Once the data are prepared, the translation dictionary training procedure can start.

4 Translation Dictionary Training

We implemented models 1 and 2 described in (Brown *et al.*, 1993) of sentence translation probability, and used iterative EM algorithm for maximizing the likelihood of generating the Czech translation from the English text.

Basic definitions:

Let $\mathbf{c} \equiv \mathbf{c_1}, \ldots, \mathbf{c_{l_c}}$ and $\mathbf{e} \equiv \mathbf{e_1}, \ldots, \mathbf{e_{l_e}}$ be Czech and English sentences, and the word alignment be $\mathbf{a} \equiv \mathbf{a_1}, \ldots, \mathbf{a_{l_c}}$, such that $0 \leq a_i \leq l_\mathbf{e}$, represents the information that the Czech word c_i is a translation of English word e_{a_i}. Symbols c_0 and e_0 stand for an empty word. The training corpus S is a set of pairs $(\mathbf{c}, \mathbf{e})$.

Model 1 in (Brown *et al.*, 1993) is based on word-by-word translation probability $t(c|e)$ and approximates the probability of translating the English sentence $\mathbf{e}$ into the Czech sentence $\mathbf{c}$ following a word alignment $\mathbf{a}$. This model also assumes, that all the possible word alignments are equally probable and also that all the possible lengths of Czech sentences have the same probability ϵ:

$$\Pr(\mathbf{c}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l_\mathbf{e}+1)^{l_\mathbf{c}}} \prod_{i=1}^{l_\mathbf{c}} t(c_i|e_{a_i}). \tag{5}$$

The re-estimation formulas for EM algorithm are:

$$\bar{c}(c|e; \mathbf{c}, \mathbf{e}) = \frac{t(c|e)}{t(c|e_0) + \ldots + t(c|e_{l_\mathbf{e}})} \sum_{i=1}^{l_\mathbf{c}} \sum_{j=0}^{l_\mathbf{e}} \delta(c, c_j)\delta(e, e_i), \tag{6}$$

English: The device driver indicates a hardware failure of equipment.

`&_device_driver_# indicates &_hardware_failure_of_equipment_# .`

Czech: Ovladač zařízení zjistil technickou závadu přístroje.

`&_ovladač_zařízení_# zjistit &_technická_závada_přístroje_# .`

English: Just then, they saw cowboys coach Eddie Sutton walk toward the court with a man pushing a kid in a wheelchair.

`just then they saw &_cowboys_coach_# eddie sutton walk toward &_court_# with &_man_# pushing &_kid_# in &_wheelchair_# .`

Czech: V tu chvíli zahlédli, jak na hřiště přichází jejich trenér Eddi Sutton s mužem, který na invalidním vozíku vezl malého chlapce.

`ten &_chvíle_# zahlédnout jak na &_hřiště_# přicházet jeho &_trenér_# eddi sutton s &_muž_# který na &_invalidní_vozík_# vézt &_malý_chlapec_# .`

Figure 2: Sentences with marked groups.

$$\lambda_e = \sum_c \sum_{(\mathbf{c},\mathbf{e})\in S} \bar{c}(c|e;\mathbf{c},\mathbf{e}), \tag{7}$$

$$t(c|e) = \sum_{(\mathbf{c},\mathbf{e})\in S} \frac{\bar{c}(c|e;\mathbf{c},\mathbf{e})}{\lambda_e}. \tag{8}$$

Model 2 in (Brown *et al.*, 1993) is an extension of model 1 by using word alignment probabilities $a(a_i|i, l_{\mathbf{c}}, l_{\mathbf{e}})$:

$$\Pr(\mathbf{c},\mathbf{a}|\mathbf{e}) = \epsilon \prod_{i=1}^{l_{\mathbf{c}}} t(c_i|e_{a_i}) a(a_i|i, l_{\mathbf{c}}, l_{\mathbf{e}}). \tag{9}$$

.

The re-estimation formulas for the EM algorithm are:

$$\bar{c}(c|e;\mathbf{c},\mathbf{e}) = \sum_{i=1}^{l_{\mathbf{c}}} \sum_{j=0}^{l_{\mathbf{e}}} \delta(c_i,c)\delta(e_j,e) \frac{t(c|e)a(j|i;l_{\mathbf{c}},l_{\mathbf{e}})}{\sum_{j'=0}^{l_{\mathbf{e}}} t(c_i|e_{j'})a(j'|i;l_{\mathbf{c}},l_{\mathbf{e}})}, \tag{10}$$

$$\bar{c}(i|j;l_{\mathbf{c}},l_{\mathbf{e}});\mathbf{c},\mathbf{e}) = \sum_{i=1}^{l_{\mathbf{c}}} \sum_{j=0}^{l_{\mathbf{e}}} \delta(c_i,c)\delta(e_j,e) \frac{t(c|e)a(j|i;l_{\mathbf{c}},l_{\mathbf{e}})}{\sum_{j'=0}^{l_{\mathbf{e}}} t(c_i|e_{j'})a(j'|i;l_{\mathbf{c}},l_{\mathbf{e}})}, \tag{11}$$

$$\lambda_e = \sum_c \sum_{(\mathbf{c},\mathbf{e})\in S} \bar{c}(c|e;\mathbf{c},\mathbf{e}), \tag{12}$$

$$\lambda_{i,l_{\mathbf{c}},l_{\mathbf{e}}} = \sum_{0<j<l_{\mathbf{e}}} \sum_{(\mathbf{c},\mathbf{e})\in S} \bar{c}(i|j;l_{\mathbf{c}},l_{\mathbf{e}}), \tag{13}$$

$$t(c|e) = \sum_{(\mathbf{c},\mathbf{e})\in S} \frac{\bar{c}(c|e;\mathbf{c},\mathbf{e})}{\lambda_e}, \tag{14}$$

$$t(i|j;l_{\mathbf{c}},l_{\mathbf{e}}) = \sum_{(\mathbf{c},\mathbf{e})\in S} \frac{\bar{c}(i|j;l_{\mathbf{c}},l_{\mathbf{e}})}{\lambda_{i,l_{\mathbf{c}},l_{\mathbf{e}}}}. \tag{15}$$

The EM algorithm works as follows:

1. Set the consistent initial values for $t(\cdot) > 0$ and $a(\cdot) > 0$.
2. Compute all $c(\cdot)$.
3. Compute the Lagrange multiplier λ_e for each English word e and $\lambda_{i,l_\mathbf{c},l_\mathbf{e}}$ for each position of word word alignment $\langle i, l_\mathbf{c}, l_\mathbf{e}\rangle$.
4. Re-estimate $t(\cdot)$ using equation 8 for model 1 or equations 14 and 15 for model 2.
5. Repeat steps 2–4 until $t(\cdot)$ converge.

Subcorpus of sentences of type 1–1 was selected from the computer oriented corpus. Subcorpus of 1–1, 1–2, 2–1 and 2–2 sentences was selected from the fiction corpus.

5 Significance Filtering and the Evaluation of Results

The training procedure described in the previous section results in a probabilistic dictionary which assigns translation probability to every pair of Czech and English words, which have ever been seen together in corresponding sentences. It is necessary to "clean up" the probabilistic dictionary by filtering out most of the translations to produce a useful dictionary. An obvious solution to reduce translations is to set a threshold on probabilities. Absolute thresholds work poorly, we use them only for rough pruning of translations with negligible probabilities.

The principle for significant filtering is to find a combination of just a few filtering criteria that affects the quality of the representative sample of the dictionary in the best way. This combination is used to filter the whole dictionary.

At the beginning we set the dictionary quality indicators and manually mark the quality of translations in the representative sample of the dictionary. We use two obvious quality indicators Precision and Recall, and a third one, Share, defined as follows:

Let $\mathcal{T}$ be a set of all translations in the input dictionary, $\mathcal{G}$ be a set of good translations for corresponding entries (i.e. these translations were marked as good by hand in the representative part of the dictionary), and $\mathcal{S}$ be a set of translations which were marked as good by the combination of filtering criteria (i.e. these translation were marked as good by the automatic filtering). We denote the translation probability of a particular translation (x) by $\Pr(x)$. Let $\mathcal{S}^*$ and $\mathcal{G}^*$ be similar sets including only very good translations. Good translations in contrast to very good ones are acceptable only in some context.

$$\mathsf{Precision}(\mathcal{T}) = \frac{card(\mathcal{S} \cap \mathcal{G})}{card(\mathcal{S}),} \tag{16}$$

$$\mathsf{Share}(\mathcal{T}) = \frac{\sum_{x \in \mathcal{S} \cap \mathcal{G}} \Pr(x)}{\sum_{x \in \mathcal{S}} \Pr(x),} \tag{17}$$

$$\mathsf{Recall}^*(\mathcal{T}) = \frac{card(\mathcal{S}^* \cap \mathcal{G}^*)}{card(\mathcal{G}^*).} \tag{18}$$

The quality indicator Precision grows with the elimination of bad translations. Share is a weighted Precision, which takes into account probabilities assigned to translations by the training procedure. Recall indicates a success in recognizing manually marked good translations by the automatic filtering.

The representative sample of the dictionary was about 4% of all entries.

Filtering criteria used are defined in Table 3.

criterion	description
Frst(n)	First n translations selected.
Thd(p)	Only translations accounting for the top of the threshold p are retained.
MC(n)	Works only for entries with low occurrence. Translations having count higher than n are excluded, if they have not been selected as groups. Translation probabilities for each entry are recomputed.
MPr(p)	Translations with the translation probability lower than p are excluded. Translation probabilities for each entry are recomputed.
NonT(p)	Works only for entries selected as groups. Translations with translation probabilities lower than p are excluded, if they have not been selected as groups. Translation probabilities for each entry are recomputed.

Table 3: Definiton of Filtering criteria.

Combination of Criteria	Computer Oriented R*/P/S (in %)	Fiction R*/P/S (in %)
Thd(0.85) ~ input dictionary	100.0/38.0/63.6	100.0/10.5/26.0
Frst(1)	56.6/88.1/88.1	57.3/61.7/61.7
Thd(0.7)	94.1/49.8/69.8	96.7/13.6/30.4
Thd(0.7) → MPr(0.08)	82.4/60.6/73.1	73.4/47.3/55.9
Thd(0.7) → MPr(0.05) → MPr(0.09)	81.3/62.1/73.8	81.3/38.7/49.7
Thd(0.7) → MPr(0.05) → MPr(0.09) → NonT(0.3)	81.3/68.5/78.4	80.9/41.7/54.4
MC(1800) → Thd(0.7) → MPr(0.05) → MPr(0.09) → NonT(0.3)	84.2/72.6/83.7	81.7/51.5/63.6
Thd(0.7) → MC(1800) → MPr(0.05) → MPr(0.09) → NonT(0.3)	83.8/74.8/84.9	81.3/51.4/63.9
Thd(0.7) → MC(1200) → MPr(0.05) → MPr(0.09) → NonT(0.3)	82.4/75.2/85.1	81.7/53.0/65.9

Table 4: Criteria combination and their influence on dictionary quality indicators.

	Recall*	Precision	Share
	Cz-En / En-Cz	Cz-En / En-Cz	Cz-En / En-Cz
Computer Oriented	86.8 / 83.8	75.1 / 74.8	86.4 / 84.2
Fiction	83.8 / 81.7	54.6 / 53.0	70.7 / 65.9
Computer Or. (noun groups only)	93.4 / 94.6	82.5 / 78.4	91.0 / 86.9
Fiction (noun groups only)	83.2 / 74.7	62.6 / 69.4	76.1 / 76.0

Table 5: Quality indicators in the output dictionaries.

Applying miscellaneous combinations of filtering criteria (changing the order of filtering criteria or criterion thresholds) we observe progress in dictionary quality indicators. An example can be seen in Table 4, where we show dictionary quality rates for English-Czech dictionaries (Recall*/ Precision/ Share). On the first line there are the rates for input dictionary (that is 100% for Recall*). The combination of filtering criteria with balanced values of dictionary quality indicators is chosen as optimal (the last row in Table 4). The whole dictionary is processed by this optimal combination of filtering criteria. Each output dictionary (Cz–En and En–Cz for computer and fiction) contains about 6.000 entries (see quality indicators in Table 5).

Figures 3 and 4 are examples of the filtered dictionaries. Entries and translations marked by * were recognised as groups by our noun group extracting tool. Numbers in square brackets are frequency counts in the corpus. Numbers in round brackets are translation probabilities (normalised for each entry). Good translations are underlined.

In Figure 3 there is a sample of English-Czech computer oriented lexicon. For instance entries **map** and **map*** or **mark** and **mark*** distinguish verbs and nouns translations. An example of a common error is the translation of the entry **manual IPL***, where the noun group was not recognised in the corresponding Czech sentence.

The frequency of good translations in the sample of Czech-English fiction translation lexicon in Figure 4 is lower than in the previous sample. The reason is the richness of tokens in the fiction part of corpora: for the same amount of data there were three times more tokens than in the computer oriented one. The same error of the non-recognised noun group appears for entry **rentgen*** and its translation *X-rays*. Just for fun realise that the entry **režisér*** (director) is more probable translated as *Spielberg* than to the exact translation *director*.

6 Conclusion

The reported experiments are, to our knowledge, the first demonstration of methods mentioned above for Czech and English parallel corpora. The results of automatic paragraph and sentence alignment on the computer oriented corpora reach a similar quality (96%) as those achieved on Canadian Hansards. Results on fiction corpora are worse (85%) because of the lower quality and non-literality of translations. The results of the dictionary extraction, for the computer oriented corpora are of unexpectedly high share (weighted precision) rates about 85% and for the terminology dictionary (that contains only noun groups) they are even better: 87%–91%. Soon, the results of this work will be used in practice for translation purposes.

manage [177] spravovat (0.47), řídit (0.31), správa* (0.22)
managed [21] řídit (0.36), spravovat (0.27), program (0.18), server/400 (0.18)
management* [37] management*(0.78), řízení* (0.22)
manager's maintenance operating* [10] operating (0.77), podrobnější informace* (0.23)
manager* [76] manager* (1.00)
manager software operating* [13] operating (0.34), nalézt (0.32), SC19 (0.18), program* (0.16)
manages [14] řídit (1.00)
managing [87] řízení* (0.36), správa* (0.27), spravující stroje* (0.22), spravující stroj* (0.16)
managing system* [13] řídicí systém* (1.00)
manual* [105] manual* (0.44), manuál* (0.36), příručka* (0.21)
manually [130] ručně (0.79), manuálně (0.21)
manuals* [11] příručky* (0.57), knihovna* (0.21), vyhledávání informací* (0.21)
manual installation* [13] ruční instalace* (1.00)
manual installation process* [22] proces ruční instalace* (0.44), ruční instalace* (0.41), proces* (0.15)
manual install display* [10] obrazovka manual install* (0.60), objevit (0.40)
manual IPL* [14] IPL (0.54), manuální (0.46)
manual mode* [26] režim manual* (1.00)
manufacturer* [10] výrobce* (0.83), zařízení IBM* (0.17)
many [404] mnoho (0.87), kolik (0.13)
map [12] mapovat (0.51), AS/400 (0.28), datové typy* (0.21)
map* [31] mapa* (0.68), map* (0.32)
mapped [22] mapovat (1.00)
mapping* [45] mapování* (0.45), macintosh (0.30), přiřazení* (0.25)
maps [19] mapy* (0.56), instalační (0.22), jeho (0.22)
maps* [10] mapy* (0.62), aplikace* (0.38)
margins* [13] okraje* (0.85), řádek* (0.15)
mark [19] označit (1.00)
mark* [18] označit (0.54), značka* (0.46)
marked [43] označit (0.83), označený (0.17)
marketing representative* [62] obchodní zástupce* (1.00)
marks [13] uvést (0.40), klíčové slovo* (0.40), uvozovky* (0.20)
mask* [35] maska* (0.59), maska podsítě* (0.41)
master* [13] master* (1.00)
master installation list* [50] hlavní instalační formulář* (1.00)
match [177] odpovídat (0.87), souhlasit (0.13)
match* [46] odpovídat (0.41), odpovídající protějšek* (0.31), shoda* (0.28)
matched [11] odpovídat (0.23), za (0.23), nalezený (0.18), další příkazy* (0.12), splňovat (0.12), vyhovět (0.12)
matches [56] odpovídat (0.85), souhlasit (0.15)
matching* [13] odpovídající* (0.63), odpovídat (0.37)
material* [30] materiál* (1.00)
materials* [11] materiály* (0.60), materiál* (0.40)
matrix [16] matice* (1.00)
max [41] max (0.79), maximálně (0.21)
maximum* [137] maximum* (0.52), maximálně (0.48)
maximum length* [18] maximální délka* (0.72), maximální délka parametru* (0.28)

Figure 3: Sample of English–Czech computer oriented dictionary extracted from a parallel corpus.

rachot* [11] crashing (0.61), its (0.39)
rada* [33] never (0.44), make (0.30), kids* (0.27)
raději [136] rather (0.53), better (0.29), prefer (0.18)
radio* [18] radio* (1.00)
radit [82] told (0.67), how (0.33)
radnice* [13] city hall* (0.55), day* (0.45)
radost* [67] joy* (0.74), happiness* (0.26)
radovat [24] joy* (0.53), well (0.47)
rady* [26] current article text* (0.58), advices* (0.42)
rajčata* [11] tomatoes* (0.57), tomato (0.43)
raketa* [10] rocket* (1.00)
raketoplán* [11] shuttle* (1.00)
rakety* [10] missiles* (0.59), missile* (0.41)
rakev* [10] husband* (0.54), coffin* (0.46)
rakovina* [96] cancer* (1.00)
rakovina plic* [16] lung cancer* (1.00)
rakovina prsu* [22] breast cancer* (0.74), cancer* (0.26)
ralph [41] ralph (0.81), jaymee (0.19)
ramena* [59] shoulders* (0.81), arm* (0.19)
rameno* [70] shoulder (0.83), right (0.17)
ranč* [22] ranch* (0.75), years* (0.25)
raul [17] raul (0.85), dad* (0.15)
ravussin [10] fat calories* (0.26), she's (0.26), sanchez (0.25), carbohydrates (0.23)
razit [14] quickly (0.55), wrong (0.45)
rád [345] love (0.66), loved (0.19), don't (0.15)
rádius* [21] radio* (0.71), disc* (0.29)
rámec* [11] companies* (0.55), endometrial (0.45)
rána* [53] hole* (1.00)
ráno* [57] morning* (0.69), wound* (0.16), girls* (0.15)
rány* [19] wounds* (0.53), blows* (0.47)
readers [10] german (0.34), again (0.22), american (0.22), those (0.22)
reader [20] digest (0.39), reader's (0.32), year* (0.15), nearly (0.15)
reagovat [108] respond (0.69), react (0.31)
reakce* [26] reaction* (0.71), something* (0.29)
recept* [13] recipe* (0.57), prescription* (0.43)
reeves [13] reeves (0.82), scott's (0.18)
reid [11] reid (0.84), golf balls* (0.16)
reklama* [23] advertising* (1.00)
rentgen* [10] rays* (0.50), x (0.50)
republikáni* [11] gingrich (0.60), showing (0.20), divorce* (0.20)
respektovat [10] respect* (0.53), got (0.47)
respirátor* [12] respirator* (1.00)
restaurace* [51] restaurant* (1.00)
rezervace* [16] park* (0.67), elephants* (0.33)
režisér* [10] spielberg (0.59), director* (0.41)

Figure 4: Sample of Czech–English fiction dictionary extracted from a parallel corpus.

References

Brown, Peter F.; DellaPietra, S. A.; DellaPietra, V. J.; Mercer, Robert L. (1993), "The Mathematics of Statistical Machine Translation: Parameter Estimation". In *Computational Linguistic*, 19(2): 263–331.

Cuřín, Jan. (1998), "Automatická extrakce překladu odborné terminologie." *MSc Thesis*, Institute of Formal and Applied Linguistic, Charles University, Prague. 89 pp. (in Czech)

Čmejrek, Martin. (1998), "Automatická extrakce dvojjazyčného pravděpodobnostního slovníku z paralelních textů". *MSc Thesis*, Institute of Formal and Applied Linguistic, Charles University, Prague. 82 pp. (in Czech)

Gale, Wiliam A.; Church, Kenneth W. (1993), "A Program for Aligning Sentences in Bilingual Corpora". In *Computations Liguistic*, 19(1): 75–102.

Hajič, Jan; Hladká, Barbora. (1998), "Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset". In *Proceedings of Coling/ACL'98*, Montreal, Canada.

Wu, Dekai; Xia, Xuanyin. (1994), "Learning an English–Chinese Lexicon from a Parallel Corpus". *Association for Machine Translation in the Americas*, Oct. 94: 206–213, Columbia, USA.