Annotation Guidelines for Czech-English Word Alignment

Ivana Kruijff-Korbayová[*], Klára Chvátalová[†], Oana Postolache[*]

[*] Saarland University, Saarbrücken, Germany
{korbay,oana@coli.uni-sb.de}
[†] Charles University, Prague, Czech Republic
klara.chvatalova@centrum.cz

Abstract

We report on our experience with manual alignment of Czech and English parallel corpus text. We applied existing guidelines for English and French (Melamed 1998) and augmented them to cover systematically occurring cases in our corpus. We describe the main extensions covered in our guidelines and provide examples. We evaluated both intra- and inter-annotator agreement and obtained very good results of Kappa well above 0.9 and agreement of 95% and 93%, respectively.

1. Introduction

Parallel multilingual corpora aligned at the sentence- or word-level are a valuable resource for developing machine translation systems and, recently, projecting annotations across word alignments. Our goals are in the latter group. In particular, we experiment with the projection of *information structure* on Czech-English parallel corpus, namely a portion of the Prague Czech-English Dependency Treebank version 1.0 (Čmejrek et al. 2004). We annotate information structure in Czech automatically (Postolache et al. 2005).

In order to project the annotation, we need an alignment of the tree nodes or at least of the surface words. We first created automatic word alignment of the PCEDT data by GIZA++ (Och and Ney 2000). However, an informal examination established that the quality is too low for our purposes. Therefore, we decided for manual alignment. Since there existed no guidelines for aligning Czech and English, we took the Annotation Style Guide of the Blinker Project (henceforth BASG) (Melamed 1998) as a starting point, because it has been reused in several projects dealing with word alignment.

In this paper we report on our experience with applying BASG to word alignment of Czech and English text and our extensions thereof. Overall, we found that the general rules in BASG which were originally developed for English and French can be applied for English and Czech as well. We identified a range of systematically occurring differences between the two languages, for which we felt the need to add more specific guidelines. We evaluated both intra- and inter-annotator agreement and obtained very good results of Kappa well above 0.9 and agreement of 95% and 93%, respectively.

In the rest of this paper, we first briefly describe the corpus, the annotation tool we used and the annotation process (Section 2); we overview our extensions of BASG (Section 3); we present intra- and inter-annotator agreement results (Section 4) and conclude (Section 5).

2. Manual Word Alignment on the PCEDT

Manual word alignment was performed on the text part of the Prague Czech-English Dependency Treebank 1.0 (PCEDT) (Čmejrek at al. 2004). The English sentences originate from the Wall-Street Journal part of the Penn Treebank corpus. They were translated by native speakers of Czech, who were instructed to translate sentence-by-sentence, and keep the translation both accurate and as close to the English original as possible.

We used a word alignment annotation and visualization tool implemented by Chris Callison-Burch (University of Edinburgh). The tool presents each pair of sentences as a matrix of clickable squares. Aligned word-pairs (or phrases) are represented by filled squares. The filling has two color degrees, black and grey (grey is printed white in the screen shots in Section 3 to increase their visibility), representing whether the annotator is *sure* or *unsure* of the alignment link, respectively. We encountered systematically occurring cases for which we wished to be able to distinguish between *strong* and *weak* alignment. The main ones are discussed in Section 3. In the current version of the tool we used the grey squares for weak alignment, thus overloading their semantics to encode both weak alignment and annotator's uncertainty.

The process leading to the formulation of the present guidelines involved a coordinator (the first author) supervising the project and one annotator (the second author), both native speakers of Czech proficient in English. First the coordinator annotated a trial set of 20 sentences according to BASG and sketched several additional rules for the annotator. The annotator then annotated the same trial set. The annotations were automatically compared and the differences and rules discussed. The annotator wrote the first version of the additional guidelines, and then annotated two more data sets. The annotator discussed additional guidelines with the coordinator regularly, and updated the guidelines.

The aligned data set consists of 285 sentences. This covers all files in the PCEDT development set, and some of the training set. The Czech files contain 7,706 words, the English files 7,902 (including punctuation marks).

3. Guidelines – Extensions of BASG

In our guidelines we discuss about 20 types of cases for which we extended or elaborated BASG, plus a few miscellaneous instances, and some additional examples; the guidelines contain 99 examples of alignment. They are organized similarly to BASG and we use similar headings when possible. The complete guidelines are available online: `www.coli.uni-sb.de/~korbay/alignment/`.

In section 3.1., we summarize the general principles we applied, based on BASG and augmented by introducing a distinction between strong and weak alignment. Then we discuss specific phenomena in more detail in the section 3.2 with illustrative examples in the

form of screenshots from the annotation tool, with English transliteration glosses of the Czech versions. In section 3.3. we discuss problematic instances of alignment.

3.1. General Principles

In line with BASG, we align as much possible. We believe that aligning only word-to-word according to the lexicon would lead to loosing information about the means to express the same meaning across the two languages. We thus adopt the general BASG rule saying that words should only remain unaligned "when you can answer 'Yes' to the following question: If the seemingly extraneous words were simply deleted from their verse, would the two verses become more similar in meaning?" (Melamed, 1998: 3).

However, we also encountered many cases where words are extraneous based of word-to-word lexical alignment, but deleting them would corrupt the correctness of the sentence. We generally align extraneous synsemantic words to the head noun. These cases are due to systematic differences between Czech and English, including articles and other determiners, case markers, zero subject and nominal substitution means and those cases of reflexive pronouns which have no correspondence in English.

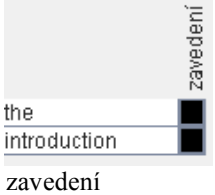
As for autosemantic words, we identified and described four specific cases where additional words mostly in Czech are required, including a difference in noun attribute types, additional common nouns, names of persons and a conjunction of a subordinate clause expressed in the other language by non-finite verbal form. On the one hand, these words are extraneous from a strictly semantic viewpoint, but on the other hand, the patterns are encountered systematically. We treat the additional words as weakly aligned (annotated by grey squares, which are show white below). Distinguishing strong and weak alignment allows us to capture more information about correspondences.

To prevent unnecessarily unaligned words, we apply the BASG rules for conjunctive non-parallelism and for resumptive pronouns. To prevent unnecessary phrase alignment, we also apply the BASG rules for auxiliary verbs and passivization. In correspondence with BASG, we align as phrases only idioms, composed prepositional constructions and miscellaneous instances of phrases with the same meaning in the given context.

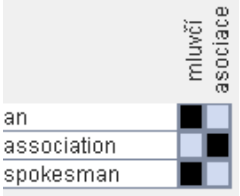
3.2. Specific Phenomena Discussed

Included among the cases discussed in our guidelines are the following phenomena:

Articles and Determiners English uses articles whereas Czech does not. Typically, we align English definite and indefinite articles with the corresponding head noun in Czech, e.g., 'the' in (1) and 'an' in (2).

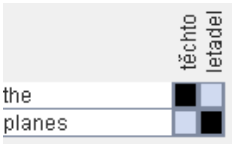


(1) zavedení
introduction

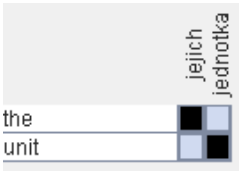


(2) mluvčí asociace
spokesman association

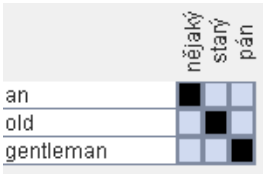
However, there are also instances where the English article corresponds to a demonstrative (ex. 3), possessive (ex. 4) or indefinite (ex. 5) pronoun in Czech, or to the Czech numeral "one" with indefinite meaning (ex. 6). Then we align the corresponding words one to another.



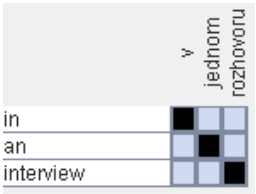
(3) těchto letadel
these planes



(4) jejich jednotka
their unit

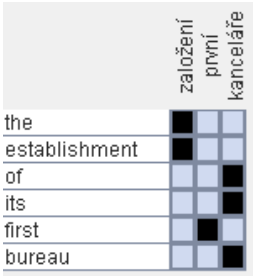


(5) nějaký starý pán
some old gentleman

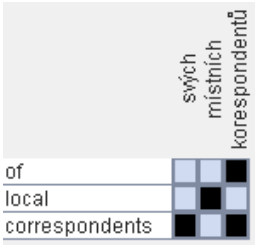


(6) v jednom rozhovoru
in one interview

Czech and English also differ in the use of possessive pronouns as determiners. We align possessive determiners present only in one language to the head of the corresponding noun phrase, e.g. 'its' in (7) and 'svých' (their: reflexive possesive) in (8).

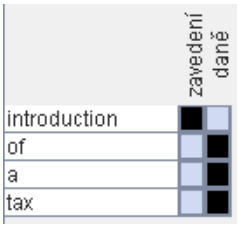


(7) založení první
establishment first
kanceláře
bureau

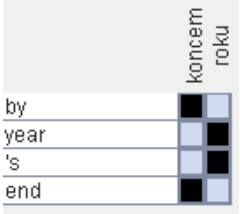


(8) svých místních
their local
korespondentů
correspondents

Case marking English often uses prepositions or possessive markers where Czech inflects the head noun of a phrase. We therefore align the former with the head noun, e.g., 'of' in (9) and 'by' in (10). Dependent adjectives are also inflected in Czech, but this is the result of adjective-noun agreement within the Czech NP, and thus not a reason for alignment. The English case markers are aligned to the adjectives only in those cases where the adjective stands in for the head noun, such as 'of' in (14).

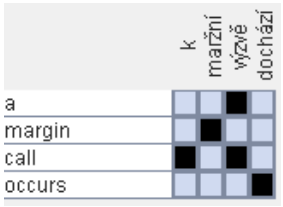


(9) zavedení daně
introduction tax



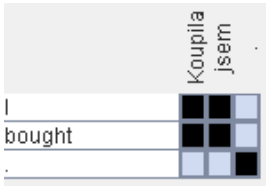
(10) koncem roku
end year

A Czech verb may require a prepositional case if there is no corresponding preposition in the English sentence. We align these additional prepositions in Czech to the head noun as well, e.g., 'k' in (11).

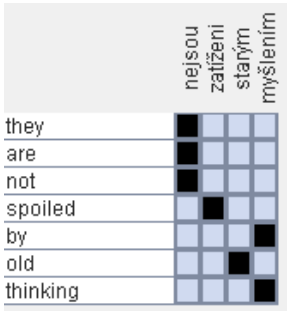


(11) k maržní výzvě dochází
to margin call comes

Zero subject Czech is a subject pro-drop language. We align the subject in English with the corresponding main and auxiliary verbs in Czech, because they also carry the agreement features (ex. (12)). However, we do not align the English subject to a predicative noun, predicative adjective or passive participle although they may also reflect number or gender , e.g., 'spoiled' in (13).



(12) Koupila jsem.
Bought be-aux.



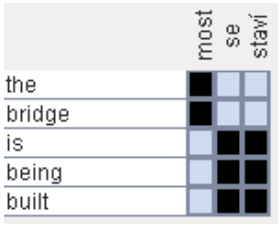
(13) nejsou zatíženi
are-not spoiled
starým myšlením
old thinking

Nominal substitution If nominal substitution with "one / ones" is used in English and an adjective as a NP head is used in Czech, we align both the English adjective and "one / ones" with the Czech adjective (ex. (14)).

(14) dvou nejdůležitějších
two most-important

Czech reflexive pronouns "se" and "si" We distinguish five different functions of Czech reflexive pronouns "se" and "si" with specific rules for alignment (Daneš, 1987: 175). In general, if there is no corresponding word in the English sentence, we align the reflexive pronoun with the English main verb (ex. (15)). The only exception is the reflexive form of passive, where we align the entire passive forms as phrases (ex. (16)).

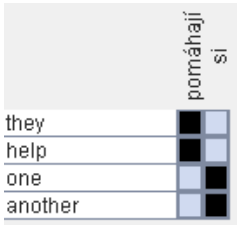
(15) se snaží
refl tries



(16) most se staví
bridge refl builds

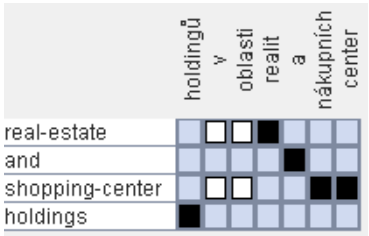
Sometimes there is a word corresponding to the Czech reflexive pronoun in the English sentence, such as a reflexive pronoun, "one another" or "each other"; then we align these words (ex. (17) and (18)). We use analogous rules for deverbal adjectives.

(17) přejmenovat se
rename itself



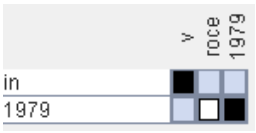
(18) pomáhají si
help one-another

Different types of attributes Attributes are often expressed differently in the two languages, particularly when English uses nominal premodifiers of nouns. Czech then often uses a construction with a noncongruent attribute, which involves an additional head noun. We treat such additional nouns as weakly aligned with the corresponding head noun –shown as white squares in the figure in example (19).

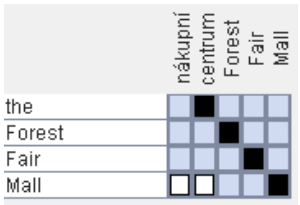


(19) holdingů v oblasti realit a nákupních center
holdings in area real-estate and shopping centres

Additional common nouns Czech often uses a common noun in addition to a proper noun, such as the name of an institution, company, date expression, etc. in order to avoid inflecting them. We treat the common noun as weakly aligned with the corresponding head noun – shown as white squares in the figure in example (20). However, the additional common noun is the head noun in this case and we align relevant articles, case markers etc. to this common noun with black squares representing strong alignment (ex. (21)).

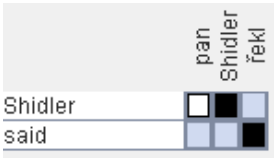


(20) v roce 1979
in year 1979

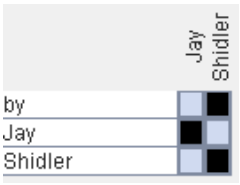


(21) nákupní centrum FFM
shopping center FFM

Names of persons In Czech a common noun often accompanies the name of a person to avoid its inflection. It is an attribute not in syntactic agreement (Grepl and Karlík, 1998: 326) and therefore we treat the proper noun, typically a surname, as a head noun in such case, e.g., 'Shindler' in (22) and (23). This is also supported by the fact that the surname seems to be the least often omitted part of the name of a person (at least in newspaper text).



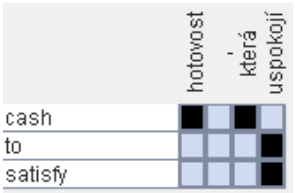
(22) pan Shidler řekl
Mister Shidler said



(23) Jay Shidler
Jay Shidler

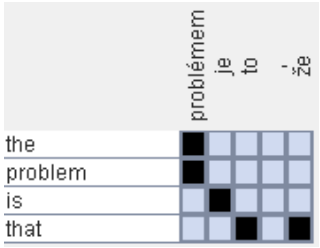
Various types of subordinate clauses We describe five types of alignment concerning subordinate clauses:

1. Relative clauses: For relative pronouns we apply the BASG rule for aligning resumptive pronouns with the respective head noun, e.g., 'která' (which) in (24).



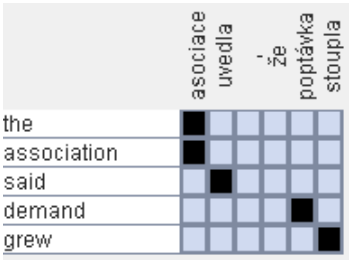
(24) hotovost, která uspokojí
cash which satisfies

2. Complex subordinating conjunctions formed by a referring word and a conjunction: We always align both parts of the correlative conjunction of this type with the English conjunction, e.g., 'to že' (that that) in (25).



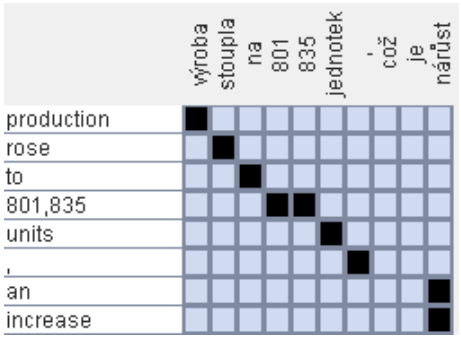
(25) problém je to, že…
problem is that, that…

3. Subordinate clauses following verbs of communication: The conjunction is very often omitted in English. The conjunction in Czech thus remains unaligned, e.g., 'že' (that) in (26).



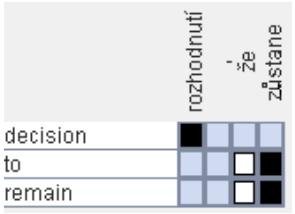
(26) asociace uvedla, že poptávka stoupla
association said that demand grew

4. When the main verb of the subordinate clause has no correspondent in the other language, we follow the meaning of the words. Non-corresponding words remain unaligned, e.g., 'což je' (which is) in (27).



(27) výroba stoupla na 801 835 jednotek, což je
production rose to 801,835 units, which is
nárůst
increase

5. Main verb in one language corresponds to a non-finite verb form in the other: We treat the whole non-finite form as strongly aligned to the main verb of the subordinate clause –black squares, and as weakly aligned to the conjunction –white squares in the figure in example (28).

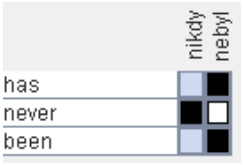


(28) rozhodnutí, že zůstane
decision that remains

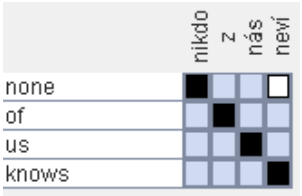
3.2.1. Problematic cases

In this section, we describe problematic cases where more than one alignment approach seems appealing, depending on whether one focuses on semantic vs. lexical/syntactic correspondence. The decision may also depend on the further use of the aligned data.

Negation involving pronouns Unlike English, Czech employs negative congruence: it uses a negative verb form and a negative pronoun. The pronouns and the verb forms are straightforward to align. However, this results in aligning a negative verb form in Czech with a positive one in English. In order to explicitly encode the involvement of all negation parts, we decided to additionally weakly align as a phrase all words reflecting the negation –white squares in the figures in examples (29) and (30).

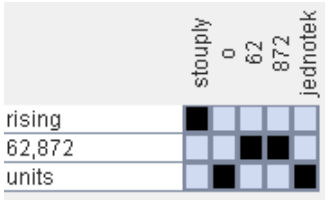


(29) nikdy nebyl
never was-not



(30) nikdo z nás neví
none of us knows-not

Expressions of quantity In Czech, certain quantitative phrases involve a systematic mismatch between semantic meaning and syntactic form: The expression of quantity is the syntactic head of the quantitative phrase, whereas the noun stating the quantified entity is an attribute not in syntactic agreement (Daneš 1987: 153; Karlík et al. 2003: 308n.). We decided to align according to the semantic meaning, which is constant cross-linguistically and is easy to decide also for non-inflected nouns (ex. (31)).



(31) stouply o 62 872 jednotek
raised by 62,872 units

Non-finite verbal forms We already mentioned subordinate clauses corresponding to non-finite verbal forms above. We also encountered non-finite verb forms corresponding to each other. However, there have not yet been enough examples to deduce a general guideline for such cases. We thus rely on the annotators' intuitions to deduce parallels with cases solved in guidelines.

4. Evaluation

4.1. Automatic vs. Manual

To compare the automatic and manual alignment, we computed the Alignment Error Rate (AER) (Och and Ney 2000) for GIZA++ against the annotator: The average AER is 0.348 with a standard deviation of 0.071. As we discuss below in more detail, we also measured the Kappa statistic (Carletta 1996). Here we obtained Kappa values below 0.6 and thus below the threshold of 0.67 for tentative conclusions (Krippendorff 1980).

4.2. Intra- and Inter- annotator agreement

In order to measure the reliability of our annotation guidelines, we measured intra- and inter- annotator agreement. We used two files with 23 and 27 sentences, respectively (all together 1,416 Czech words and 1,367 English words). These 2 test files were re-annotated by the annotator after an intermittent period of several months. The same two file were also annotated by a second annotator, who had not been involved in the project earlier. The second annotator was instructed as follows: he read BASG and our extensions, and annotated the same trial set of 20 sentences that we had used initially (cf. Section 2). Discrepancies and cases where he had doubts were discussed once. Then he annotated the two test files without further interaction. The values we obtained for intra- and inter-annotator agreement are shown in Table 1.

To compute agreement, we considered as instances all Czech-English word-pairs. If a pair was aligned (the corresponding square was colored, with either black or grey) we consider it to have the category True; if it was not aligned it has the category False. We used two measures described below.

The first measure takes into account only the True (aligned) instances. For each two annotations we computed A_1 and A_2 the number of True instances for the annotation 1 and annotation 2, respectively. Then we have computed the intersection I between the two annotations. The final agreement was computed as

$$AGR = 2*I / (A_1 + A_2)$$

The intra-annotator agreement was about 95%, inter-annotator agreement about 93%, both very high.

For comparison, the inter-annotator agreement published for the Blinker Project (Melamed and Marcus, 1988) had an overall average of 81.87. They used as a metric a weighted F-measure considering only the True instances. The weights were introduced in order to avoid placing undue importance to the words that were linked more than once.

As a second measure we used the Kappa statistic (Carletta 1996). We considered all True and False instances and computed Kappa by the two well known methods based on (Cohen 1960) and (Siegel and Castellan 1988). However, in our case the difference between the two values is negligible (in the order of 10^{-8}). The intra-annotator agreement reached Kappa of about 0.95, and inter-annotator agreement above 0.93, both well above the threshold of 0.8 for reliability (Krippendorff 1980).

One known issue for the Kappa statistic is that it does not account for varying difficulty among instances. Considering our representation of the instances as a matrix of cells, where, the True instances are usually placed along or near the diagonal, the False instances in the left-bottom and right-top corner can be considered 'easy' in the sense that they typically do not pose any difficulty when annotators make the decision (not to mark them). Because of this, taking into account all these 'easy' False instances leads to more favourable Kappa values. In order to get a feeling of how much the Kappa values are due to the agreement on the True instances, rather then to the agreement on the False instances, we also computed Kappa between each annotator and GIZA++. The resulting values below 0.6 are considerably lower than those for inter-annotator agreement (the difference is about 0.3). We thus conclude that agreement on positive instances contributed significantly to the high agreement among the annotators.

	Anot 1 vs Anot 1		Anot 1 vs Anot 2		Anot 1 vs GIZA++		Anot 2 vs GIZA++	
	AGR	Kappa	AGR	Kappa	AGR	Kappa	AGR	Kappa
File 1	95.27	.9507	94.21	.9397	61.54	.6020	60.95	.5958
File 2	94.81	.9461	93.49	.9324	55.97	.5453	55.98	.5453
Average	95.04	.9484	93.85	.9360	58.75	.5736	58.46	.5705

Table 1: Intra- and inter-annotator agreement. The AGR values are percentages.

5. Conclusions

We presented our experience from manual word alignment of a Czech-English parallel corpus. We aligned 285 sentences from the PCEDT corpus. We used existing guidelines for aligning English and French (Melamed 1998) and extended them in order to deal with a range of cases that reflect systematic differences between Czech and English. Our additional guidelines were collected gradually during the annotation, generalizing on the basis of the encountered cases. Given that the annotation of the final portion of the corpus did not require any new guidelines, we believe that the corpus size was sufficient for identifying most common phenomena. Aligning data of different genres may nevertheless still lead to additions.

We evaluated both intra- and inter-annotator reliability, and obtained very good results of Kappa well above 0.9 and agreement of 95% and 93%, respectively. This compares favourably to another annotation effort without such explicit guidelines as ours, which has resulted in an error rate of 18% (Bojar, p.c.).

We have experienced the need to make a distinction between strong and weak alignment, in order to adequately represent certain systematically occurring cases of cross-lingual correspondence. Typically, this involves one part which fits the concept of word-to-word semantic equivalence, and another part where the relationship is weaker, e.g., added words. Leaving the weakly equivalent part unaligned means information loss, but annotating such cases as phrase alignment also means losing the information about the strongly equivalent parts. Therefore, we propose to include a labelling of strong vs. weak alignment besides the already commonly used labelling of sure vs. unsure alignment.

When evaluating reliability using the Kappa statistic, one faces the problem that the statistic may be skewed because there are relatively many more False instances than Positive ones. One might consider a non-uniform distribution of the expected agreement values depending on the position of word pairs in the alignment matrix: the closer a pair of words is to the diagonal, the higher the expected alignment likelihood, and vice versa. It remains to be seen how to determine whether such Kappa calculation would give more informative results.

6. Acknowledgments

The project was supported by the International Post-Graduate College "Language Technology and Cognitive Systems". We would also like to thank Václav Němčík for volunteering to do the second alignment.

7. References

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2):249–46.

J. Cohen. 1960. A coefficient of agreement for nom-inal scales. *Educational and Psychological Measurement*, 20:37–46.

M. Čmejrek, J. Cuřín, J. Havelka, J. Hajič, and V. Kuboň. 2004. Prague Czech-English Dependency Treebank: Syntactically annotated resources for machine translation. In *Proc. of LREC*, Lisabon.

F. Daneš et al. 1987. *Mluvnice češtiny 3. Skladba.* Praha, Academia.

M. Grepl and P. Karlík. 1998. *Skladba češtiny.* Olomouc, Votobia.

P. Karlík and M. Nekula and Z. Rusínová (eds.). 2003. *Příruční mluvnice češtiny.* Praha, Nakladatelství Lidové noviny.

K. Krippendorff. 1980. *Content Analysis: An introduction to its Methodology*. Sage Publications, Beverly Hills.

D. Melamed. 1998. *Annotation Style Guide for the Blinker Project. Version 1.0.4.* Philadelphia, University of Pennsylvania.

D. Melamed and M. P. Marcus. 1998. Automatic Construction of Chinese-English Translation Lexicons, IRCS Technical Report #98-28.

F.J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proc. of the 38th Annual Meeting of the ACL*, pages 440–447, Hong Kong, China. Sidney Siegel and John N. Castellan. 1988. Nonparametric statistics for the behavioral sciences. McGraw Hill, Boston.

O. Postolache, I. Kruijff-Korbayová, and G.-J. Kruijff. 2005. Data-driven approaches for information structure identification. In *Proc. of the EMNLP'05 Conference*, Oct. 6–8 2005, Vancouver.