A Structural Similarity Measure

Petr Homola and **Vladislav Kuboň**
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
110 00 Praha 1, Czech republic
{homola,vk}@ufal.mff.cuni.cz

Abstract

This paper outlines a measure of language similarity based on structural similarity of surface syntactic dependency trees. Unlike the more traditional string-based measures, this measure tries to reflect "deeper" correspondences among languages. The development of this measure has been inspired by the experience from MT of syntactically similar languages. This experience shows that the lexical similarity is less important than syntactic similarity. This claim is supported by a number of examples illustrating the problems which may arise when a measure of language similarity relies too much on a simple similarity of texts in different languages.

1 Introduction

Although the similarity of natural languages is in principal a very vague notion, the linguistic literature seems to be full of claims classifying two natural languages as being more or less similar. These claims are in some cases a result of a detailed comparative examination of lexical and/or syntactic properties of languages under question, in some cases they are based on a very subjective opinion of the author, in many other cases they reflect the application of some mathematical formula on textual data (a very nice example of such mathematical approach can be found at (Scannell, 2004)).

Especially in the last case the notion of language similarity is very often confused with the notion of text similarity. Even the well known paper (Lebart and Rajman, 2000) deals more with the text similarity than language similarity. This general trend is quite understandable, the mathematical methods for measuring text similarity are of a prominent importance especially for information retrieval and similar fields. On the other hand, they concentrate too much on the surface similarity of word forms and thus may not reflect the similarity of languages properly. This paper tries to advocate different approach, based on the experience gained in MT experiments with closely related (and similar) languages, where it is possible to "measure" the similarity indirectly by a complexity of modules we have to use in order to achieve a reasonable translation quality. This experience led us to formulating an evaluation measure trying to capture not only textual, but also syntactic similarities between natural languages.

2 Imperfections of measures based on string similarity

There are many application areas in the NLP in which it is useful to apply the measures exploiting the similarity of word forms (strings). They serve very well for example for tasks like spellchecking (where the choice of the best candidates for correction of a spelling error is typically based upon the Levenshtein metrics) or estimating the similarity of a new source sentence to those stored in the translation memory of a Machine Aided Translation system. They are a bit controversial in a "proper" machine translation, where the popular BLEU score (Papineni et al., 2002), although widely accepted as a measure of translation accuracy, seems to favor stochastic approaches based on

Proceedings of the Workshop on Linguistic Distances, pages 91–99,
Sydney, July 2006.

an n-gram model over other MT methods (see the results in (Nist, 2001)).

The controversies the BLEU score seems to provoke arise due to the fact that the evaluation of MT systems can be, in general, performed from two different viewpoints. The first one is that of a developer of such a system, who needs to get a reliable feedback in the process of development and debugging of the system. The primary interest of such a person is the grammar or dictionary coverage and system performance and he needs a cheap, fast and simple evaluation method in order to allow frequent routine tests indicating the improvements of the system during the development of the system.

The second viewpoint is that of a user, who is primarily concerned with the capability of the system to provide fast and reliable translation requiring as few post-editing efforts as possible. The simplicity, speed and low costs are not of such importance here. If the evaluation is performed only once, in the moment when the system is considered to be ready, the evaluation method may even be relatively complicated, expensive and slow. A good example of such a complex measure is the FEMTI framework (Framework for the Evaluation of Machine Translation). The most complete description of the FEMTI framework can be found in (Hovy et al., 2002). Such measures are much more popular among translators than among language engineers and MT systems developers.

If we aim at measuring the similarity of languages or language distances, our point of view should be much more similar to that of a human translator than of a system developer, if we'll stick to our MT analogy. When looking for clues concerning the desirable properties of a language similarity (or distance) measure, we can first try to formulate the reasons why we consider the simple string-based (or word-form-based) measures inadequate.

If we take into account a number of languages existing in the world, the number of word forms existing in each of those languages and a simple fact that a huge percentage of those word forms is not longer than five or six characters, it is quite clear that there is a huge number of overlapping word forms which have completely different meaning in all languages containing that particular word form. Let us take for illustration some language pairs of non-related languages.

For example for Czech and English (the languages very different with regard both to the lexicon and syntax) we can find several examples of overlapping word forms. The English word *house* means a *duckling* in Czech, the English indefinite article *a* is in Czech also very frequent, because it represents a coordinating conjunction *and*, while *an* is an archaic form of a pronoun in Czech. On the other hand, if we look at the identical (or nearly identical) word forms in similar languages, we can find many examples of totally different meaning. For example, the word form *život* means *life* in Czech and *belly* in Russian; *godina* means *year* in Serbo-Croatian while *hodina* is an hour in Czech (by the way, an hour in Russian is *čas* — and the same word means *time* in Czech).

The overlapping word forms between relatively distant languages are so frequent that it is even possible to create (more or less) syntactically correct sentences in one language containing only word forms from the other language. Again, let us look at the Czech-English language pair. The English sentences *Let my pal to pile a lumpy paste on a metal pan.* or *I had to let a house to a nosy patron.* consist entirely of word forms existing also in Czech, while the Czech sentence *Adept demise metal hole pod led.* — [A resignation candidate was throwing sticks under the ice.] consists of English word forms.

Creating such a Czech sentence is more complicated — as a highly inflected language it uses a wide variety of endings, which make it more difficult to create a syntactically correct sentence from word forms of a language which has incomparably smaller repertoire of endings. This fact directly leads to another argument against the string similarity based measures — even though two languages may have very similar syntactic properties and their basic word forms may also be very similar, then if the languages are highly inflective and the only difference between those languages are different endings used for expressing identical morphosyntactic properties, the string similarity based methods will probably show a substan-

tial difference between these languages.

This is highly probable especially for shorter words — the words with a basic form only four or five characters long may have endings longer or equal to the length of the basic form, for example: ***nová***/***nov****ata* "new" (Cze/Mac), ***vid****ěný*/***vid****imyj* "seen" (Cze/Rus), ***fotograf****ující*/***fotograf****uojantysis* "photographing" (Cze/Lit).

The last but not least indirect argument against the use of string-based metrics can be found in (Kuboň and Bémová, 1990). The paper describes so called transducing dictionary, a set of rules designed for a direct transcription of a certain category of source language words into a target language. The system has been tested on two language pairs (English-to-Czech and Czech-to-Russian) and although there was a natural original assumption that such a system will cover substantially more expressions when applied to a pair of related languages (which are not only related, but also quite similar), this assumption turned to be wrong. The system covered almost identical set of words for both language pairs — namely the words with Greek or Latin origin. The similarity of coverage even allowed to build an English-to-Russian transducing dictionary using Czech as a pivot language with a negligible loss of the coverage.

3 Experience from MT of similar languages

The Machine Translation field is a good testing ground for any theory concerning the similarity of natural languages. The systems dealing with related languages usually achieve higher translation quality than the systems aiming at the translation of more distant language pairs — the average MT quality for a given system and a given language pair might therefore also serve as some kind of a very rough metrics of similarity of languages concerned.

Let us demonstrate this idea using an example of a multilingual MT system described in several recently published papers (see e.g. (Hajič et al., 2003) or (Homola and Kuboň, 2004)). The system aims at the translation from a single source language (Czech) into multiple more or less similar target languages, namely into Slovak, Polish, Lithuanian, Lower Sorbian and Macedonian.

The system is very simple — it doesn't contain any full-fledged parser, neither rule based, nor stochastic one. It relies on the syntactic similarity of the source and target languages. It is transfer-based with the transfer being performed as soon as possible, depending on the similarity of both languages. In its simplest form (Czech to Slovak translation) the system consists of the following modules:

1. Morphological analysis of the source language (Czech)
2. Morphological disambiguation of the source language text by means of a stochastic tagger
3. Transfer exploiting the domain-related bilingual glossaries and a general (domain independent) bilingual dictionary
4. Morphological synthesis of the target language

The lower degree of similarity between Czech and the remaining target languages led to an inclusion of a shallow parsing module for Czech for some of the language pairs. This module directly follows the morphological disambiguation of Czech.

The evaluation results presented in (Homola and Kuboň, 2004) indicate that even though Czech and Lithuanian are much less similar at the lexical and morphological level (e.g. at both levels actually dealing with strings), the translation quality is very similar due to the syntactic similarity between all languages concerned.

4 Typology of language similarity

The experience from the field of MT of closely related languages presented in the previus sections shows that it is very useful to classify the language similarity into several categories:

- typological
- morphological
- syntactic
- lexical

Let us now look at these categories from the point of view of machine translation,

4.1 Typological similarity

The first type of similarity is probably the most important one. If both the target and the source language are of a different language type, it is more difficult to obtain good MT quality. The notions like word order, the existence or non-existence of articles, different temporal system and several other properties have direct consequences for the translation quality. Let us take Czech and Lithuanian as an example of the language pair, which doesn't belong to the same group of languages (Czech is a Slavic and Lithuanian Baltic language). Both languages have rich inflection and very high degree of word order freedom, thus it is not necessary to change the word order at the constituent level. On the other hand, both languages differ a lot in the lexics and morphology.

For example, both (1) and (3) mean approximately "*The father read a/the book*". What these sentences differ in is the information structure. (1) should be translated as "*The father read a book*", whereas (3) means in fact "*The book has been read by the father*".[1] The category of voice differs in both sentences because of strict word order in English, although in both Czech equivalents, active voice is used.[2] We see that in the Lithuanian translation, the word order is exactly the same.

(1) *Otec* *četl* *knihu*
father-NOM read-3SG,PAST book-ACC
"The father read a book." (Cze)

(2) *Tėvas* *skaitė* *knygą*
father-NOM read-3SG,PAST book-ACC
"The father read a book." (Lit)

(3) *Knihu* *četl* *otec*
book-ACC read-3SG,PAST father-NOM
"The father read a book." (Cze)

(4) *Knygą* *skaitė* *tėvas*
book-ACC read-3SG,PAST father-NOM
"The father read a book." (Lit)

4.2 Lexical similarity

The lexical similarity does not mean that the vocabulary has to have the same origin, i.e., that words have to be created from the same (proto-)stem. What is important for shallow MT (and for MT in general), is the semantic correspondence (preferably one-to-one relation).

Lexical similarity is the least important one from the point of view of MT, because the lexical differences are solved in the glossaries and general dictionaries.

4.3 Syntactic similarity

Syntactic similarity is also very important especially on higher levels, in particular on the verbal level. The differences in verbal valences have negative influence on the quality of translation due to the fact that the transfer thus requires a large scale valence lexicon for both languages, which is extremely difficult to build. Syntactic structure of smaller constituents, such as nominal and prepositional phrases, is not that important, because it is possible to analyze those constituents syntactically using a shallow syntactic analysis and thus it is possible to adapt locally the syntactic structure of a target sentence.

4.4 Morphological similarity

Morphological similarity means similar structure of morphological hierarchy and paradigms such as case system, verbal system etc. In our understanding Baltic and Slavic languages (except for Bulgarian and Macedonian) have a similar case system and their verbal system is quite similar as well. Some problems are caused by synthetic forms, which have to be expressed by analytical constructions in other languages (e.g., future tense or conjunctive in Czech and Lithuanian). The differences in morphology can be relatively easily overcomed by the exploitation of full-fledged morphology of both languages (source and target).

Similar morphological systems simplify the transfer. For example, Slavonic languages (except of Bulgarian and Macedonian) have 6-7

[1] Note that in the first sentence, an indefinite article is used, whereas in the latter one, a definite article stands in front of "book". The reason is that in the first sentence, the noun "book" is not contextually bound (it belongs to the focus), in the latter one it belongs to the topic.

[2] Passive voice (except of the reflexive one) occurs rarely in Czech (and most other Slavonic languages). It can be used if one would like to underline the direct object or if there is no subject at all (for example, *Kniha byla čtena* "The book has been read").

cases. The case system of East Baltic languages is very similar, although it has been reduced formally in Latvian (instrumental forms are equal as dative and accusative and the function of instrumentral is expressed by the preposition *ar* "with", similarly as in Upper Sorbian). (Ambrazas, 1996) gives seven cases for Lithuanian, but there are in fact at least eight cases in Lithuanian (or ten cases but only eight of them are productive[3]). Nevertheless the case systems of Slavonic and East Baltic languages are very similar which makes the languages quite similar even across the border of different language groups.

Significant differences occur only in the verbal system, East Baltic languages have a huge amount of participles and half-participles that have no direct counterpart in Czech. The Lithuanian translation of an example from (Gamut, 1991) is given in (5):

(5) *Gimė* *vaikas,*
was-born-3SG child-NOM
valdysiantis *pasaulį*
ruling-FUT,MASC,SG,NOM world-ACC
"A child was born which will rule the world." (Lit)

The participle *valdysiantis* is used instead of an embedded sentence, because Lithuanian has future participles. These participles have to be expresses by an embedded sentence in Slavonic languages.

5 An outline of a structural similarity measure

In this section, we propose a comparatively simple measure of syntactic (structural) similarity. There are generally two levels which may serve as a basis for such a structural measure, the surface or deep syntactic level. Let us first explain the reasons supporting our choice of surface syntactic level.

Compared to deep syntactic representation, the surface syntactic trees are much more closely related to the actual surface form of a sentence. It is quite common that every word form or punctuation sign is directly related to a single node of a surface syntactic tree. The deep syntactic trees, on the other hand, usually represent autosemantic words only, they may even actually contain more nodes than there are words in the input sentence (for example, when the input sentence contains ellipsis). It is also quite clear that the deep syntactic trees are much more closely related to the meaning of the sentence than its original surface form, therefore they may hide certain differences between the languages concerned, it is a generally accepted hypothesis that transfer performed on the deep syntactic level is easier than the transfer at the surface syntactic level, especially for syntactically and typologically less similar languages.

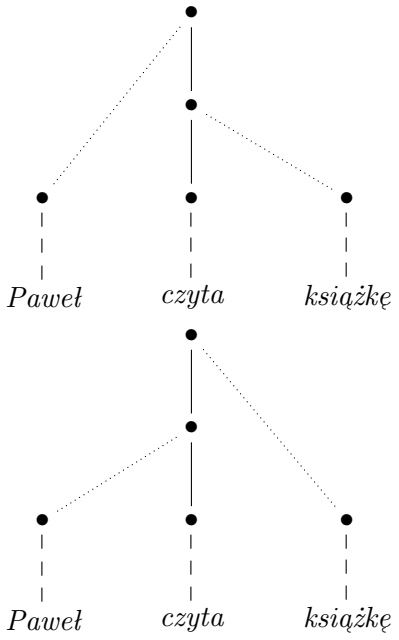
The second important decision we had to make was to select the best type of surface syntactic trees between the dependency and phrase structure trees. For practical reasons we have decided to use dependency trees. The main motivation for this decision is the enormous structural ambiguity of phrase structure trees that represent sentences with identical surface form. Let us have a look at the following Polish sentence:

(6) *Paweł* *czyta*
Paweł-NOM read-3SG
książkę
book-FEM,SG,ACC
"Paweł is reading a/the book."

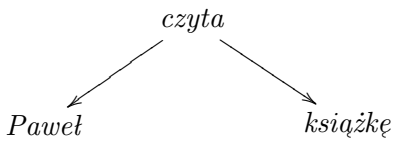
The syntactic structure of this sentence can be expressed by two phrase structure trees representing different order of attaching nominal phrases to a verb.[4]

[3] Although some Balticists argue that illative forms are adverbs, it is a fact that this case is productive and used quite often (Erika Rimkutė, personal communication), though it has been widely replaced by prepositional phrases. Allative and adessive are used only in some Lithuanian dialects, except of a few fixed allative forms (e.g., *vakarop(i)* "in the evening", *velniop(i)* "to the hell".)

[4] The full line denotes the head of the phrase, the dotted line a dependent.



There is no linguistically relevant difference between these two trees. Although generally useful, the information hidden in both trees is purely superfluous for our goal of designing a simple structural metrics. The dependency tree obtained from the phrase structure ones by contraction of all head edges seem to be much more appropriate for our purpose. In our example, we therefore get the following form of the dependency tree:



The nodes of the dependency trees representing surface syntactic level directly correspond to word forms present in the sentence. For the sake of simplicity, the punctuation marks are not represented in our trees. They would probably cause a lot of technical problems and might distort the whole similarity measure. The node of a tree are ordered and reflect the surface word-order of the sentence. Different labels of nodes in both languages (see the example below) don't influence the value of the measure, however they are important for the identification of corresponding nodes (a bilingual dictionary is used here).

The structural measure we are suggesting is based on the analogy to the Levenshtein measure. It is therefore pretty simple — the distance of two trees is the minimal amount of elementary operations that transform one tree to the other. We consider the following elementary operations:

1. adding a node,
2. removing a node,
3. changing the order of a node,
4. changing the father of a node.

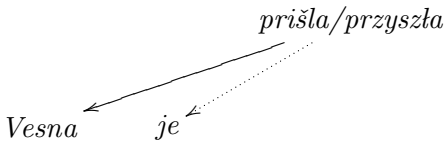
The similarity of languages can be obtained as an average distance of individual sentences in a parallel corpus.

The following examples show the use of the measure on individual trees. The correspondence between individual nodes of both trees can be handled by exploiting the bilingual dictionary wherever necessary:

(7) *Vesna* *je*
Vesna-NOM is-3SG
prišla
come-RESPART,FEM,SG
"Vesna has come." (Slo)

(8) *Vesna* *przyszła*
Vesna-NOM come-RESPART,FEM,SG
"Vesna has come." (Pol)

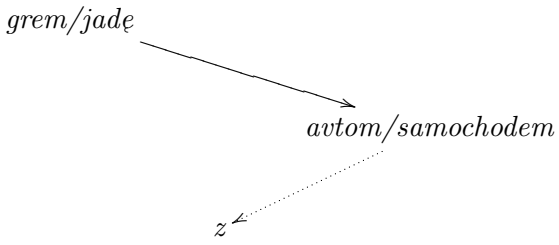
The distance between (7) and (8) is equal 1, since one node has been removed (the dotted line gives the removed node).



(9) *Grem* *z* *avtom*
go-1SG with car-MASC,SG,INS
"I am going by car." (Slo)

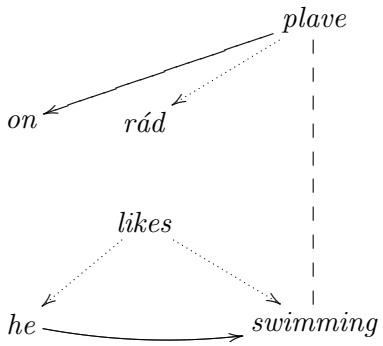
(10) *Jadę* *samochodem*
go-1SG car-MASC,SG,INS
"I am going by car." (Pol)

The distance between (9) and (10) is equal 1, since one node has been removed (the dotted line gives the removed node).



5.1 Formalization

(11) *On* *rád* *plave*
he-NOM with-pleasure swims-3SG
"He likes swimming." (Cze)



The Czech-English example (11) shows two sentences which have a mutual distance equal to 3 — if we start changing the Czech tree into an English one, then the first elementary operation is the deletion of the node *rád*, the second operation adds the new node corresponding to the English word *likes* and the third and last operation is the change of the father of the node corresponding to the personal pronoun *on* [he] from *swimming* to *likes*. As mentioned above, the node labels are not taken into account, the fact that the Czech finite verbal form *plave* changes into an English gerund has no effect on the distance.

A similar case are sentences with a dative agent, for example:

(12) *Je* *mi* *zima*
is me-DAT cold-F,SG,NOM
"I am cold" (Cze)

In this sentence, the Czech *mi* does not match to *I* since it is no subject. Similarly, the substantive *zima* does not match to *cold*, since it is a different part of speech. Hence two nodes are removed and two new nodes are added, which gives us a distance of 4. This example demonstrates that the measure tends to behave naturally - even short sentences containing syntactically different constructions get a relatively high score.

To formalize the process described above, let us introduce a notion of lexical and analytical equality of nodes in analytical trees:

- Two nodes equal lexically if and only if they share the same meaning in the given context. Nevertheless to simplify automatic processing, we treat two nodes as lexically equal if they share a particular meaning (defined e.g. as a non-empty intersection of Wordnet classes).
- Two nodes equal analytically if and only if they have the same analytical label (e.g. subject, spacial adverbial etc.).

As for the measure, two nodes match to each other if they 1) occur at the same position in the subtree of their parent and 2) equal lexically and analytically.

If a subtree (greater than 1) is added or removed, the operation contributes to the measure with the size of the subtree (the amount of its nodes), for example in the following idiomatic phrase:

(13) *puścić* *z* *dymem*
leave-INF with smoke-MASC,SG,INS
"burn down" (Pol)

(14) *zapálit*
burn-down-INF
"burn down" (Cze)

In the above example, the distance is equal 2.

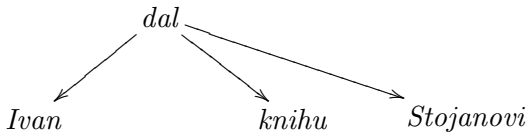
The automatic procedure can be described as follows (given two trees):

1. Align all sons of the root node.
2. Count discrepancies.
3. For all matched nodes, go to step 1 to process subtrees and sum up distances.

5.2 Discussion

It is obvious that our measure expresses the typological similarity of languages. We get comparatively high values even for genetically related languages if their typology is different. Let us demonstrate this fact on Czech and Macedonian examples.

(15) *Ivan* *dal*
Ivan-NOM gave-RESPART,MASC,SG
knihu *Stojanovi*
book-FEM.SG,ACC Stojan-DAT
"Ivan gave the book to Stojan." (Cze)



(16) *Ivan* *mu* *ja*
Ivan-NOM him her-FEM,SG,ACC
ima *dadeno*
has-3SG given-PPART,NEUT,SG
knigata *na* *Stojan*
book-FEM.SG,DEF on Stojan
"Ivan gave the book to Stojan." (Mac)

The distance equals 5. The score is relatively high, taken into account that both languages are related. It indicates again that for a given purpose the measure seems to provide consistent results.

The proposed measure takes into account only the structure of the trees, completely ignoring node and edge labels. Let us analyze the following example:

(17) *Ta* *książka*
this-FEM,SG,NOM book-FEM.SG,NOM
się *często* *czyta*
REFL well read-3SG
"This book is read often."

(18) *Tę* *książkę*
this-FEM,SG,ACC book-FEM.SG,ACC
się *często* *czyta*
REFL well read-3SG
"This book is read often."

The syntactic trees of both sentences have the same structure, but (17) is passive and (18) active (with a general subject). This is of course a significant difference and as such it should be captured in the measure, nevertheless our simple measure doesn't reflect it. There are several reasons why a current version of the measure doesn't include morphological and morphosyntactic labels. One of the reasons is a different nature of the problem — to design a reliable measure combining structural information with the information contained in node labels is very difficult. From the technical point of view, a great obstacle is also the variety of systems of tags used for this purpose for individual languages, which may not be compatible. For example, Macedonian has almost no cases at nouns, therefore it would make no sense to use cases in the noun annotation, while for other Slavic languages (and not only for Slavic ones) is this information very important. To find a good integration of morphosyntactic features into the structural measure is definitely a very interesting topic for future research.

6 Conclusions

This paper contains an outline of a simple language similarity measure based upon the surface syntactic dependency trees. According to our opinion, such a measure expresses more adequately the similarity of languages than simple string-based measures used for the text similarity. The measure is defined on pairs of trees from a parallel corpus. In its current form it doesn't account for differences in morphosyntactic labels of corresponding nodes or edges, although it is an important parameter of language similarity. The proper combination of our basic structural similarity measure with some measure reflecting the differences of labels opens a wide range of options for a future research. Equally important seems to be a task of gathering properly syntactically annotated parallel corpora of a reasonable size. The only corpus of such kind which we have at our disposal, the Prague Czech-English Dependency Treebank (Cuřín et al., 2004) relies on imperfect automatic annotation which might distort the results. The human annotation of the PCEDT is just starting, so there's a

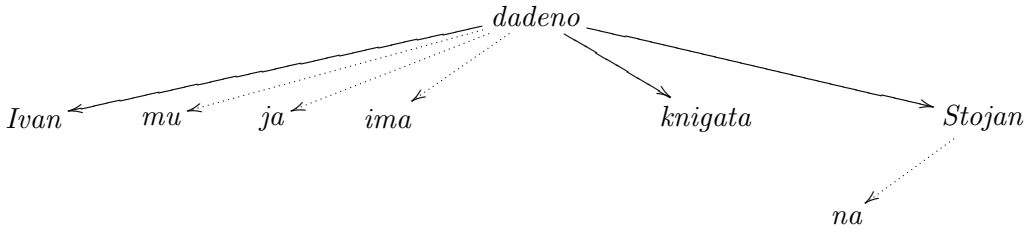


Figure 1: The dependency tree of (16)

good chance that the measure will bring some reliable results at least for those two lenguages soon.

7 Acknowledgements

This research was supported by the Ministry of Education of the Czech Republic, project MSM0021620838, by the grant No. GAUK 351/2005 and by the grant No. 1ET100300517. We would like to thank the anonymous reviewers for their valuable comments and recommendations.

References

Vytautas Ambrazas. 1996. *Dabartinės lietuvių kalbos gramatika*. Mokslo ir enciklopedijų leidykla, Vilnius.

Jan Cuřín, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. 2004. Prague Czech-English Dependency Treebank Version 1.0. Linguistic Data Consortium.

LTF Gamut. 1991. *Login, loanguage and meaning 2: Intensional logic and logical grammar*. University of Chicago Press, Chicago.

Jan Hajič, Petr Homola, and Vladislav Kuboň. 2003. A simple multilinguale machine translation system. In *Proceedings of the MT Summit IX*, New Orleans.

Petr Homola and Vladislav Kuboň. 2004. A translation model for languages of accessing countries. In *Proceedings of the 9th EAMT Workshop*, La Valetta, Malta.

Eduard Hovy, Margaret King, and Andrei Popescu-Beli. 2002. Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 1(17).

Vladislav Kuboň and Alevtina Bémová. 1990. Czech-to-Russian Transducing Dictionary. In *Proceedings of the XIIIth conference COLING '90*, volume 3.

Ludovic Lebart and Martin Rajman, 2000. *Handbook of Natural Language Processing*, chapter Computing similarity. Dekker, New York.

Nist. 2001. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Technical report, NIST.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia.

Kevin P. Scannell. 2004. Corpus building for minority languages. http://borel.slu.edu/crubadan/index.html.