HunOr: A Hungarian–Russian Parallel Corpus

Martina Katalin Szabó[1], Veronika Vincze[2], István Nagy T.[3]
[1]University of Szeged, Department of Hungarian Linguistics
[2]Hungarian Academy of Sciences,
Research Group on Artificial Intelligence
[3]University of Szeged, Department of Informatics
E-mail: szabomartinakatalin@gmail.com,{vinczev, nistvan}@inf.u-szeged.hu

Abstract

In this paper, we present HunOr, the first multi-domain Hungarian–Russian parallel corpus. Some of the corpus texts have been manually aligned and split into sentences, besides, named entities also have been annotated while the other parts are automatically aligned at the sentence level and they are POS-tagged as well. The corpus contains texts from the domains literature, official language use and science, however, we would like to add texts from the news domain to the corpus. In the future, we are planning to carry out a syntactic annotation of the HunOr corpus, which will further enhance the usability of the corpus in various NLP fields such as transfer-based machine translation or cross lingual information retrieval.

Keywords: Hungarian–Russian parallel corpus, sentence level alignment, named entity recognition

1. Introduction

Parallel corpora are of primary importance in many areas of computational linguistics like machine translation, cross lingual information retrieval etc. Moreover, they can enhance research in certain fields of the humanities such as contrastive linguistics or translational studies (Klaudy, 2001; Szabómihály, 2003; Dobrovolsky et al., 2005; Horváth, 2008). Research in these academic fields can surely exploit such corpora, however, Hungarian is, unfortunately, an underresourced language as far as parallel corpora are concerned. To the best of our knowledge, two Hungarian–English parallel corpora have been already created: Hunglish (Varga et al., 2005) and SzegedParalell (Tóth et al., 2008). As for Russian, there are several parallel corpora containing Russian as one language, e.g. in the Russian National Corpus, there are English–Russian and German–Russian parallel sections (Dobrovolsky et al., 2005) and UMC 0.1 contains texts in Czech, Russian and English (Klyueva and Bojar, 2008). Among the languages used in the MULTEXT-EAST project, we can find Hungarian and Russian as well, i.e. there exists an annotated version of Orwell's *1984* for both languages. However, no digitalised Hungarian–Russian parallel corpus that contains texts from multiple domains has been made so far.

In this paper, we present HunOr[1], the first multi-domain Hungarian-Russian parallel corpus. We discuss the difficulties concerning corpus building and alignment for the given language pair, then we provide some statistical data on the corpus. We conclude with a description of applicability of the corpus and future work.

2. Composition of the HunOr corpus

The HunOr corpus currently comprises approximately 800 thousand words, but is undergoing continuous enlargement. Texts of the corpus are from various sources, for instance, printed version, electronic publication etc.

The HunOr corpus consists of three subcorpora on the basis of the text genres: literature, scientific and official language subcorpora. Nevertheless, the corpus is going to be extended with a newspaper subcorpus within a short period of time.

2.1. The literature subcorpus

The literature subcorpus currently contains five books written in Russian and their versions translated to Hungarian and five books written in Hungarian and their versions translated to Russian: Boris Akunin – Grigory Chartishvili *Kladbisenskie istorii* 'Cemetery Stories' published in 2005 (the Hungarian translation was made by Ibolya Bagi and Csaba Sarnyai); Fyodor Mikhaylovich Dostoevsky *Zapiski iz podpolya* 'Notes from Underground' published in 1864 (the Hungarian translation was made by Imre Makai); Ilya Ilf, Yevgeny Petrov *Dvenadtsat stulyev* 'The Twelve Chairs' published in 1928 (the Hungarian translation was made by Hugó Gellért); Isaak Emmanuilovich Babel *Konarmija* 'Red Cavalry' published in 1926 (the Hungarian translation was made by János Elbert and László Wessely); Nikolay Vasilyevich Gogol *Zapiski sumasshedshego* 'Diary of a Madman' published in 1835 (the Hungarian translation was made by József Czimer); Frigyes Karinthy *Tanár úr, kérem* 'Please Sir' published in 1916 (the Russian translation was made by A. Gerskovic); Ferenc Móra *Aranykoporsó* 'The Gold Coffin' published in 1933 (the Russian translation was made by V. Malihin); Géza Gárdonyi *Egri csillagok* 'Stars of Eger' published in 1899 (the Russian translation was made by A. Kun); Kálmán Mikszáth *A fekete város* 'The Black Town' published in 1911 (the Russian translation was made by G. Leybutin); Jenő Rejtő *A tizennégy karátos autó* 'The 14-carat roadster' published in 1940 (the Russian translation was made by I. Aleksandrov).

Most of the texts are from the internet but some of them were available only in a printed version, therefore had to be digitalised.

2.2. The scientific subcorpus

The scientific subcorpus consists of essays on literary

[1] The acronym consists of two parts: *Hun* (*Hungarian*) and *Or* (*Orosz* 'Russian').

works. One of the essays is the paper by Vitaly Orlov published under the title *Hranitel nenuzhnih veshey* 'The keeper of needless things' in 1999, the other one is an extract of a longer essay by Nikolay Berdyaev published under the title *O vecno-babyom v russkoy duse* 'About the „eternal femininity" in the Russian soul' in 1990. The essays were translated into Hungarian by György Zoltán Józsa and Ildikó Régéczi. Texts written in Russian are from the internet but the texts translated into Hungarian had to be digitalised.

2.3. The official language subcorpus

Texts of the official language subcorpus are from the website of the Hungarian Ministry of Foreign Affairs. An electronic publication of the Ministry, *Tények Magyarországról* 'Facts about Hungary' is translated into several languages, among others into Russian. The official subcorpus of HunOr currently consists of the following texts of the publication and their translations: *A magyar kultúra ezer esztendeje* 'One thousand years of Hungarian culture'*; Nemzeti jelképek, nemzeti ünnepek* 'National symbols, national days'*; Magyar Nobel-díjasok egy jobb világért* 'Nobel laureates from Hungary for a better world'*; Törvény a szomszédos államokban élő magyarokról: érdekek és célok* 'Act on Hungarians living in neighbouring countries: interests and goals'. Regarding the authors and the translators of the texts only the following information is at our disposal: *A magyar kultúra ezer esztendeje* was written by Béla Pomogáts, and *Magyar Nobel-díjasok egy jobb világért* by Ferenc Nagy.
Table 1 demonstrates the basic statistical data on the current version of the HunOr corpus:

Text genre	Tokens		Sentences	
	Rus	Hun	Rus	Hun
Literature	789,001	798,641	67,021	61,505
Scientific	6,683	7,228	370	348
Official	14,774	13,522	668	568
Total	810,458	819,391	68,059	62,421

Table 1: Statistical data on the HunOr corpus.

As can be seen, there are more tokens in the Hungarian part of the corpus, however, they are organized into less sentences than the Russian tokens. Still, there is no significant difference between the average length of sentences: in Russian, a sentence contains 11.9 tokens while in Hungarian, this number is 13.1. It should be noted that in general, the scientific and official texts contain longer sentences (the above rate being about 20) but due to the large size of the literature subcorpus, which consists of shorter sentences, this rate is about 12 at the corpus level.

2.4. Directions of corpus enlargement

As it was mentioned before, we would like to extend the corpus with newspaper texts and pieces of news of miscellaneous topics. On the other hand, as there are certain texts that are included in the SzegedParalell corpus or they are available in English as well, we would like to build an English–Hungarian–Russian trilingual subcorpus of HunOr, which will comprise of the following texts:

- Jenő Rejtő: *A tizennégy karátos autó*
- Béla Pomogáts: *A magyar kultúra ezer esztendeje*
- Frigyes Karinthy: *Tanár úr, kérem*

As for now, the text *A magyar kultúra ezer esztendeje* has been aligned in all the three languages, which contains about 160 sentences.
Thus, the HunOr corpus can be expanded with regard to the domain and language of the texts.

3. Alignment

Corpus texts are processed as follows: after digitalisation, we split the texts into sentences, which are then aligned, and finally we supply the corpus with morphological annotation.
First, the texts have been converted to txt format and conversion errors have been corrected manually. The texts have been split into sentences, which have been aligned and Named Entities have been annotated in some of the texts.

3.1 Manual annotation

In order to test the efficiency of automatic sentence splitters and aligners, manual annotation was carried out on a small part of corpus texts. To enhance the usability of the corpus, Named Entities are also annotated in the database. At the moment, the manually aligned texts constitute one text from each subcorpus, furthermore, the annotation of the named entities is carried out on all of the following texts: the scientific subcorpus, *A magyar kultúra ezer esztendeje* and *Kladbisenskie istorii*. Two linguists annotated the four classical NE types, i.e. PERSON, ORGANISATION, LOCATION, MISCELLANEOUS (Tjong Kim Sang and De Meulder, 2003) in the texts. Their agreement rates were 0.8695 and 0.9609 on Hungarian whereas 0.7995 and 0.9318 on Russian data (given in κ-measure and micro-averaged F-measure, respectively). The annotation also makes it possible to train and test Hungarian and Russian NER applications on HunOr.
Statistical data shown in Table 2 demonstrate that the two languages differ in the frequency of named entities. On the one hand, this might be due to interlingual differences, i.e. names of holidays, historical events and periods are written with capital letters in Russian like *Рождество* 'Christmas' or *Великая Октябрьская социалистическая революция* 'Great October Socialist Revolution', which are considered a named entity in Russian but their Hungarian equivalents, *karácsony* and *a nagy októberi szocialista forradalom* are not (Bolla et al., 1977; Laczkó and Mártonfi, 2006). On the other hand, there are stylistic differences in translation: for instance, a pronoun can stand for the proper name in the other language.

	Russian	Hungarian
Person	1704	1656
Location	732	603
Organisation	148	116
Miscellaneous	327	253
Total	2910	2628

Table 2: Statistical data on named entities.

3.2 Sentence level alignment

When aligning source and target language sentences, six types of correspondence are typically distinguished (Klaudy, 2007). Moreover, during the segmentation of the HunOr corpus we detected a 7th type of correspondence as well, listed here as g). The seven types of the translation unit are the following:

a) correspondence „1-1”: one source language sentence corresponds to one target language sentence;

Gorcsev Iván, a Rangoon teherhajó matróza még huszonegy éves sem volt, midőn elnyerte a fizikai Nobel-díjat. ‘Ivan Gorchev, sailor on the freight ship 'Rangoon', was not yet twenty-one when he won the Nobel Prize in physics.’

Иван Горчев – матрос фрахтера «Рангун» – получил Нобелевскую премию по физике, когда ему не было и двадцати одного года. ‘Ivan Gorchev – sailor of the freight ship 'Rangoon' – got the Nobel Prize in physics when he was not yet twenty-one.’

(Subcorpus: Literature, type: novel, author: Jenő Rejtő, title: A tizennégy karátos autó, date: 1940, source: Hungarian Electronic Library, internet, translator: I. Aleksandrov, title: Zolotoy avtomobil, date: 1989, source: Librusek, internet)

b) correspondence „0-1”: addition of sentence(s);

(no example in the HunOr corpus)

c) correspondence „1-0”: omission of sentence(s);

(no example in the HunOr corpus)

d) correspondence „1-N”: separation of sentences;

Впервые я почувствовал, что она жива, в ранней молодости, когда служил в тихом учреждении, расположенном неподалеку от Донского монастыря, и ходил с коллегами на древние могилки пить невкусное, но крепкое вино "Агдам". ‘I had a feeling for the first time in my early youth that she was alive, when I was serving in a quiet institute not far from the monastery of Don, and I often went with my colleagues to the ancient graves to drink unsavoury but strong wine 'Agdam'.’

Még egészen fiatal voltam, amikor először megéreztem, hogy életben van. Egy csendes intézetben dolgoztam, nem messze a Doni kolostortól, és gyakran kijártunk a kollégáimmal az ősi sírok közé Agdamot, ezt a vacak ízű, de annál erősebb bort inni. ’I was quite young when I had a feeling for the first time that she is alive. I was working in a quiet institute not far from the monastery of Don, and we often went to the ancient graves to drink 'Agdam', an awful tasting but all the stronger wine.’

(Subcorpus: Literature, type: novel, author: Boris Akunin – Grigory Chartishvili, title: Kladbisenskie istorii, date: 2005, source: Librusek, internet, translators: Ibolya Bagi, Csaba Sarnyai, title: Temetői történetek, date: 2008, source: printed version)

e) correspondence „N-1”: conjoining of sentences;

И там тяжело заболел - результат голода, обморожения, истощения. Когда рукопись была перепечатана и готовилась к отправке в Москву, кто-то опять донес на Домбровского. ’And there he became heavy ill in the consequence of starvation, frostbite and exhaustion). When the manuscript was typed and ready for dispatch to Moscow, somebody reported Dombrovsky again.’

Ott pedig súlyos betegség tört rá (az éhezés, az elfagyások, a legyengülés következménye), s mire a kézirat szép rendben legépelve csak arra várt, hogy Moszkvába küldjék, valaki ismét csak feljelentést tett Dombrovszkij ellen. ’And there he became seriously ill (in the consequence of starvation, frostbites and weakening) and by the time the manuscript was finely typed and was waiting only to be sent to Moscow, somebody reported Dombrovsky again.’

(Subcorpus: Scientific, type: essay, author: Vitaly Orlov, title: Hranitel nenuzhnih veshey, date: 1999, source: Vestnik, internet, translator: György Zoltán Józsa, title: A szükségtelen tárgyak részlegének őrzője, date: 2009, source: Mária Fonalka (ed). Visszavonások könyve, printed version)

f) correspondence „N-M”: shifting sentence borders;

Ha megfigyeltük eddig hősünket, egy különös tulajdonságát ismerhettük fel: sohasem mondott igazat, de nem is hazudott. Csak éppen habozás nélkül kimondott mindent, ami eszébe jutott, és ez sok, elképesztő bonyodalomba sodorta életében. Egyik szavától a másikig, egyik tettétől a következőkig ritkán vezetett valamiféle okszerűség. ’If we have observed our hero, then we could have noted a peculiar attribute of him: he has never told the truth, but he has never lied either. It was just that he said without hesitation, everything that came to his mind and this habit plunged him into many astounding situations. From one of his words to another, from one of his actions to another has rarely been a kind of rationality.’

Как вы успели, вероятно, заметить, наш герой отличался замечательным качеством: он не говорил правды и не лгал, а просто и порывисто излагал все, что приходило в голову. Такое свойство уже не раз вовлекало его в невообразимые истории, поскольку довольно редко наблюдалась логическая связь между его словами или поступками. ’As for sure you managed to observe, our hero differed in a remarkable attribute: he has never told the truth and he has never lied, but he simply and abruptly reported everything that came to his mind. This attribute has already plunged him into incredible stories many times, since logical connection could be quite rarely observed between his words or his actions.’

(Subcorpus: Literature, type: novel, author: Jenő Rejtő, title: A tizennégy karátos autó, date: 1940, source: Hungarian Electronic Library, internet, translator: I. Aleksandrov, title: Zolotoy avtomobil, date: 1989, source: Librusek, internet)

g) correspondence „N=M": transposition of the order of the sentences.

Лемносский бог тебя сковал
Для рук бессмертной Немезиды,
Свободы тайный страж, карающий кинжал,
Последний судия Позора и Обиды.

Это стихотворение Домбровский очень любил.

'God of Lemnos hammered you / To the hands of the immortal Nemesis, / Secret guardian of the freedom, retributiving dagger, / Supreme judge of scandal and injury.

Dombrovsky liked this poem very much.'

Dombrovszkij ezt a verset igen szerette.

Kit vulkán edzett jó előre
S a Nemezis kezébe tett:
A bosszú kése vagy szabadság titkos őre,
Bírák bírája bűn és jogtiprás felett!

'Dombrovsky liked this poem very much.

Who was hardened by a volcano in advance / And was taken into the hands of Nemesis: / The knife of the retribution or the secret guardian of the freedom, / Supreme judge of guilt and injustice.'

(Subcorpus: Scientific, type: essay, author: Vitaly Orlov, title: Hranitel nenuzhnih veshey, date: 1999, source: Vestnik, internet, translator: György Zoltán Józsa, title: A szükségtelen tárgyak részlegének őrzője, date: 2009, source: Fonalka Mária (ed.) Visszavonások könyve, printed version)

3.3 Sentence Splitting

As a part of our corpus has manually annotated sentence boundaries, we could test several sentence splitter tools. We evaluated five different tools (Dragon (Zhou et al., 2007), magyarlanc (Zsibrita et al., 2009), LingPipe (Alias-I, 2008), MorphAdorner (Kumar, 2009) and Stanford (Toutanova and Manning, 2000)) with ten different models on the Hungarian part of the corpus. Unfortunately these approaches could not work on Russian texts. Therefore, we evaluated the Punkt sentence splitter (Kiss and Strunk, 2006) from the NLTK toolkit (Bird et al., 2009) with their Russian model.
magyarlanc**,** which tool was designed for Hungarian, achieved the best results on the Hungarian part of corpus (96.39/97.78/97.08 in terms of recall, precision and F-score), but the average results of the different devices are not much worse than the best (94.92/94.55/94.61). The result for Russian was 97.99/63.76/77.25. Thus, the differences between the two language sentence splitting results shows that it is not a trivial task to adapt existing tools to another character set (in this case, Cyrillic). In addition, fewer tools are available for Russian, which also caused that we could not experiment with more splitters.

3.4 Alignment by Using Named Entities as Anchors

Named entities are successfully applied as anchors in the automatic synchronisation of texts written in different languages since algorithms rely efficiently on language elements identical with each other (Tóth et al., 2008). However, during the creation of the corpus, we encountered several difficulties. First of all, translators totally transform the named entities of the source language in many cases, for instance, they substitute proper nouns with common nouns or they omit them (Vermes, 2005). In other cases, translators substitute common nouns of the source language with proper nouns in the target language (or substituting a personal pronoun with a proper name). These operations in translation limit the applicability of the named entities as anchors in automatic synchronisation.
Moreover, the character sets of the two languages are not the same for Hungarian uses Latinate characters whereas Russian uses Cyrillic characters. This results in the fact that finding anchors in texts is not trivial.
Another complication is that foreign proper nouns are not literally transcribed into Russian but according to their pronunciation (to some extent). The following examples from the HunOr corpus demonstrate this peculiarity:

New York Times (Eng.) → *Нью-Йорк Таймс* [Nyu York Tayms]
Francois de la Chaise (Fr.) → *Франсуа де ла Шез* [Fransua de la Shez]

Bilingual lists of proper names and a NER system may help to identify the other language equivalents of the given named entity.
Besides these specific transliteration rules, the forms of named entities might also differ due to inflection as well. Hungarian lemmas typically do not change when suffixes are added to them (Törkenczy, 2005), for instance, adding a dative suffix to names ending in a consonant typically does not change the lemma: *Gábor* → *Gábort* ('Gábor' → 'for Gábor'). However, there are many exceptional cases. In Hungarian, words ending in *a, e, o* or *ö* become lengthened before most suffixes, which is true for Named Entities as well, for instance: *Anna* → *Annával* ('Anna' → 'with Anna'). In the case of multiword Named Entities, it is only the last member that gets inflected, the other members remain unchanged: *Magyar Köztársaság* →*Magyar Köztársasággal* ('Republic of Hungary' → 'with the Republic of Hungary').
With respect to the inflectional behaviour of the named entities, the Russian language (Rozental' and Telenkova, 1984; Pehlivanova, 1989; Beloshapkova, 1997) shows similar characteristics to Hungarian language. In many cases Russian lemmas do not change when suffixes are added to them. For instance, in dative case, if a Russian male first name ends in a consonant, we generally add *y* to the stem: *Владимир* → *Владимиру* ('Vladimir' → 'for Vladimir'). However, some of the first names have a

second stem which is derived from the main stem by deletion of the final vowel. For instance, in dative case the final vowel of the female first names ending in *а* or *я* is replaced with *е*: *Анна → Анне* ('Anna' → 'for Anna'). In addition, the second stem of some Russian nouns is derived from the main stem by elision of the final vowel preceding the stem-final consonant, which is true for some of the named entities as well, for instance: *Павел → Павлу* ('Pavel' → 'for Pavel'). However, in contrast to Hungarian language, if the Russian named entity consists of more than one element, each element of the named entity has to be inflected in most of the cases, for instance: *Российская Федерация → в Российской Федерации* ('Russian Federation'→ 'in the Russian Federation').
In such cases, the assumption that the first character n-grams (case suffixes are disregarded now) are required to match might prove useful in automatic alignment. We would like to experiment with alignment techniques based on named entity recognition as future work.

4. Morphological analysis

In order to enhance the usability of the corpus, texts were automatically POS-tagged. For Russian, we used the TreeTagger morphological analyzer and POS-tagger (Schmid, 1994, 1995) with the tagset of Sharoff et al. (2008) and for Hungarian, we applied the toolkit magyarlanc (Zsibrita et al. 2010). The trilingual part of the corpus was also POS-tagged: for the English texts, the Stanford POS-tagger was utilized (Toutanova and Manning, 2000).
Statistical data on the frequency of parts-of-speech in the subcorpora can be seen in Table 3.

POS	Russian	Hungarian
Noun	185,930	157,379
Verb	122,917	111,040
Adjective	51,781	61,052
Article	--	79,925
Adverb	44,214	87,875
Numeral	4,591	9,641
Pronoun	83,755	42,086
Conjunction	55,311	54,590
Pre/postposition	67,731	8,536
Punctuation	183,487	168,895
Other	10,701	38,372

Table 3: Statistical data on parts of speech.

5. Conclusions and future work

In this paper, we have presented HunOr, the first multi-domain Hungarian-Russian parallel corpus. Some of the corpus texts have been manually aligned and split into sentences, besides, named entities also have been annotated. The other parts of the corpus are automatically split and aligned and the entire corpus is automatically POS-tagged. The current version of the corpus consists of approximately 800,000 tokens and 60,000 sentence alignment units from the domains literature, official language use and science, however, we would like to add texts from the news domain to the corpus. Furthermore, we would like to add the English version of texts to the corpus – wherever available – in order to create a trilingual subcorpus. The corpus is freely available at http://www.inf.u-szeged.hu/rgai/corpus_hunor.
In the future, we are planning to add syntactic annotation to the HunOr corpus. In this way, the parallel corpus will certainly prove useful in the development of Hungarian-Russian transfer-based machine translation systems. In addition, applications in the field of cross language information retrieval can also profit from the database. Moreover, as a consequence of the several layers of linguistic annotation (named entities, morphology, syntax) the HunOr corpus will be a powerful help for various linguistic fields such as translational studies or mono- or bilingual corpus-based syntactic research.

6. Acknowledgements

This work was supported in part by the National Innovation Office of the Hungarian government within the framework of the project MASZEKER.

7. References

Alias-i. (2008). *LingPipe 4.1.0*. http://alias-i.com/lingpipe
Beloshapkova = Белошапкова, В.А. (1997). *Современный русский язык*. 3-е изд., Москва: Азбуковник
Bird, S., Loper, E., Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.
Bolla et al. = Болла, К., Палл, Э., Папп, Ф. (1977). *Курс современного русского языка*. Budapest: Tankönyvkiadó.
Dobrovolsky et al. = Добровольский, Д.О., Кретов, А.А., Шаров, С.А. (2005). Корпус параллельных текстов. Архитектура и возможности использования. In *Национальный корпус русского языка: 2003–2005*. Москва: Индрик. pp. 263–296.
Horváth, P. I. (2008). Személynevek a szakfordításban [Person names in domain translation]. *Névtani Értesítő* 30, pp. 35–40.
Kiss, T., Strunk, J. (2006). Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics* 32:4, pp. 485–525.
Klaudy, K. (2001). Mit tehet a fordítástudomány a magyar nyelv „korszerűsítéséért"? [What can translational studies do for the „modernization" of the Hungarian language?] *Magyar Nyelvőr* 125, pp. 145–52.
Klaudy, K. (2007). *Languages in Translation. Lectures on the theory, teaching and practice of translation. With illustrations in English, French, German, Russian and Hungarian*. Budapest: Scholastica.
Klyueva, N., Bojar, O. (2008). UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proceedings of International Conference Corpus Linguistics,* pp. 188–195.
Kumar, A. (2009). *MONK Project: Architecture Overview*. Technical Report of the Northwestern University.
Laczkó K., Mártonfi A. (2006). *Helyesírás [Orthography]*. Budapest: Osiris.
Pehlivanova = Пехливанова, К.И., Лебедева, М.Н. (1989). *Грамматика русского языка в иллюстрациях (для иностранцев, изучающих русский язык)*. Москва: Русский язык
Rozental' and Telenkova = Розенталь, Д.Э., Теленкова, М.А. (1984). *Словарь трудностей русского языка*. 3-е изд., Москва: Русский язык
Schmid, H. (1994). Probabilistic Part-of-Speech Tagging

Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A. and Divjak, D. (2008). Designing and Evaluating a Russian Tagset. In *Proceedings of LREC 2008*, pp. 279−285.

Szabómihály, G. (2003). A szlovákiai magyar szakfordítások minőségének javításáról és az objektív fordításkritika megteremtésének feltételeiről [On the development of the quality of Slovakian Hungarian domain translations and the conditions of creating an objective critique of translation]. *Fórum Társadalomtudományi Szemle* 4, pp. 55−68.

Tjong Kim Sang, E., De Meulder, F. (2003) Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*.

Tóth, K., Farkas, R., Kocsor, A. (2008). Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora. *Acta Cybernetica* 18(3), pp. 463–478.

Toutanova, K., Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of EMNLP 2000*, pp. 63–70.

Törkenczy, M. (2002). *Practical Hungarian Grammar*. Budapest: Corvina.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*. pp. 590–596.

Vermes, A. P. (2005). *Proper names in translation: A relevance-theoretic analysis*. Kossuth Egyetemi Kiadó: Debrecen.

Zhou, X., Zhang, X., Hu, X. (2007). Dragon Toolkit: Incorporating Auto-learned Semantic Knowledge into Large-Scale Text Retrieval and Mining. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*.

Zsibrita, J., Nagy, I., Farkas, R. (2009). Magyar nyelvi elemző modulok az UIMA keretrendszerhez [Hungarian language processing modules for the UIMA framework]. In *VI. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 394-395.

Zsibrita, J., Vincze, V., Farkas, R. (2010). Ismeretlen kifejezések és a szófaji egyértelműsítés [Unknown expressions and POS-tagging]. In *VII. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 275–283.