The Effects of Multi-Word Tagging on Text Disambiguation

Miloš Utvić[1], Ivan Obradović[2], Cvetana Krstev[1], Duško Vitas[3]
[1] University of Belgrade, Faculty of Philology
[2] University of Belgrade, Faculty of Mining and Geology
[3] University of Belgrade, Faculty of Mathematics
E-mail: misko@matf.bg.ac.rs, ivano@rgf.bg.ac.rs, cvetana@matf.bg.ac.rs, vitas@matf.bg.ac.rs

Abstract

This paper outlines an experiment aimed at assessing the effects that tagging of multi-words in a text has on text disambiguation. The experiment was performed on the Serbian translation of Verne's novel 'Around the world in 80 days', and consisted of two steps: in the first step we applied only resources for single words, while in the second step we included available resources for multi-word tagging. We have assessed the effects of using the multi-word resources through several measures pertaining to overall ambiguity, ambiguity of lemmas and ambiguity of grammatical categories. The results confirmed that the tagging with multi-word units reduces the ambiguity of a text, which was to be expected, but also showed that despite considerable benefit obtained in specific cases the overall reduction of ambiguity was not substantial, at least for the given example and the available resources. We further analyze the possible reasons for such results and ways of to improve them.

Keywords: text disambiguation, text tagging, multi-word units, morphological electronic dictionaries, Unitex.

1 Motivation

The ambiguity, or rather false ambiguity, of a text processed by morphological electronic dictionary is high for Serbian, as for many other languages. Unfortunately, with the enlargement of e-dictionaries the rate of ambiguity rises as well. The automatic elimination of false ambiguities without eliminating the correct grammatical solutions is a difficult task in Natural Language Processing. It seems that correct recognition and tagging of multi-word units could help in this direction. Some authors have observed that usage of multi-word units helps in various natural language processing tasks. Villavicencio in (Villavicencio et al., 2007) states that usage of multi-word units increases the grammar coverage in syntactic processing while Alegria in (Alegria et al., 2004) obtained a significantly higher precision of POS tagging when using multi-word units.

This point can best be illustrated by the following example. If we analyze the title of Verne's novel *Put oko sveta za 80 dana* 'Around the world in 80 days' taking into consideration simple words only, we obtain the results shown in Table 1.

Table 1. Potential grammatical tags of the phrase *Put oko sveta za 80 dana* when considering only simple word units.

Put	**oko**	**Sveta**	**Za**	**8**	**0**	**dana**
put,.ADV	oko,.N:ns1q	sveta,svet.A:aefs1g	za,.PREP+p2			dana,dan.A+PP:aefs1g
put,.N:fs1q	oko,.N:ns4q	sveta,svet.A:aefs5g	za,.PREP+p4			dana,dan.A+PP:aefs5g
put,.N:fs4q	oko,.N:ns5q	sveta,svet.A:aemw2g	za,.PREP+p6			dana,dan.A+PP:aemw2g
put,.N:ms1q	oko,.PREP+p2	sveta,svet.A:aemw4g				dana,dan.A+PP:aemw4g
put,.N:ms4q	oko,oka.N:fs5q	sveta,svet.A:aenw2g				dana,dan.A+PP:aenw2g
put,.PREP+p2		sveta,svet.A:aenw4g				dana,dan.A+PP:aenw4g
		sveta,svet.A:aenp1g				dana,dan.A+PP:aenw4g
		sveta,svet.A:aenp4g				dana,dan.A+PP:aenp1g
		sveta,svet.A:aenp5g				dana,dan.A+PP:aenp4g
		sveta,svet.A:akms2g				dana,dan.A+PP:aenp5g
		sveta,svet.A:akms4v				dana,dan.A+PP:akms2g
		sveta,svet.A:akns2g				dana,dan.A+PP:akms4v
		sveta,svet.N+Ek:mw2q				dana,dan.A+PP:akns2g
		sveta,svet.N+Ek:mw4q				dana,dan.N:mw2q
		sveta,svet.N+Ek:ms2q				dana,dan.N:mw4q
						dana,dan.N:mp2q
						dana,dan.N:ms2q
						dana,dati.V+Perf+Tr+Iref+Ref:Tfs
						dana,dati.V+Perf+Tr+Iref+Ref:Tmw
						dana,dati.V+Perf+Tr+Iref+Ref:Tnw
						dana,dati.V+Perf+Tr+Iref+Ref:Tnp
6 grammatical solutions/ 4 lemmas	5 grammatical solutions/3 lemmas	15 grammatical solutions/2 lemmas	3 lemmas	1	1	21 grammatical solutions/3 lemmas

If we analyze the same title with multi-word units as well, this analysis yields the results presented in Table 2.

Table 2. Potential grammatical tags of the phrase *Put oko sveta za 80 dana* when considering only multi-word units.

Put oko sveta	**za**	**80**	**dana**
put oko sveta,.N+Comp:s1qm	za,.PREP+p2	80,.NUM+C+v5	dana,dan.A+PP:aefs1g
put oko sveta,.N+Comp:s4qm	za,.PREP+p4		dana,dan.A+PP:aefs5g
	za,.PREP+p6		dana,dan.A+PP:aemw2g
			...
2 grammatical solution/1 lemma	3 lemmas	1 lemma	21 grammatical solutions/3 lemmas

This example shows that when only simple words were used the analysis tokenized the title into 7 tokens. The average number of lemmas per token was 17/7=2.43, while the average number of grammatical solutions per token was 52/7=7.43. The number of possible paths in the sentence graph was 28,350. The analysis that included dictionaries of compounds and dictionary graphs for multi-token and multi-word numerals tokenized the title into 4 tokens. The average number of

lemmas per token was 8/4=2, while the average number of grammatical solutions per token was 27/4=6.75. The number of possible paths in the sentence graph was 126.

When judging these results one has to bear in mind that although the average numbers in the latter case were not dramatically lower, they were computed on almost a half of the number of tokens – 4 vs. 7. On the other hand, the number of paths in the sentence graph was reduced considerably.

We could quite easily eliminate some other false ambiguities with the use of, for instance, ELAG grammars (Laporte 1999). One such grammar would distinguish from the set of 21 offered grammatical solutions for *dana* the correct one dana,dan.N:mp2q, since it follows the numeral 80 that agrees with nouns (and other part-of-speech) in the plural genitive case. However, in this experiment we will not consider the effects of such solutions to the disambiguation process.

2 The set-up of the experiment

For our experiment we have chosen the Serbian translation of Verne's novel *Le tours du monde en quatre-vingt jours* 'Around the world in 80 days'. We have decided to work with this text for two reasons. First of all, many compound words from this novel were previously identified and added to the Serbian dictionary of compounds. Namely, the work presented in (Laporte et al., 2008) served as the basis for detecting a number of compounds in Serbian and adding them in the Serbian dictionary of compounds (Vitas et al., 2008), – a total of 143 DELAC entries. Since Serbian dictionaries of compounds are still under-developed, using a novel with good compound coverage produced a more realistic environment for the experiment. Second, since Serbian version of the text is already part-of-speech and grammatically lemmatized with simple words, and partly with multi-word units as well (Tufiş et al., 2008), we plan to combine the results of this experiment with the lemmatized text for some future research.

The Serbian version of the novel processed by Unitex has 4,226 sentence delimiters, 279 digits and 58,724 simple forms, compared to the French original that has 4,458 sentences, 438 digits and 71,859 simple forms. Lexical resources for simple words used for processing the text consist of general dictionaries, dictionaries of proper names (personal and geographic names), a rudimentary dictionary of encyclopedic knowledge and a small dictionary specific to this text. This last dictionary consists mainly of proper names and derived adjectives specific to this text,

like *Paspartu* 'Passepartout'. However, this dictionary is not exhaustive so 276 simple words occurring 620 times remained unrecognized. Most of them are also proper names and derived adjectives. When the French original of the novel was processed by Unitex using French lexical resources 384 simple words remained unrecognized.

Lexical resources for multi-word units applied to the text were of various types. For recognition of numerals several dictionary transducers were applied. This way we recognized numerals written with digits only – 60.000, spelled numerals - *tri hiljade devet stotina i devedeset i devet* 'three thousand nine hundred and ninety and nine', as well as some abbreviated ordinal numbers – *21-og* '21st'. In addition to that we used our dictionary of common compounds (nouns and adjectives), as well as proper compound names.

While the French original was associated with 2,044 compound lexical entries (vs. 13,031 simple lexical entries), the Serbian text was associated with 802 compound lexical entries (vs. 43,092 simple lexical entries). Among them were 131 numerals, 23 adjectives, 35 adverbs, 19 conjunctions, 35 prepositions, and 559 nouns (82 proper names). These compound lexical entries were associated to 1080 compound text tokens. The most frequent (49 times) compound token was the conjunction *kao da* 'as if' followed by several tokens related to the compound proper name *Hong-Kong* in three different cases with a total frequency of 68 (27+22+19). The former could be considered as text independent but the latter is obviously strongly dependent on the subject of the analyzed text. For our analysis we assumed that all recognized multi-word units take precedence over simple word constituents. This does not mean that multi-word units are unambiguous since some of them (nouns and adjectives) can represent different grammatical realizations. All recognized multi-word units were manually checked, and our assumption was proven wrong in 13 cases – one noun, three adverbs and nine prepositions. One example is the sequence *na to* which can be a compound adverb 'upon that' as in *Na to udovica ode iz grada*... 'Alors la veuve quitta la ville, …' ('(Upon that) the widow left the town…'), while in some cases it represents a preposition followed by a demonstrative pronoun 'to that', like in *Na to je gospodin Fog odgovarao*... 'A quoi Mr. Fogg répondait…' ('Mister Fogg responded to that…'). However, we believe that a small number of incorrect cases will not seriously affect our results.

3 Results of the experiment

3.1 Overall assessment of ambiguity

We divided our experiment in two parts: in the first part we applied to our text only resources for single words (we will refer to these results as 'swu') while in the second part we included in processing all available resources for tagging with multi-word units (these results will be referred as 'mwu'). In the second part of the experiment ('mwu'), the text was tokenized in 13,003 different tokens (types) with a total frequency of 71,986; among these tokens the tag {S} had the frequency 4,226, while 12,546 different simple tokens (simple word forms and punctuation marks) had the frequency 66,680 (an average of 5.3 by token), while 456 different multi-word tokens had the frequency 1080 (an average of 2.4 by token).

In our experiment we were interested in word units rather than in all text tokens (such as punctuation marks). Table 3 shows that the number of word units in the text was reduced by 2.39% in the second experiment. The last two columns in the table show how many grammatical solutions and lemmas were on the average associated as potential tags to each word unit by applied lexical resources.

Table 3. Tokenization of text in simple word tokens vs. multiword tokens

	word units	gs/wu	lemma/wu
Swu	59007	3.49	1.72
Mwu	57598	3.44	1.71
Diff	1409		
Diff %	2.39%		

Let us first look at the overall ambiguity of word types in terms of lemmas assigned to them in the 'swu' case. Out of 12,475 different types (after discarding all punctuation marks, digits and unknown words), in 4040 cases (32.3%) a type was unambiguously a noun, in 2962 cases (23.7%) it was an unambiguous verb, in 1976 cases (15.8%) an unambiguous adjective, in 212 cases (1.7%) an unambiguous pronoun, and in 173 cases (1.3%) an unambiguous adverb. In all, about three quarter of text types an unambiguous lemma is assigned – although it does not mean that grammatical tags are unambiguous as well.

We will now look at the overall results obtained by applying multi-word lexical resources, summarized in Table 4 for different POS categories. The 'swu' columns give the number of possible grammatical solutions and candidate lemmas for each specific POS without the application of multi-word resources, and the 'mwu' columns give the numbers after the application. The difference is given in absolute numbers (diff) and percents (diff %). The final

two columns give the ratio of possible grammatical solutions and candidate lemmas in the "swu" and 'mwu' case, respectively. With the application of multi word resources there is a reduction in all POS categories except for abbreviations, which is only natural. Besides unknown words (UNK) which have been reduced by 20%, the best results were obtained for numerals (NUM) where both possible grammatical solutions and candidate lemmas were reduced by more than 13% and for adjectives (around 8% for both categories). It is interesting to note that there was no significant change in the ratio of possible grammatical solutions and candidate lemmas.

Table 4. Ambiguities in the text per different parts-of-speech

PoS	**Grammatical solutions**				**Lemmas**				**Gs/Lemmas**	
	swu	**mwu**	**diff**	**diff%**	**swu**	**mwu**	**diff**	**diff%**	**swu**	**mwu**
UNK	620	496	124	20.00%	620	496	124	20.00%	1.00	1.00
A	41478	38125	3353	8.08%	7641	7046	595	7.79%	**5.43**	**5.41**
ABB	56	56	0	0.00%	56	56	0	0.00%	1.00	1.00
ADV	7431	7231	200	2.69%	7431	7231	200	2.69%	1.00	1.00
CONJ	6187	6024	163	2.63%	6187	6024	163	2.63%	1.00	1.00
INT	3167	3010	157	4.96%	3167	3010	157	4.96%	1.00	1.00
N	59834	58015	1819	3.04%	24235	23672	563	2.32%	**2.47**	**2.45**
NUM	4387	3808	579	13.20%	1612	1391	221	13.71%	**2.72**	**2.74**
PAR	7005	6817	188	2.68%	7005	6817	188	2.68%	1.00	1.00
PREP	11858	11345	513	4.33%	11858	11345	513	4.33%	1.00	1.00
PRO	28783	28620	163	0.57%	11873	11793	80	0.67%	**2.42**	**2.43**
V	34996	34549	447	1.28%	19733	19528	205	1.04%	**1.77**	**1.77**
Total	205802	198096	7706	3.75%	101418	98409	3009	2.97%	**2.03**	**2.01**

3.2 The ambiguity of lemmas

Further analysis showed that the most ambiguous type in terms of the assigned lemmas in both experiments – 'swu' and 'mwu' is the word form *Po.* It had a maximum of seven different lemmas assigned, however, only when appearing with the uppercase 'P'. Namely, both *po* and *Po* were assigned an adverb lemma (*jedno po jedno* 'one by one'), a particle lemma (*devet i po* 'nine and a half'), two preposition lemmas (a preposition with the instrumental case *po lepom vremenu* 'in nice weather' and a preposition with the accusative case *po detektiva* 'for detective'). However, to the uppercase *Po* three additional proper nouns were assigned: the English surname 'Poe' transcribed to Serbian, the Italian river *Po*, and the French town 'Pau', also transcribed to Serbian. All of the four former possibilities were realized in our text, but none of the proper nouns effectively appeared.

Six different lemmas were assigned to eight types in the 'swu' case and to seven types in the 'mwu' case, as shown in Table 5. It should be noted that all of them are case sensitive again;

namely, they had six lemmas assigned only when written in uppercase. The type *mile* which accounts for the reduction from eight types in the 'swu' case to seven in the 'mwu' case is particularly interesting for our research: it occurred only once in our text and as a component of the compound adverb *do mile volje* 'at their ease'. This adverb was recognized in the 'mwu' case and thus *mile* did not display its "high ambiguity" in this case. Unfortunately for our cause, the majority of other types with six assigned possible lemmas occurred more frequently (see Table 5), as did *Po* (16) with its seven lemmas.

Table 5. Types with six assigned possible lemmas

Type	Freq.	SWN	MWU
Bar	1	5 N, 1 PAR	5 N, 1 PAR
Bila	6	4 N, 2 V	4 N, 2 V
Dobro	32	1 A, 1 ADV, 4N	1 A, 1 ADV, 4N
Gore	4	2 A, 1 ADV, 1 N, 2 V	2 A, 1 ADV, 1 N, 2 V
Kose	2	1 A, 4 N, 1 V	1 A, 4 N, 1 V
Mile	1	1 A, 5 V	
Pola	2	1 ADV, 5 N	1 ADV, 5 N
Više	8	2 A, 1 ADV, 1 N, 1 PREP, 1 V	2 A, 1 ADV, 1 N, 1 PREP, 1 V

Five different lemmas were assigned to 26 types (both in the "swu" and the "mwu" case). It is interesting to note that *više* which appeared in the previous group with uppercase "V" appeared with lowercase "v" in this group as well, with the same possibilities except for the (proper) noun. By far the most frequent in this group was the type *da* (1 ADV, 1 CONJ, 1 INT, 1 PAR, 1 V) – it occurred 1386 times. Four of these five possibilities were actually realized in our text:

SR: *Nije prestajao* ***da*** {da,da.CONJ} *misli o tome.*

FR: Cela ne laissait pas de le préoccuper.

EN[1]: ... and these thoughts did not cease worrying him for a long time.

*„**Da*** {da,.da.INT} *!" odgovori Fileas Fog.*

FR: Oui, répondit Phileas Fogg.

EN: "Yes," returned Phileas Fogg.

SR: ***Da*** {da,da.PAR} *niste nešto zaboravili?*

FR: Vous n'avez rien oublié?

EN: You have forgotten nothing?

[1] English translations were obtained from http://www.gutenberg.org/files/103/103.txt.

SR: *U podne vodič* ***da*** {da,dati.V} *znak za polazak.*

FR: A midi, le guide donna le signal du départ.

EN : At noon the Parsee gave the signal of departure.

In the case of types with different lemmas assigned, the most frequent combination was a verb lemma and an adjective lemma – 615 cases, or 21% of the 2924 types with ambiguous lemmas. The cases that follow are: noun and verb (433; 14.8%), noun and noun (332; 11.4%), adjective and adverb (288; 9.8%), and verb and verb (222; 7.6%). All other combinations fall below 4% of types with ambiguous lemmas. The highest ambiguity between verbs and adjectives does not surprise: some verbal forms, like active and passive past participle and present and past gerund are often used as adjectives as well. In our texts some of these verbal forms, for some verbs, occurred in both functions: verbal or adjectival. Some examples are:

SR: *Zaista, idući* {idući.ići.V:S} *na istok, Fileas Fog je išao prema suncu,...* (present gerund of the verb *ići* 'to go')

FR: En effet, en marchant vers l'est, Phileas Fogg allait au-devant du soleil,...

EN: In other words, while Phileas Fogg, going eastward,...

SR: *A kad će proći idući* {idući.idući.A:adms1g} *voz? upita Fileas Fog.*

FR: Et le train suivant, quand passera-t-il? demanda Phileas Fogg.

EN: "And when does the next train pass here?" said Phileas Fogg.

SR: *Da nije bilo bure, usled koje je izgubljeno* {izgubljeno,izgubiti.V:Gns} *nekoliko sati...* (passive past participle of the verb *izgubiti* 'to lose')

FR: Sans cette tempête, pendant laquelle il perdit plusieurs heures,...

EN: Had there been no storm, during which several hours were lost,...

SR: *Bilo je moguće nadoknaditi izgubljeno* {izgubljeno,izgubljen.A:aens1g} *vreme.*

FR: Il n'était pas impossible que le retard fût regagné.

EN: It was not impossible that the lost time might yet be recovered;

3.3 The ambiguity of grammatical tags

The type with the greatest number of assigned grammatical tags was *gore* – a total of 32 grammatical tags: 20 A (for two adjectives with irregular comparative forms - *rđav* 'bad' and *zao* 'evil'), 2 ADV (adverbs *gore* 'worse' and *gore* 'up'), 6 N (all for one noun *gora* 'mountain'), 4 V (for two verbs - *goreti* and its jekavian equivalent *gorjeti* 'to burn'). This type appeared four

times in our text: in the ‘swu’ case as the two adverbs, and in the ‘mwu’ case as the simple adverb ‘worse” and as a component of the compound adverb *gore-dole* ‘up and down’:

SR: *Time **gore** {gore,.ADV} po sunce, gospodine!*

FR: Tant pis pour le soleil, monsieur!

EN: So much the worse for the sun, monsieur.

SR: *Paspartu je neprestano išao **gore-dole** {gore-dole,.ADV} po stepenicama kuće u Sevil-rou.*

FR: Passepartout ne cessa de monter et de descendre l'escalier de la maison de Saville-row.

EN: Passepartout continually ascended and descended the stairs.

Table 6 shows some other types with a high number of grammatical tags assigned. All of them appear in the same way in both the ‘swu’ and the ‘mwu’ case, except for the last example *železnička*. Namely, in the ‘mwu’ case *železnička* appears only as a component of two compounds: *železnička pruga* ‘railway’ and *železnička stanica* ‘railway station’. Moreover, to both of these multi-word unit types only one grammatical tag is assigned.

Table 6. Types with more than 24 grammatical tags assigned

Type	freq.	No. of tags	PoS	No. of grammatical tags
Čuvena	2	28	A,V	24 A, 4 V
Ostala	3	28	A, A,V	24 A, 4 V
Kose	2	27	A,N,N,V	7 A, 19 N, 1 V
Soli	2	26	N,N,V	22 N, 4 V
Ista	1	24	N,N,A,PRO	3N, 12A, 9PRO
Poveća	3	24	A,A,V	3V, 21A
Razna	1	24	A,A	24 A
Više	8	24	A,A, ADV, N, PREP, V	20 A, 1 ADV, 1 N, 1 PREP, 1 V
železnička	14	24	A,A	24A

4 Conclusions and further work

As we have expected, tagging with multi-word units reduces the ambiguity of a text. The reduction obtained in the given example, assessed by means of different measures is, however, not substantial. Our experiment showed that the considerable benefit obtained for particular cases fades away when the text is perceived as a whole. One of the reasons is the under-development of dictionaries of compounds and tools for multi-word units like dictionary graphs. Besides enlarging the general dictionaries and tools, production and application of tools specific to an analyzed text could be of great use. For instance, the correct recognition in our text of only three compounds *Fileas Fog* ‘Phileas Fogg’, *gospodin Fileas Fog* ‘mister Phileas Fogg’, *gospodin Fog*

'mister Fogg', occurring 306, 8 and 309 times respectively, would improve our results significantly, since the number of simple words (and tokens) would be reduced by an additional 631. Our results also pointed out the types on which we should concentrate our efforts, e.g. looking for multi-word units with the most ambiguous components, such as *po* or *gore*, constructing ELAG grammars around these types, etc.

Our future work will proceed in several directions. First we will try to extract more statistical information on the dependencies of ambiguous solutions (in the line of work presented in (Baptista & Faisca 2007)). Next, we will compare the ambiguous text with the text that has already been manually disambiguated in order to extract some useful information that could help to speed up the disambiguation process. Finally, we will try to add some practical solutions to the disambiguation process that rely on the high frequency of some specific two and three word sequences. For instance, the sequence *kao što je* occurs 4 times in our text and is always tagged as `{kao,.CONJ} {sxto,.CONJ} {je,jesam.V+Imperf+It+Iref+Aux:Pzsi}`, although lexical resources offer two solutions for *kao*, four solutions for *što*, and three solutions for *je*. The recognition of such unambiguously tagged sequences can have positive effects to the disambiguation process. The main benefit of this preliminary work is a step towards establishment of measures of ambiguity elimination that could be used for various approaches that we plan to work on in the future.

References

Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., Urizar, R. (2004). "Representation and treatment of multiword expressions in Basque". In *Second ACL Workshop on Multiword Expressions: Integrating Processing*", pp. 48-55, Barcelona Spain.

Baptista, J. and Faisca, L. (2007). "Mapping, filtering and measuring impact of ambiguous words in Portuguese", in *Formaliser les langues avec l'ordinateur : De INTEX à Nooj*, eds. Svetla Koeva, Denis Maurel, Max Silberztein, pp. 305-316, Presses Universitaires de Franche Comté, Besancon.

Laporte, E. and Monceaux, A. (1999). "Elimination of lexical ambiguities by grammars. The ELAG system", *Lingvisticae Investigationes* XXII, Amsterdam-Philadelphie : Benjamins, pp. 341-367.

Laporte, E., Nakamura, T., Voyatzi, S. (2008), "A French Corpus Annotated for Multiword Nouns", in: *Towards a Shared Task for Multiword Expressions* (MWE 2008), in scope of the *Sixth International Conference on Language Resources and Evaluation* (LREC'08).

Tufiş, D., Koeva, S., Erjavec, T., Gavrilidou, M., Krstev, C. (2008). "Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages". In Tadić, M., Dimitrova-Vulchanova, M. and Koeva, S. (eds.) *Proceedings of the Sixth International Conference Formal Approaches to*

South Slavic and Balkan Languages (FASSBL 2008), pp. 145-152, Dubrovnik, Croatia, September 25-28.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., Ramisch, C. (2007). "Validation and evaluation of automatically acquired multiword expressions for grammar enginiering". In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*, pp. 1034-1043, Prague, Czech Republic.

Vitas, D., Koeva, S., Krstev, C., Obradović, I. (2008) "*Tour du monde* through the dictionaries", *Actes du 27eme Colloque International sur le Lexique et la Gammaire*, L'Aquila, 10-13 septembre 2008, eds. M. Constant, T, Nakamura, M. De Gioia, S. Vecchiato, pp.249-256, Universite Paris-Est, Institut Gaspard-Monge.