Tour du monde through the dictionaries

Duško Vitas[1], Svetla Koeva[2], Cvetana Krstev[1], Ivan Obradović[1]
[1] Faculty of Mathematics, University of Belgrade, Studentski trg 16, RS - 110000 Belgrade
[2] Bulgarian Academy of Sciences, 52 Shipchenski prohod, Bl. 17, BG - 1113 Sofia

Abstract

In this paper we present a multilingual corpus generated from a novel and its translation into several languages that is aligned on the level of equivalent segments. The main purpose of this corpus is analyzed along with the possibilities of application of monolingual lexical resources that are implemented for the system Unitex 2.0 in its exploitation. We also look into the preconditions necessary for formulating general queries for extracting information from aligned texts.

Keywords : aligned corpora, electronic dictionaries, morpho-syntactic annotation

1. Introduction

A general review of methods for text alignment is outlined in (Laporte: 2006). The integration of a monolingual corpus processor such as Unitex into the exploitation of multilingual resources is particularly stressed. The version 2.0[1] of the Unitex system has the option for the alignment of two texts as well as the concordancer that enables the necessary correction of alignment to be performed and queries to be posed over the bitext - parallel aligned text). For all texts that form a bitext Unitext produces in the preprocessing phase the corresponding dictionaries of simple words, compounds and unrecognized words. Such an enhancement of the Unitex system poses the following questions: (a) resources that a user can use to experiment with bitexts; and (b) comparison of annotations incorporated in the particular monolingual resources. In this paper we present the first steps made in order to find answer to these questions. Text chosen to perform these experiments is Jules Verne's novel *Around the world in eighty days* that, first of all, presents the sample text for the French distribution of the Unitex system and for which the translations exist in digital form for all European languages for which e-dictionaries exist in LADL format.

The envisaged system should for a given query, regardless of whether it is formulated as a regular expression or a graph, generate bilingual concordances of a bitext on the basis of the information found in monolingual dictionaries. Besides the query response in a given language, these concordances would also contain the equivalent segment from one (or more) aligned texts. In this paper we will analyze the problems related to text segmentation with the system Unitex (which is the important issue in the alignment problem), creation of aligned versions with the integrated aligner, and finally, the possibility of application of existing monolingual lexical resources in LADL format to exploitation of bitexts. The aim of this paper is to analyze the preconditions necessary for the formulation of a query over a bitext that would extract equivalent segments in two or more languages. It is very likely that in the

---

[1] http://www-igm.univ-mlv.fr/~unitex/download3.html

first step the named entities represent the most suitable domain for such an experiment, since they are potential cognates in a bitext. Namely, the supposition is that the majority of named entities must appear in equivalent segments of a bitext.

A previous systematic approach to the development of multilingual corpora was initiated within the Multext project[2], which subsequently included East-European languages through the Multext-East project[3] (Erjavec: 2003). These projects provided two important resources, a multilingual annotated aligned corpus Orwell's 1984, and a proposal for standards of morpho-syntactic annotation. Results of these projects are incorporated in the largest multilingual corpus composed of European Union legislation in 22 languages[4]. The annotation model developed within the Multex-project played an important role in various applications during the past decade, but the concept of annotation is nevertheless subject to different criticisms (Krstev: 2008). On the other hand, monolingual resources for the majority of European (and not only European) languages, which enable a considerably more complex processing of corpora than in the case of statistically based methods, have been developed within the RELEX network[5]. It is thus natural to raise the question of the possibility of the use of these monolingual resources in the processing of multilingual corpora. It would be particularly important if the concept of local grammar could be generalized in such a way that one local grammar extracts (approximately) the same objects from texts in different languages, using the information stored in the system of electronic dictionaries. The solution of this problem depends on at least two components. One of them is how "faithful" is the translation compared to the original, and the other – the manner in which specific information is marked in the system of electronic dictionaries.

2. The Corpus

Jules Verne's novel was chosen for this experiment for two basic reasons: the text is available in digital form for the majority of languages that are relevant for this experiment, and on the other hand, regarding its content, it represents a suitable text for different types of analysis, especially in the domain of named entity recognition (geographical concepts and different measures).

During the research process, versions of Verne's novel in fifteen languages have been acquired, namely: French, English, German, Spanish, Portuguese, Italian, Romanian, Russian, Serbian (including a separate translation in Croatian), Bulgarian[6], Macedonian, Polish, Hungarian and Greek. Texts were aligned for twelve languages for which appropriate dictionaries existed (all except Hungarian, Italian and Romanian). Besides these, versions in Dutch and Slovenian are also available, although appropriate dictionaries for these languages do not exist (Vitas: 2006).

---

2 http://aune.lpl.univ-aix.fr/projects/MULTEXT/

3 http://nl.ijs.si/ME/

4 http://langtech.jrc.it/JRC-Acquis.html

5 http://infolingu.univ-mlv.fr/Relex/Relex.html

6 Part of the SEE ERANET project *Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages.*

In the preparatory phase each translation was marked in accordance with the TEI-standard in XML, and the title (<head>), paragraph ( ) and "sentence" (<seg>) were included as units of text logical layout. Before alignment, each text was transformed to the TEI-conformant format[7]. The XAlign system[8], that is now being incorporated in the version 2.0 of the Unitex system, was used for the alignment process. The strategy used by this program is, basically, to attempt to align in the first step the tree structure of texts within bitexts (encoded by XML-tags), and then, in the second, to align segments of different length.

Starting from the French version, the goal of the alignment was to establish 1:1 relations on the segment level (<seg> tag) with all other languages. This type of text alignment of bitexts required an intensive manual control of the output of the XAlign system, using the attached concordancer[9] that shows the aligned pairs. In this way, the missing segments or the inconsistencies between the source text and its translations were also identified.

In a certain number of translations there are sentences which do not exist in the original, which leads to the supposition that another version of the original text exists[10]. There are other differences of diverse nature, which are a consequence of either translators' liberties (or flaws) or specific features of languages the text was translated to. In that regard, among the correct translations the freest translation is the one in English. As an example, to the French segment:

<seg>A l'un des plus grands orateurs qui honorent l'Angleterre, succédait donc ce Phileas Fogg, personnage énigmatique, ...

corresponds in this translation,[11] which is compliant to the paper edition,[12] only the segment *an enigmatical personage*,..., while the part that precedes is omitted.

Another example is the translation of the segment:

<seg>-- Le Morning Chronicle assure que c'est un gentleman.</seg>

<seg>"The Daily Telegraph says that he is a gentleman."</seg>

where the named entity *Morning Chronicle* was replaced by *Daily Telegraph* although *Daily Telegraph* later appears in the French text as well.

More serious problems occur with the translation in Italian obtained from the site of the Carmel[13] project, which in its greater part represents a very free translation.

In each of the translations, a number of segments representing either a part of the sentence or an entire sentence are missing. Such parts were marked manually as comments in the text and replaced by the appropriate sequence from the original. For example, in the Serbian

---

7 http://www.tei-c.org/index.xml

8 http://led.loria.fr/download/source/Xalign.zip

9 http://led.loria.fr/download/source/concordancier.zip

10 On http://gallica.bnf.fr/ark:/12148/bpt6k89828c the 1873 and 1874 editions can be found, which are compliant with the version distributed under Unitex.

11 http://www.gutenberg.org/etext/103

12 Jules Verne: *Around the World in Eighty Days*, Penguin Popular Clasics, Penguin Books, 1994.

13 http://www.projetcarmel.org

translation a total of 16 segments are missing, while the number of missing segments in Bulgarian is 36, in Portuguese 4, in Spanish 49, in Greek 7, in Polish 16, etc. The biggest discrepancies are found in the German translation[14] where entire paragraphs are omitted in several places. The alignment process cannot identify such segments by itself, and in the majority of cases they became the source of 1:n (n>1) relations that do not represent equivalent segments.

Finally, differences in segmentation can result from orthographic differences or differences in the sequence of words among languages. In that sense, the German translation is the most interesting, because it is written according to old orthography, which considerably differs from current German editions of this novel.

3. Comparison of text sizes

Table 1 demonstrates the length of unmarked texts for different languages processed by the system Unitex 1.2, expressed in the total number of simple forms N, the number of different simple forms V, as well as the number of appearances of decimal digits within the text. It is obvious at the first sight that the more or less same content is expressed in different languages by texts of different lengths. The text in Polish is about 25% shorter than the French original, but the number of different words is about 60% greater than in French. This result is partly a consequence of grammatical differences (absence of article, null subject, omitting of auxiliary verbs in some composite tenses), as well as of derivational potential of Polish.

The counting results would be different if the counting process included additional conventions regarding the writing of words as simple words or compounds. For example the French segment *brut dix-sept cent soixante-dix tonnes* has its counterpart *водоизместимост хиляда седемстотин и седемдесет тона* in Bulgarian, *bruto hiljadu sedam stotina tona* in Serbian, etc. In other words, seven French words were translated by six words in Bulgarian and five in Serbian in which the translation of *soixante-dix* is missing.

	Total		**Total**	**Coverage (words)**		
Language	**N**	**V (N/V)**	**digits**	**Simple**	**Comp.**	**Unkn.**
French (FR)	71793	9433 (7.6)	440	13437	2171	336
English (EN)	63743	7434 (8.8)	262	10695	281	268
Spanish (ES)	65064	9959 (6.5)	277	12351	0	1241
Portuguese (PG)	65037	10271 (6.3)	435	11929	286	725
German (DE)	62726	10228 (6.1)	315	17691	258	2120
Italian (IT)	68450	11599 (5.9)	582	10624	216	3530
Greek (EL)	68615	11809 (5.8)	295	3894	69	7581
Bulgarian (BG)	58678	11217 (5.2)	400	10850	0	755
Serbian (SR)	58722	12733 (4.6)	279	14947	347	89

[14] http://gutenberg.spiegel.de/

Russian (RU)	56293	14708 (3.8)	232	16668	435	689
Polish (PL)	54871	15406 (3.6)	265	9196	13	8812

Table 1

In the column ***Coverage (words)*** the results of the recognition of simple words (Simple) and compounds (Comp.) by the system Unitex, as well as the number of words that remain unrecognized (Unkn.).

On the other hand, differences in the usage of digits originate partly from different systems of numeration in texts (Roman or Arabic digits), but also from different writing of numbers, such as in the following example:

FR: *avec promesse, en cas de succès, d'une prime de deux mille livres (50 000 F) et cinq pour cent de la somme qui serait retrouvée*

EN: *inspired by the proffered reward of two thousand pounds, and five per cent. on the sum that might be recovered*.</seg>

BG: *с обещанието, че при успех ще има награда от две хиляди лири (50 000 франка) и пет процента от намерената сума*

SR: *uz obećanje da će u slučaju uspeha dobiti nagradu od dve hiljade livara i pet od sto od sume novca koji bude nađen*

The example demonstrates that the English and Serbian translator omitted the conversion of the sum of 2000 pounds into francs, and that the Bulgarian translator expanded the abbreviation F into *франк*.

A similar example is the way the toponym *Hong Kong* is written in different translations, namely as one or two words: in French, Serbian and Spanish as *Hong-Kong*, in Portuguese, English and Croatian as *Hong Kong*, in Greek as *Χονγκ Κονγκ*, in German, Bulgarian and Polish as *Hongkong* (*Хонгконг*), and in Russian as *Гонконг*. The example of this toponym points to another problem in word counting: neither *Hong* nor *Kong* exist as a separate lexical unit in cases when they are written as two words, so in a more correct counting they should be counted as one.

One other aspect of the comparison of the word number counts in different languages is revealed when the *multiword nouns* which in the French text represent one lexical unit (Laporte: 2008) are linked with their equivalents in different languages. For instance, a number of structures of the form NDN in French is translated in Serbian and Bulgarian in the units of the form AN or NN.

4. Numbers, numerical expressions and proper names

It could be expected that expression written in digits will be literally transferred to the translation, thus making their recognition potential cognates in the correction of alignment. However, each of the translations has its specific features in this sense.

The first difference is illustrated by the previous example of conversion of pounds into francs, and this is the main cause for differences in the column *Total Digits* in Table 1. The second source is the localization in writing numbers. For example, in the sentence from the original:

<seg id="n3113">...pesanteur ... est de **1 170**, celle de l'eau ...étant **1 000**.</seg>

numbers in translations are written in different ways: 1 000 and 1 176 - PL, 1,000 and 1,170 – ES, EN, 1.000 and 1.170 – PT, EL, 1000 and 1170 SR, IT, Удельный ...- тысяча сто семьдесят, ... воду за тысячу – RU, eintausendeinhundertundsiebenzig gegen eintausend des ...Wassers – DE, missing segment in Bulgarian.

Besides numbers written by digits, Verne's text is abundant with numerical expression written in words. Processing of such expressions depends on the way information on the number is expressed in the dictionary. As an example, let us consider the sentence from the original text:

<seg id="n140">Reform-Club mettait à sa disposition deux bibliothèques</seg>

The way these numbers are written in the text, as well as the appropriate codes, are language depend and reflect a specific view of this type of words. In a certain number of dictionaries, the number two is defined as a determinator having the subcategory number, in some others as an adjective or noun, and in others still, it represents a particular type of word (NUM), for example:

FR: deux,.DET+Dnum+z1:mp:fp or N+Nnum+z1:ms:mp,

EN: two,.DET+Dnum:p or two,.N:s,

DE: zwei,.NUM or zwei,.Num+FF

BG: две,два.NUM+ORD+DVE:s,

SR: dve,dva.NUM+v2+Ek:fp1g:fp4g:fp5g,

RU: две,два.NUM+pauc:nF:ajF, etc.

We can conclude that a query that identifies a number in each particular language cannot be expressed by a general pattern, due to the absence of standardized code, as well as different views on the way this word should be marked.

A more complex example is found in the sequence:

<seg id="n186">...avec promesse ... d'une prime de deux mille livres ...</seg>

where the number deux mille could be interpreted as a sequence of two numbers (*deux* + *mille*), as a simple word (in German: *zweitausend* described as DET+ADJ+Num:X) or as a compound (in Serbian: *dve hiljade*,.NUM+C+v5).

Two classes of proper names appear in the text of the novel: names of real entities and names of novel heroes. The latter category, as a rule, belongs to the group of unknown words (Phileas Fogg, Aouda, Passepartout).

In the first group, toponyms, such as San Francisco and New York appear. Dictionaries that come with Unitex 1.2 (Paumier, 2002) process them in different ways. These two toponyms are recognized in dictionaries of French, Polish, Russian, Serbian, German (San Francisco only) and Italian. Other dictionaries either do not recognize these two proper names (Greek) or recognize them partially (e.g. in Spanish only York and Francisco are marked as proper names, in English New is an adjective, York a proper name and in German New is unknown and York is English toponym). As in the case of numbers, differences can be found in assignment of attributes which describe these entities with respect to their possible description in a database of the Prolex type (Grass: 2004), (Maurel; 2007). Names of the novel heroes, as we have already mentioned, are unrecognized words, but their frequency in the text is not negligible, and it amounts to more than 2% of the text.

5. Exploitation of aligned texts

The described variability of translations, both in length and in the number of different words, indicates that cases of biunivocal correspondence between two languages on the word level will be rare, even in the case of closely related languages, such as Romance or Slavic languages. On the other hand, despite all indicated differences and discrepancies between the original and its translations, all translations nevertheless express the content of the original text. It is thus natural to raise the question whether partial or complete identification of equivalent parts within the <seg> type tags is possible without bilingual dictionaries being consulted. The answer to this question directly depends on the completeness of dictionaries available for experimental purposes, as well as on the theoretical framework in which the lemma and its properties are interpreted in a particular language. Some experiments in recognition of equivalent sequences can be realized using the model of local grammars. They will be illustrated by an example of the recognition of one class of named entities.

For example, let us observe the example of annotation of named entities for some measures on the aligned texts of Verne's novel. An expression for a measure in French and Bulgarian/ Serbian is depicted by the graphs in Figure 1 which describe it as a structure of a sequence of numbers written by words followed by a measure indicator (kilometer, grade, mile, foot, etc) which is obviously language dependent.

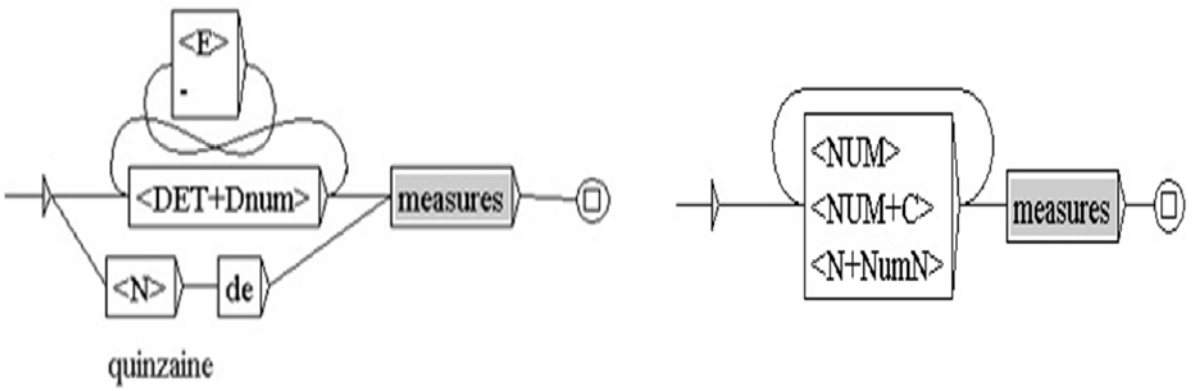

Figure 1. The graphs for the recognition of measure expressions in French and Serbian/Bulgarian

The examples of the first fifteen sequences retrieved by these graphs are presented in *Table 1.* It should be stressed that the further development of such an aligned resource relies heavily on the development of some additional resources with aim to standardize the field of syntactic and semantic markers in electronic dictionaries.

quatre-vingt-quatre degrés Fahrenheit	osamdeset četiri stepena Farenhajtovih	осамдесет и четири градуса по Фаренхйт
deux mille huit cents tonnes	dve hiljade osam sto tona	две хияади и осемстотин тона
dix milles	deset milja	десет мили
neuf milles	devet i po milja	девет мили
cent soixante kilomètres	sto šezdeset kilometara	сто и шестдест километра
deux milles mètres	dve hiljade metara	две хияади метра
treize cent dix milles	hiljadu tri stotine deset milja	хияада триста и десет мили
treize cent dix milles	hiljadu tri stotine deset milja	хияада триста и десет мили
seize cent cinquante milles	hiljadu šest stotina pedeset milja	хияада шестстотин и петдесет мили
quatorze cent mille milles carrés	četiri stotine hiljada kvadratnih milja	четиристотин квадратни мили
sept cent mille milles carrés	sedam stotine hiljada kvadratnih milja	седамстотин хияадиквадратни мили
onze cents milles	hiljadu sto milja	хияада и сто мили
soixante-dix-sept degrès	sedamdeset i sedam stepeni	седамдесет и седем градуса

quinze milles	petnaest milja	петнадесет мили
cinquantaine de milles	pedeset milja	петдесетина мили
vingtaine de milles	dvadeset milja	двадесетина мили
vingt-cinq milles	dvadeset i pet milja	двадесет и пет мили
douze milles	dvanaest milja	дванадесет мили
deux milles	dve milje	две мили
Quatre-vingts milles	osamdeset milja	осамдесет мили

Table 2 The first fifteen occurences in bitext recognized by graphs in Figure 1

6. Conclusion

The analysis of the multilingual aligned corpus containing Jules Verne's novel *Around the world in eighty days* in the original and a dozen translations gives a very interesting insight in the possibilities of exploitation of aligned texts. It has been shown that many pitfalls exist on the path to formulating a unique query which could extract information from aligned texts. Substantial differences in translations were found even in places where they would not have been expected, such as in numbers, numerical expressions and proper names. However, it has been demonstrated that nevertheless possibilities exist for formulating unique queries for some language combinations, for closely related languages in the first place, such as French and Portuguese or Slavic languages (Serbian, Bulgarian, Russian, Polish), using the existing monolingual resources.

References

GELBUKH, A., SIDOROV, G. VERA-FÉLIX, J.-A. (2006) « A Bilingual Corpus of Novels Aligned at Paragraph Level ». In proc. *FinTAL*-2006. *Lecture Notes in Artificial Intelligence*, no. 4139, Springer-Verlag, pp. 16–23

GRASS T., MAUREL D. and TRAN M. (2004), « Prolexbase : Une ontologie pour le traitement multilingue des noms propres », in *Linguistica Antverpiensia*, NS3:293-309.

ERJAVEC T., KRSTEV C., PETKEVIČ V., SIMOV K., TADIĆ M., VITAS D. (2003), « The MULTEXT-East Morphosyntactic Specifications for Slavic Languages », in Erjavec T. and Vitas D, *Proceedings of the Workshop on Morphological Processing of Slavic Languages* : 10th Conference of the European Chapter, EACL 2003, Budapest, Hungary, pp. 25-32

KRSTEV C., KOEVA S. and VITAS D. (2008) « A Dictionary-based Model for Morpho-Syntactic Annotation » in: *Proceedings of the 2nd Linguistic Annotation Workshop*, in scope of the *Sixth Interantional Conference on Language Resources and Evaluation* (LREC'08), Marrakech, Morocco, 26 May 2008, European Language Resources Association (ELRA)

LAPORTE, E., KRSTEV, C., VITAS, D. (2006), Preparation and exploitation of bilingual texts, *Lux Coreana* no. 1, Han-Seine, pp. 110-132

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 Interantional Conference on Language Resources and Evaluation* (LREC'08), http://multiword.sourceforge.net/download/MWE2008-papers/8_Laporte.pdf

MAUREL D., VITAS D. KRSTEV C., KOEVA S. (2007) « Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian », in: *Bulag - Bulletin de Linguistique Appliquée et Générale*, Les langues slaves et le français : approches formelles dans les études contrastives, eds. Aleksandra Dziadkiewicz et Izabella Thomas, No. 32, pp. 55-72, Presses Universitaires de Franche Comté, Besancon

PAUMIER, S. (2002), *Manuel d'utilisation du logiciel Unitex,* IGM, Université de Marne-la-Vallée. http://www-igm.univ-mlv.fr/~unitex/manuelunitex.pdf.

VITAS D. KRSTEV C. (2006), « Literature and Aligned Texts », in: *Readings in Multilinguality*, eds. Milena Slavcheva, Galia Angelova and Kiril Simov, pp. 148-155, Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria