Parallel French-Slovak Corpus

Dorota Vasilišinová and Radovan Garabík

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences,
Bratislava, Slovakia
{dorota,garabik}@korpus.juls.savba.sk

Abstract. Presented French-Slovak parallel corpus «FRASK» is a sizeable corpus consisting of European Union legislative texts and fiction in both French and Slovak languages. Texts are sentence-aligned, lemmatized and contain morphological information. The searching mechanism includes the possibility to query single words, phrases, lemmas and morphology tag, using regular expressions. The corpus is publicly available on the internet.

1 Introduction and choice of texts

The intended scope of the corpus is twofold: first, to create an aligned corpus of French and Slovak text for general purposes, and second, to support cross-language terminology research, especially with emphasis on legal and economic texts of the European Union legislature. The corpus therefore consists of two kinds of texts, the first part consisting of fiction and the second consisting of a collection of texts of European Union law. At the moment, the fiction part of the corpus contains three French novels and their translation into Slovak. Texts of European Union law include The Official Journal of the European Union, treaties, legislation, case law, preparatory acts and parliamentary questions. These texts were obtained from the JRC-ACQUIS Multilingual Parallel Corpus, Version 3.0 [jrc07], where the texts were already downloaded from the European Union information portal and conveniently converted into the XML format – but without any additional linguistic annotation, nor language-aligned.

The size of the corpus is 334 021 French and 226 990 Slovak words for the fiction part and 65 797 270 French and 59 076 782 Slovak words for the EU law part, totaling 66 131 291 French and 59 303 772 Slovak words (punctuation included).

2 Text format and processing

Texts in the corpus are processed in several phases using a modular system where each conversion step is applied to the previous level of conversion. There are several levels of conversion:

1. Conversion from the original file format (HTML, MS Word, etc.) into a simple text format (UTF-8 encoding, paragraphs separated by a blank line).

2. Manual editing of the document, where applicable (not in the case of the EU subcorpus). Stray texts at the beginning and end of the documents were compared and brought into agreement – there are often differences across the translations in the format of the document title, author, editorial prologues or epilogues.
3. Conversion into TEI XML format, with paragraphs marked by a corresponding XML tag.
4. Lemmatization and part-of-speech (or full morphological) tagging, converting the document into TEI XML format with sentence delimiters and grammar information for each word.
5. Conversion into simple text format suitable for the hunalign aligning program (using only lemmas, to help the aligning process), with a special sign ‘¶’ as a paragraph separator.
6. Adding the alignment back to the TEI XML format as an attribute for the sentence XML tag, linking to the corresponding sentence(s) in the opposite language document.
7. Converting the data into a vertical file format, suitable for the Manatee corpus manager indexing.

Before lemmatization, the texts were typographically normalized – different quotation marks (Slovak „ “ ” and French “ ” « ») were all internally translated into simple straight quotes `" (U+0022 QUOTATION MARK)` and various kinds of dashes were translated into `U+002D HYPHEN-MINUS` for the benefit of TreeTagger, which works internally in the Windows-1252 codepage and cannot properly deal with rich typographical characters.

2.1 French lemmatization and POS tagging

French texts have been lemmatized and morphologically annotated with TreeTagger, a tool for annotating text with part-of-speech and lemma information. The part-of-speech tag system used is described in [Ste03]. POS tags for the French language include 33 tags which describe major word classes and some of their inflectional variants (e.g. verbs in conditional, future tense, imperative etc.), tags for special word forms (abbreviations, acronyms), miscellaneous symbols and certain punctuation marks.

The French letter (ligature) *e dans l’o (Œ, œ)* has been retained in the corpus. Although the majority of the texts used the simple *oe* character sequence (probably due to inadequate historical use of the ISO/IEC 8859-1 character encoding), we decided to keep the *œ* character, if present in the source texts. This means that both the variants (e.g. *coeur* and *cœur*) are considered to be two different words and special care has to be taken when querying the corpus (e.g. by using the appropriate regular expression `"c(oe|œ)ur"`). Lemmatization contains the orthographically correct *œ* form regardless of the original variant, so when querying the lemma attribute only the canonical form needs to be used: `[lemma="cœur"]` (compare with `[lemma="moelleux"]`).

2.2 Slovak lemmatization and POS tagging

Slovak texts contain complete morphological information. Each word is assigned a lemma and a morphological tag, containing all the relevant grammar information (such as gender, case, number, tense, aspect). The tagset used is described in [Gar06], and for homonymy disambiguation, we are using the Hunpos tagger [HKO07] trained on a manually annotated corpus of about 511 thousand tokens.

3 Alignment accuracy

Texts were aligned using the hunalign [VNH$^+$05] software, which works on a sentence level, using a combination of length and dictionary based similarities to align the parallel texts. Although hunalign is able to work without a supplied dictionary, using one can improve the alignment dramatically. Since no French-Slovak dictionary was available, we bootstrapped a dictionary from automatically generated aligned word pairs, manually correcting the entries, obtaining an initial dictionary of 1 505 entries, and then running the alignment again, generating a new automatic dictionary and correcting it again manually. At the end, we obtained a dictionary of 6 858 manually verified word pairs. Alignment accuracy was estimated by choosing several (Slovak) words and randomly choosing several hundred concordances semi-uniformly dispersed throughout the corpus and manually counting the number of matching bisentences. We considered only 'perfect' matches, i.e. only those, where one source language sentence was translated by one target language sentence and correctly aligned[1]. In the following tables, we see the accuracy compared using the initial small dictionary, using the final dictionary and for the fiction corpus only, for the whole corpus, and for the whole corpus with filtered bisentences only (taking into account only those bisentences where alignment score as given by hunalign exceeds 0.5 and the lengths of original and translated sentence differ by less than 30 %).

word	dictionary	
	smaller	bigger
malý	58.5	63.2
počuť	77.8	84.4
voda	62.5	60.5
alebo	62.6	66.7
total	63.9	66.9

Table 1. Improving alignment accuracy by increasing dictionary size, the whole corpus.

[1] Obviously, using this method we can never reach 100 % accuracy, because often there is not a 1:1 correspondence between original and translated sentences, and even if correctly aligned, we do not count such translations as accurate.

word	dictionary	
	smaller	bigger
malý	76.7	91.5
počuť	69.2	83.7
voda	69.0	84.5
alebo	69.3	79.7
total	71.5	85.0

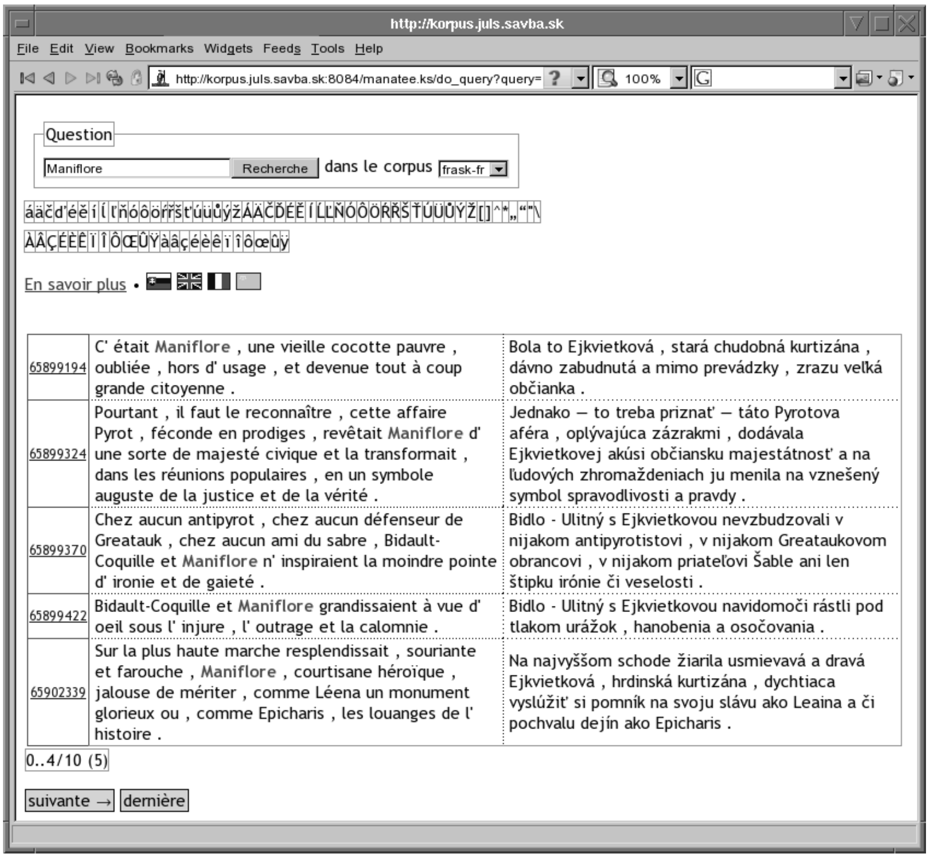
Table 2. Improving alignment accuracy by increasing dictionary size, fiction only.

word	corpus		
	fiction	whole	filtered
malý	91.5	63.2	94.5
počuť	83.7	84.4	87.1
voda	84.5	60.5	83.0
alebo	79.7	66.7	90.3
total	85.0	66.9	88.8

Table 3. Comparing alignment accuracy, bigger dictionary.

4 Query interface

Corpus backend is provided by the Manatee server [Ryc00], where each half (Slovak and French) of the corpus is indexed separately. Links between the halves are provided in form of a `link` attribute to the sentence XML tag (i.e. `<s link="5+6" id="4">...</s>` means that the 4^{th} sentence in one language corresponds to the 5^{th} and 6^{th} sentences in the other language). On top of the Manatee libraries, a custom WWW-based search interface has been built, using the Karrigell web application framework [kar07] in the Python programming language. The query interface follows the CQP syntax and provides full regular expression queries for words, lemmas and POS tags (or morphosyntactic attributes), displaying the result in a KWIC-like format, with parallel text from the other language displayed alongside.


http://korpus.juls.savba.sk
File Edit View Bookmarks Widgets Feeds Tools Help
http://korpus.juls.savba.sk:8084/manatee.ks/do_query?query=
100%
Question
Maniflore
Recherche
dans le corpus
frask-fr
áäčďéěíĺľňóôöŕřšťúüůýžÁÄČĎÉĚÍĹĽŇÓÔÖŔŘŠŤÚÜŮÝŽ[]^*„“"\
ÀÂÇÉÈÊÏÎÔŒÛŸàâçéèêïîôœûÿ
En savoir plus •

65899194	C' était Maniflore , une vieille cocotte pauvre , oubliée , hors d' usage , et devenue tout à coup grande citoyenne .	Bola to Ejkvietková , stará chudobná kurtizána , dávno zabudnutá a mimo prevádzky , zrazu veľká občianka .
65899324	Pourtant , il faut le reconnaître , cette affaire Pyrot , féconde en prodiges , revêtait Maniflore d' une sorte de majesté civique et la transformait , dans les réunions populaires , en un symbole auguste de la justice et de la vérité .	Jednako – to treba priznať – táto Pyrotova aféra , oplývajúca zázrakmi , dodávala Ejkvietkovej akúsi občiansku majestátnosť a na ľudových zhromaždeniach ju menila na vznešený symbol spravodlivosti a pravdy .
65899370	Chez aucun antipyrot , chez aucun défenseur de Greatauk , chez aucun ami du sabre , Bidault-Coquille et Maniflore n' inspiraient la moindre pointe d' ironie et de gaieté .	Bidlo - Ulitný s Ejkvietkovou nevzbudzovali v nijakom antipyrotistovi , v nijakom Greataukovom obrancovi , v nijakom priateľovi Šable ani len štipku irónie či veselosti .
65899422	Bidault-Coquille et Maniflore grandissaient à vue d' oeil sous l' injure , l' outrage et la calomnie .	Bidlo - Ulitný s Ejkvietkovou navidomoči rástli pod tlakom urážok , hanobenia a osočovania .
65902339	Sur la plus haute marche resplendissait , souriante et farouche , Maniflore , courtisane héroïque , jalouse de mériter , comme Léena un monument glorieux ou , comme Epicharis , les louanges de l' histoire .	Na najvyššom schode žiarila usmievavá a dravá Ejkvietková , hrdinská kurtizána , dychtiaca vyslúžiť si pomník na svoju slávu ako Leaina a či pochvalu dejín ako Epicharis .

Fig. 1. Example of the query interface; searching for a proper noun.

5 Conclusion and further work

From the alignment accuracy comparisons we see that the alignment depends heavily on the size (and presumably quality) of the bilingual dictionary available. Our final dictionary of 6 858 words is obviously too small to cover much of the input texts, and does not contain many specialized words frequently present in legal texts. Our first necessary task will be to increase the size of the dictionary and to add the most frequent terms present in the European Union texts.

Since the provenience of the EU translations is not very clear, it is possible that we are dealing with two parallel translations into French and Slovak, not with the original and translation (in fact, the majority of the texts are probably just translations from original English). This does not diminish the usefulness of the corpus as such, but compels us to interpret the results with care and to apply additional measures to improve the corpus accuracy. In particular, we have to implement filtering, removing misaligned sentences and eventually also sentences containing too much nontextual information – in the EU texts, there are often various lists, enumerations, tables and other elements, as well as complete texts in third unrelated languages in both the French and Slovak parts. Filtering out this content would improve the usefulness of the corpus texts and improve the aligning tools accuracy.

In the future we plan to provide the French part of the corpus with complete morphosyntactic annotation, using the FLEMM analyzer [Nam00]. In addition, an increase in the amount of texts in the corpus is a high priority, in order to augment the (rather small) fiction part to a more representative volume.

References

[Gar06] Radovan Garabík. Slovak morphology analyzer based on Levenshtein edit operations. In *Proceedings of the WIKT'06 conference*, pages 2–5, 2006.

[HKO07] Péter Halácsy, András Kornai, and Csaba Oravecz. HunPos – an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion*, pages 209–212. Association for Computational Linguistics, 2007.

[jrc07] `http://langtech.jrc.it/JRC-Acquis.html`, 2007.

[kar07] `http://karrigell.sf.net/`, 2007.

[Nam00] Fiammetta Namer. Flemm: Un analyseur Flexionnel de Français à base de règles. In Christian Jacquemin, editor, *Traitement automatique des Langues pour la recherche d'information*, pages 523–547, Paris, 2000. Hermes.

[Ryc00] Pavel Rychlý. *Korpusové manažery a jejich efektivní implementace.* PhD thesis, Faculty of Informatics, Masaryk University, Brno, 2000.

[Ste03] Achim Stein. French TreeTagger Part-of-Speech Tags, 2003. `http://www.ims.uni-stuttgart.de/ schmid/french-tagset.html`.

[VNH$^+$05] Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing 2005 Conference*, pages 590–596, 2005.