Compiling and Using a French-Slovenian Parallel Corpus

Adriana Mezeg

University of Ljubljana, Faculty of Arts

Abstract: The article presents the compilation and application of a French-Slovenian parallel corpus, the first independent parallel corpus for this language pair. The first part of the article focuses on some of the aspects of corpus design and development (text availability and collection, copyright, alignment and annotation), whereas in the second part, to demonstrate corpus use and usefulness for contrastive and translation studies research, we present a case study concerned with the translation of French sentence-initial gerundial (*i.e. en* participle) clauses into Slovenian. Due to the implicitness of syntactic and semantic elements in French non-finite clauses, which often hinder their interpretation and comprehension, we assume that Slovenian translators tend towards the explicitness of these elements. The analysis confirms this hypothesis in that more than 95% of the Slovenian translations are syntactically more explicit than their source, *i.e.* French counterparts, whereas semantically speaking the explicitness amounts to 85%.

1 Introduction

Since the development of the first corpora and the general awareness of their advantages, they have become indispensable in virtually all the areas dealing with the study of language: grammar, lexicology, lexicography, language teaching, contrastive and translation studies, etc. Through large national projects, usually financed by public and private institutions and carried out by experts in linguistics and natural language processing, numerous countries have developed large reference corpora for their respective national languages. From a national perspective, parallel corpora are generally not that vital, as they are mostly of interest to a limited circle of experts. Their compilation is thus usually undertaken by enthusiastic individuals, mostly linguists, translation scholars or PhD students, who find it necessary to provide modern contrastive descriptions of languages on the basis of large quantities of authentic data, to compile modern bilingual general and specialised dictionaries (or glossaries) or to modernise the already existing obsolete ones, to study translation or language phenomena in different text types in two or more languages and use the findings in translator training and in translation practice, etc. These are also some of the reasons for the compilation of *FraSloK*, the first French-Slovenian parallel corpus.

Making a corpus of texts in language A and their translations into a language B is a long and complex process for several reasons: (non-)availability of large quantities of (electronic) texts for less translated language pairs, securing permission from copyright holders of texts for both languages in question, alignment and annotation. These issues will be discussed with regard to the development of a 2.5-million word French-Slovenian corpus (*FraSloK*) of contemporary literary and journalistic texts, which is, in spite of its small size, an invaluable source of data for contrastive studies and exploration of translation phenomena for the language pair in question. This will be demonstrated in the second part of the paper, dealing with a contrastive translation studies analysis of French sentence-initial gerundial clauses and their Slovenian translations, extracted from the corpus with Michael Barlow's *ParaConc* (2001).

2 *FraSloK* design and development

In the first part of the paper, we focus on the compilation of a French-Slovenian parallel corpus, which was undertaken in November 2007 and completed in January 2010. It was primarily intended to serve as a basis for a contrastive analysis of French detached constructions and their

Slovenian translations which we wanted to conduct within our PhD thesis. However, we strove from the beginning for its wide-ranging usefulness.

2.1 Text availability and collection, copyright

When we started planning the compilation of our corpus, we immediately decided it should contain written contemporary complete texts, as such a corpus would meet a growing need among Francophiles in Slovenia for this kind of language resources. Being limited in time and on our own in this project, the envisaged size was one million words per language, this providing a solid basis for the envisaged research.

The decision on what text types to include in the corpus was subject to the availability of a sufficient number of Slovenian translations of original French texts, therefore genre selection could not be predetermined. Apart from this, the only prerequisites were to obtain as many texts in electronic form as possible, so no time would be lost digitising them, and that they would be of high quality (edited and, if possible, published in written form). Excluding EU documents, as a corpus of such texts already exists,[1] the following text types seemed viable options: legal and administrative texts, promotional texts, journalistic articles from *Le Monde diplomatique* and its Slovenian edition, and literary novels. However, the first two text types had to be excluded for the following reasons: holders of legal and administrative texts were not prepared to share them because of confidential data, whereas the problem with promotional texts for various French products was that most of the available material was translated from English and not French, therefore we could not collect enough material. Wanting the corpus to contain at least two different text types of proportional size for the sake of the comparability of results, we selected journalistic articles and literary novels.

We started collecting texts in November 2007 by sending a letter to the editorial board of the Slovenian edition of *Le Monde diplomatique*, issued in Slovenian since October 2005. A few months later we obtained their permission to include the articles in the corpus and signed afterwards a contract allowing us to use them for non-profit, research-only purposes. Moreover, *Le Monde* kindly provided available copies in electronic form, the rest of the articles being downloaded from the Internet. We had the same experience with the French editorial board, though all the texts were downloaded from the Internet. In the journalistic part of the corpus, we finally included 300 articles from *Le Monde diplomatique* and their translations from *Le Monde diplomatique v slovenščini*, all published between 2006 and 2009 and comprising 1,164,074 words.

Before starting to collect literary novels, we made a survey of the translations from French published in the last 15 years. Wanting to include in the corpus works by as many different authors and translators, as well as publishers,[2] the list of potential novels was fairly moderate. As with the journalistic texts, we first sent a letter (including a contract) to Slovenian publishers of all the selected translations, asking them for permission to include the texts in the corpus and, if possible, kindly provide them in electronic form. Surprisingly, the majority responded quickly and positively: after the signature of the contract by both parties,[3] we were even sent the texts by e-mail. In order to achieve balance between the subcorpora in terms of size, 12 novels were selected for inclusion in the literary subcorpus.

When we submitted the same request to French publishing houses, we were confronted with a problem, already pointed out by many corpus builders, of how difficult it is to get permission from copyright holders. Each publisher received our request accompanied by two letters of support, signed by the director of the Department of Translation and the director of the French Institute of Ljubljana. Only few responded, most of them negatively. Further communication continued via e-mail. After additional explanations and promises to use texts for research purposes only, we received permission from approximately half of the copyright holders. Wanting the corpus to become available to other researchers and interested users, we are still negotiating copyright permission with the remaining publishing houses. It goes without saying that no French novel was acquired in electronic form. The works included in the literary subcorpus span from 1995 to 2008 and total 1,302,911 words.

2.2 Pre-processing and alignment

Since most of the texts were acquired in electronic form, we only had to digitise the French novels and manually correct the scanning errors (particularly punctuation and misrecognised characters[4]), which was quite time-consuming. Once in machine-readable form, we prepared the texts for alignment with *ParaConc*, as it comprises a user-friendly alignment utility[5] and we had already decided initially to run our corpus on this concordancer, available at a low cost and offering everything for the kind of research we wanted to conduct. We first removed possible images, tables, footnotes, endnotes, tables of contents, prefaces, etc. in some translations, and then saved all the files in text-only ANSI format,[6] required by *ParaConc*. Individual parallel texts were then displayed side by side and edited so that they contained the same number of paragraphs. Afterwards, the texts were loaded to *ParaConc* and aligned automatically at sentence level. However, automatic alignment was not 100% correct, therefore manual correction was necessary as parallelism of source and target segments is pivotal for a successful search and analysis. Most of the errors occurred at the level of abbreviations, acronyms and Web site addresses, since the full stops they contained did not indicate the end of a sentence. Moreover, problems occurred when source sentences did not have a corresponding translation, so we had to insert empty lines at those places.

2.3 Annotation

Wanting to allow searches using complex syntactic patterns (*e.g.* detached constructions) and not only specific individual words, it was necessary to annotate the corpus. An expert in this field, Dr Tomaž Erjavec from the Department of Knowledge Technologies of the Jožef Stefan Institute of Ljubljana, kindly agreed to do the work. The texts in both languages were grammatically tagged, *i.e.* every token was assigned a corresponding part-of-speech tag. The French part of the corpus was annotated with *TreeTagger* (see Schmid 1994 and Stein 1994) and the Slovenian one with *ToTale* (see Erjavec et al. 2005).

When the annotation had already been completed, a new tagging system, called *MElt*$_{fr}$ (see Denis and Sagot 2009), was developed for French. Because *TreeTagger* produced some tagging errors, we wanted to test the accuracy of *MElt*$_{fr}$. A French novel was annotated with the new tagger and the results compared with those by *TreeTagger*. Figures 1 and 2 contain the same excerpt annotated with the two taggers, accompanied by a legend explaining the tags. Comparing the results, there is no considerable difference between the two, the error rate (misannotated words are in bold) being approximately the same. For this reason, as well as the fact that the subcategorisation within certain word classes (particularly the verb, which we focus on in our PhD thesis research) is more detailed in the case of *TreeTagger*, we decided not to change the annotation.

<w So>Ici-bas<w PUNCT>, <w Zo>je <w Gps>dépose <w Dp>des <w So>gerbes <w De>de <w So>mots<w PUNCT>, <w V>afin <w V>que <w Ts>ma <w So>liberté **<w R>soit <w Gss>tienne**<w PUNCT>.

<w K>1

<w So>Il <w P>court<w PUNCT>, **<w So>tacle**<w PUNCT>, **<w So>dribble**<w PUNCT>, **<w So>frappe**<w PUNCT>, **<w So>tombe**<w PUNCT>, <w Zo>se <w Gps>relève <w V>et **<w R>court** <w R>encore<w PUNCT>.

<w R>Plus <w R>vite <w PUNCT>!

<w V>Mais <w T>le <w So>vent <w Gps>a <w Gdr>tourné <w PUNCT>: <w R>maintenant<w PUNCT>, <w T>le <w So>ballon <w Gps>vise <w T>l'<w So>entrejambe <w De>de <w Sl>Toldo<w PUNCT>, <w T>le <w So>goal <w P>italien<w PUNCT>.

<w M>Oh <w PUNCT>! <w Ts>Mon <w So>Dieu<w PUNCT>, <w Gps>faites **<w**

Zn>quelque <w So>chose <w PUNCT>!

Legend:

De – preposition
Dp – preposition plus article
Gdr – verb past participle
Gps – verb present
Gss – verb subjunctive present
K – numeral
M – interjection
P – adjective
PUNCT - punctuation
R – adverb
Sl – proper name
So – noun
T – article
Ts – possessive pronoun
V – conjunction
Zn – indefinite pronoun
Zo – personal pronoun

Figure 1. Annotation accuracy of a text, morphosyntactically annotated with *TreeTagger*.

Ici-bas/NPP ,/PONCT je/CLS dépose/V des/DET gerbes/NC de/P mots/NC ,/PONCT **afin/P** que/CS ma/DET liberté/NC soit/VS **tienne/VS** ./PONCT

1/ADJ

Il/CLS court/V ,/PONCT **tacle/NC** ,/PONCT **dribble/NC** ,/PONCT **frappe/NC** ,/PONCT **tombe/NC** ,/PONCT se/CLR relève/V et/CC court/V encore/ADV ./PONCT

Plus/ADV vite/ADV !/PONCT

Mais/CC le/DET vent/NC a/V tourné/VPP :/PONCT maintenant/ADV ,/PONCT le/DET ballon/NC vise/V l'/DET entrejambe/NC de/P Toldo/NPP ,/PONCT le/DET goal/NC italien/ADJ ./PONCT

Oh/NPP !/PONCT

Mon/NC Dieu/NPP ,/PONCT **faites/VPP quelque/ADV** chose/NC !/PONCT

Legend:

ADJ – adjective
ADV – adverb
CC – coordination conjunction
CLR – reflexive clicit pronoun
CLS – subject clitic pronoun
CS – subordination conjunction
DET – determiner
NC – common noun
NPP – proper noun
P – preposition
PONCT – punctuation mark
V – indicative or conditional verb form
VPP – past participle
VS – subjunctive verb form

Figure 2. Annotation accuracy of a text, morphosyntactically annotated with *MElt$_{fr}$*.

2.4 Some statistics

In conclusion of our first part, we present some general statistics about *FraSloK*, produced with *Wordsmith Tools* (Scott 2007).

As shown in table 1, the size of the corpus is 2,466,985 words. The literary subcorpus is slightly bigger because it was difficult to achieve a perfect balance due to the fact that novels are much longer than journalistic articles. Interestingly, both Slovenian subcorpora are smaller, containing approximately one thousand words or tokens (*i.e.* running words) less than the French parts. This difference can be, *inter alia*, attributed to the fact that in Slovenian, contrary to French, nominal phrases are not preceded by articles and personal pronouns are implicit from verbal inflections and should not be explicitly present in surface structure (see note 13).

Table 1. Size of the French-Slovenian corpus (*FraSloK*) and its subcorpora.

	Journalistic subcorpus	**Literary subcorpus**	**Total/language**
French part	637,297	701,715	1.339,012
Slovene part	526,777	601,196	1.127,973
Total/subcorpus (tokens)	1.164,074	1.302,911	
Total/corpus (tokens)	**2.466,985**		**2.466,985**

Tables 2 and 3, on the other hand, show general statistics for individual subcorpora. The first considerable difference between the French and the Slovenian parts is the type-token ratio, referring to the relationship between the total number of running words and the number of different word forms used in a corpus (Olohan 2004: 80). In order to be able to compare corpora of different size, we should take into account a standardised type-token ratio, calculated for every 1,000 words (*ibid.*). In *FraSloK*, the French subcorpora have a noticeably lower type-token ratio than the Slovenian parts, which means there is more repetition and less variety in vocabulary in French texts. This difference could be attributed to the same facts mentioned at the end of the previous paragraph. Interestingly, there is virtually no difference between the French parts in this respect, even though we would expect that the language used in journalistic articles differs from that used in novels.

Table 2. General statistics for the French and Slovenian journalistic subcorpora.

Journalistic corpus	**French subcorpus**	**Slovenian subcorpus**
Tokens	637,297	526,777
Types	38,994	63,514
Type/token ratio	6,21	12,27
Standardised TTR	46,89	60,30
Mean word length	4,98	5,55
Sentences	25,420	24,002
Average sentence length in words	24,71	21,56

The tables of statistics also show values for word and sentence lengths, usually interesting to students studying languages. Comparing individual subcorpora, words tend to be slightly longer in the journalistic subcorpus, particularly in Slovenian. Interesting from a stylistic perspective are the data for the number of sentences. In the literary subcorpus, the number is almost twice as high as in the journalistic subcorpus, but the sentences contain fewer words. The number of sentences in the French and the Slovenian part of the literary subcorpus is more or less equal, whereas the Slovenian part of the journalistic subcorpus contains approximately 1,400 sentences less than the French part, meaning that the Slovenian translators frequently joined the contents of two sentences into one. Last but not least, according to the data in the last rows of tables 2 and 3, French sentences are on average longer, either because of the differences in the language systems, meaning that more words have to be used in French to convey the same meaning, or because there is more simplification (for example due to omissions, ellipsis, etc.) in the Slovenian subcorpora, which would need to be closely examined before coming to any definite conclusions.

Table 3. General statistics for the French and Slovenian literary subcorpora.

Literary corpus	**French subcorpus**	**Slovenian subcorpus**
Tokens	701,715	601,196
Types	41,976	68,919
Type/token ratio	5,99	11,47
Standardised TTR	47,77	58,61
Mean word length	4,54	4,82
Sentences	42,350	42,151
Average sentence length in words	16,55	14,25

3 Case study: Translation of French sentence-initial gerundial clauses into Slovenian

The second part of the paper is centred on a contrastive analysis of French sentence-initial gerundial clauses and their Slovenian translations. Firstly, we define this type of clauses, called ‘detached constructions’ (*constructions détachées*) in French (see also note 10), and explain why they are problematic from a French-Slovenian translation perspective. Secondly, we present the extraction of French gerundial clauses from the French-Slovenian parallel corpus.

Finally, we propose a syntactic and semantic analysis of Slovenian translations in order to discover some recurrent translation patterns or strategies which would be useful in the pedagogic as well as professional translation context.

3.1 Characteristics of French detached constructions

In teaching French grammar at the Department of Translation of Ljubljana, we notice that the majority of students, all being non-francophone, have difficulties in understanding and interpreting French detached constructions, particularly when they are placed at the beginning of a sentence. Let us take an example of a detached construction (underlined) having a gerund as a base (italicised):

(1) *En sillonnant* le pays par la route, on découvre de nombreuses affiches chantant les louanges de la FWO et de l'armée. (*Le Monde diplomatique*, January 2008)
[*On travelling across* the country by road, we discover many billboards singing praise to the FWO and the army.][7]

According to Combettes (1998: 10-13), detached constructions have three main characteristics: i) they can appear in different positions (initial, medial, final) within a sentence and are separated from the main clause by a comma, ii) they act as secondary predicates, as opposed to the main predicate inside the main clause, and iii) they comprise an underlying referent, corefering with the subject[8] of the main clause. Syntactically, then, a detached construction does not contain a subject and a finite verb form in surface structure; its nucleus can be either non-finite (present participle, past participle, gerund) or verbless (adjective, noun). Semantically, the logical relationship between a detached construction and the main clause of a sentence is obscure because the linking device is not explicitly present in surface structure.

Finally, in order for the syntactic and semantic interpretation to be correct, we also have to take into account the intra- and extrasentential context (Combettes 1998: 54) as well as our extralinguistic knowledge and competences (Havu 2002a: 11).

Detached constructions are commonly used in French written texts and have a specific stylistic role: bringing additional information to the referent in question, we condense the wording in order to avoid the complexity of a sentence using, for instance, a semantically equivalent subordinate clause (Combettes 2005: 40). This type of writing seems to be more fluid and dynamic (Jackiewicz *et al.* 2009: 4) and the French, favouring conciseness, use them spontaneously and quite frequently, according to *FraSloK*.

Even though detached constructions exist in Slovenian, they are nowadays rarely used in written texts and appear obsolete in the contemporary Slovenian language. We therefore assume they are not retained in Slovenian when translated from a foreign language, such as French. Referring to example (1), the Slovenian translation would be awkward and archaic if we kept the source language structure:

(1*) *Vozeč se* po deželi z avtomobilom, bomo opazili številne plakate, ki hvalijo družbo FWO in vojsko.

If we wanted the translation to be in line with the (stylistic) norms of the Slovenian language, we would have to use a finite verb form instead of the participle *vozeč se*[9] and join the two clauses by an appropriate linking device. According to context, the relationship between the two clauses could be temporal (1a) or conditional (1b):

(1a) *Ko se* po deželi *vozimo* z avtomobilom, opazimo številne plakate, ki hvalijo družbo FWO in vojsko.

[*When we travel across* the country by road, we notice many billboards praising the company FWO and the army.]

(1b) *Če se* po deželi *vozimo* z avtomobilom, opazimo številne plakate, ki hvalijo družbo FWO in vojsko.

[*If we travel across* the country by road, we notice many billboards praising the company FWO and the army.]

This leads us to advance a hypothesis that Slovenian translations of French detached constructions are syntactically more explicit and semantically more transparent than their French counterparts. In this paper we will focus only on detached constructions having a gerund as a base and appearing in sentence-initial position. In English grammar, these structures correspond to subjectless gerundial clauses (see Blaganje and Konte 2005: 534), therefore we mostly use this term in the rest of the paper. However, when referring to non-finite and verbless clauses at the same time, we prefer using the term ‘detached construction’.

3.2 Extracting sentence-initial gerundial clauses from FraSloK

To our knowledge, there are no commercial or publicly available software enabling automatic extraction of French detached constructions. Moreover, due to very restrictive criteria on what constitutes a detached construction (its position in a sentence, ellipsis of a subject and a finite verb form in surface structure, type of the base (adjectival, nominal, participial, etc.) which can be preceded or followed by one or several adverbs, a conjunction, etc.), it is, as stated by Benzitoun and Caddeo (2005: 307), difficult to establish patterns on the basis of defining criteria of appositions[10] in order to be able to conduct automatic search. For this reason, as well as the fact that the compilation of a suitable corpus, if such does not exist, can be very laborious, the majority of scholars investigating French detached constructions (*e.g.* Combettes 1998, Havu 2003, Neveu 1998) collect the examples manually[11] in printed texts.

For the purpose of this study, the examples of sentence-initial gerundial clauses were extracted from the French part of *FraSloK* using *ParaConc* as this tool enables complex search on the basis of syntactic patterns composed of part-of-speech tags and regular expressions.

In French, a gerund (*le gérondif*) is a non-finite verb form composed of a preposition *en* and a present participle ending in *-ant*. Taking into account all the criteria on what constitutes a detached construction, we composed three patterns[12] to get all the examples of initial gerundial clauses. After a manual elimination of examples which met all the criteria but did not constitute such a clause, we counted 257 examples of left-detached gerundial clauses in the literary subcorpus and 134 in the journalistic one. This means that there are 20% more examples in the literary subcorpus. The difference can be due, *inter alia*, to the text type and length (a journalistic article vs. a novel) as well as to authors' preference for a certain structure. However, we calculated, as a rough guide only, that for the same number of words, the relationship between the number of detached constructions in the journalistic and literary subcorpus is approximately 1/1.3 respectively. In other words, in an article from *Le Monde diplomatique*, containing approximately 2,100 words, we would find 3 such clauses, whereas in a novel we would find 4 for the same length. From this angle, the difference between the number of occurrences in the two subcorpora seems less noticeable.

In conclusion, although gerundial clauses are far from abundant in journalistic and literary texts from our corpus, we should not forget that the data refer only to those occurrences which appear at the beggining of a sentence, before the main clause, and that in linguistics, translation and language pedagogy every word and structure matter.

3.3 Translation strategies

In this section we present the results of a syntactic and semantic analysis of Slovenian translations of French gerundial clauses, extracted from *FraSloK*.

3.3.1 Syntactic analysis of Slovenian translations

All the translations of French sentence-initial gerundial clauses extracted from the parallel corpus were first analysed syntactically in order to find the equivalents of the source structures. As can be seen from Figure 1, the results confirm our conjecture about the very scarce use of detached constructions in Slovenian. In fact, only four (2%) such structures are retained in the literary subcorpus and none in the journalistic one. However, if and when a French detached structure is directly or literally transmitted into Slovenian, the translation (see (1*)) appears incomplete, incoherent and archaic, therefore it would be more acceptable and stylistically appropriate to express the semantic data of the source sentences using other structures, for example subordinate clauses.

Leaving aside the marginal categories 'Detached construction' and 'Other', 95% of French gerundial clauses from the literary subcorpus and 97% from the journalistic one undergo structural changes in the process of translation. Comparing literary subcorpus with the journalistic one, the distribution of translation strategies is quite homogeneous. Almost 70% of all the extracted examples from both subcorpora are rendered into Slovenian by a subordinate clause, wherein the contents of the source gerunds are expressed as a predicate or an attribute. In (2a), for example, the subject of the source main clause is shifted to a subordinate clause (*ameriški izvajalec* or *the American executive*); instead of a gerund the translator used a past simple tense *je zaupal* (*entrusted*) and joined the two clauses with a temporal conjunction *ko* (*when*). In comparison with the source sentence, the word order changes considerably and the translation is much more explicit and thus easily comprehensible.

(2) *En confiant* le « sale travail » à l'Éthiopie, l'exécutif américain a pris le risque de ranimer des braises mal éteintes dans la région. (*Le Monde diplomatique*, November 2007)

[*In entrusting* the »dirty work« to Ethiopia, the American executive risked fanning the flames in the region.]

(2a) *Ko je* ameriški izvajalec *zaupal* »umazano delo« Etiopiji, je tvegal, da se bo v regiji razpihala žerjavica, ki še ni dobro ugasnila.

[*When* the American executive *entrusted* the »dirty work« to Ethiopia, he risked …]

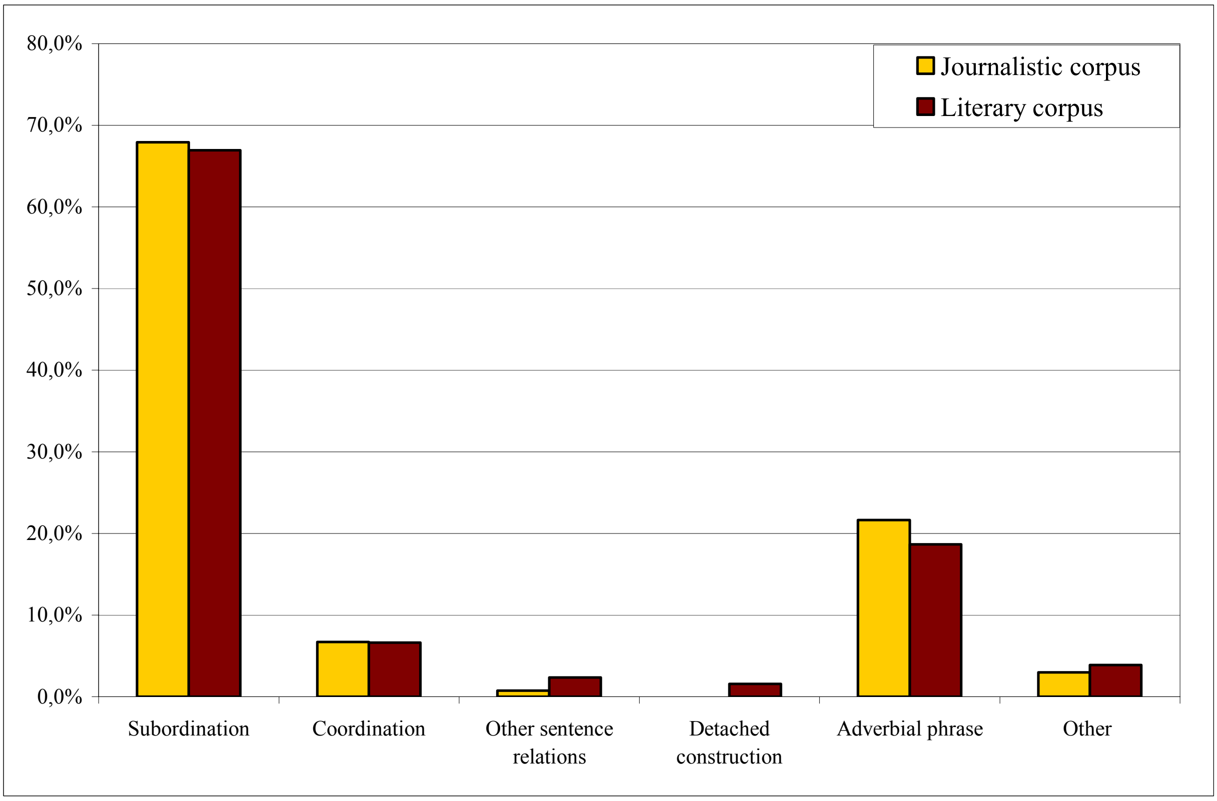

Figure 1. Distribution of Slovenian syntactic equivalents of French sentence-initial gerundial clauses.

The second most noticeable strategy in the Slovenian translations, applicable to a fifth of examples in both subcorpora, is expressing the contents of a gerund with a circumstantial adjunct (Blaganje and Konte 2005: 432-433), as in (3a) where the adjunct of manner *z nasmehom* (*with a smile*) precedes a finite verb form *je strgal*[13] (*he tore off*):

(3) *En souriant*, il arrachait sa tunique, dénouait son pantaloon de soie et dénudait son corps vigoureux. (S. Sa, *Empress*, 2003)

[*Smiling*, he tore off his tunic, undid his silk trousers and stripped naked his vigorous body.]

(3a) *Z nasmehom* je s sebe strgal tuniko, si odvezal svilene hlače in razgalil svoje krepko telo.

[*With a smile* he tore off his tunic, undid his silk trousers and revealed his vigorous body.]

In *FraSloK*, sentence-initial gerunds are sometimes rendered into Slovenian with a finite verb form incorporated into a coordinate clause (7% of examples in both subcorpora). In (4a), the two clauses are joined by the coordinator *in* (*and*) indicating a series of actions:

(4) Puis, *en secouant* sa torpeur, jeta d'une voix rauque : – Qu'est-ce que tu veux qu'on en fasse ? (A. Makine, *The French Testament*, 1995)

[Then, *getting out of* the numbness, he said pointedly: – What do you want us to do about it?]

(4a) Potem *se je zdrznil* iz odrevenelosti in rezko odvrnil: – Kaj pa bi rada, da naredim?

[Then *he got out of* the numbness and said pointedly: – What do you want me to do?]

Very rarely, a finite verb form appears in a sentence where clauses are assembled by means of other relations (category ‘Other sentence relations’), most frequently juxtaposition (1 example in the journalistic subcorpus and 6 in the literary one), where there is no linking device between the clauses. Juxtaposition can be achieved using a comma, as shown in (5a). In this example, we can also note that the word order changed, the subject of the source main clause being shifted to the beginning of the first clause:

(5) *En m’emmenant* trois jours en week-end avec son trésorier et ses dobermans, le directeur de la chaîne a cru me faire passer à jamais le goût de la gaudriole. (M. Darrieussecq, *Pig Tales. A Novel of Lust and Transformation*, 1996)

[*By taking me* for three days to his country house with his treasurer and his Dobermans, the director of the chain thought that I would repress for ever my desire for hanky-panky.]

(5a) Direktor *me je peljal* na vikend z blagajnikom in s svojimi tremi dobermani, bil je prepričan, da me bo razuzdanost za vedno minila.

[The director *took me* to his country house with the treasurer and his three Dobermans, he was convinced that I would get over my hanky-panky for ever.]

4.3.2 Semantic analysis of Slovenian translations

Considering the results of the syntactic analysis of Slovenian translations, the explicitation of a semantic relationship between a detached construction and the main clause of a sentence can only be observed within the categories ‘Subordination’, ‘Coordination’ and ‘Circumstantial adjuncts’, the relationship elsewhere being, as in the source language, implicit or non-transparent. These categories concern 92% of the examples from the literary subcorpus and 96% from the journalistic one. As a semantic interpretation of detached constructions depends on a number of different factors (*e.g.* intra- and extrasentential context, linguistic and extralinguistic knowledge, etc., see, for example, Havu 2002b and 2003) and would necessitate an in-depth analysis and discussion beyond the scope of this article, we present a brief survey of the nature of logical relations explicitated in the Slovenian translations from *FraSloK*.

Within subordination, sentence-initial gerunds most frequently establish a temporal relationship with the main clause, emphasising the simultaneity of two actions. This is generally achieved with a subordinator *ko* (*when*) followed by a predicate or an attribute expressing the contents of a source gerund (see (2)-(2a)). As shown in Figure 2, temporal relationship concerns 54% of subordinate clauses from the journalistic subcorpus and 85% from the literary one. Moreover, subordinate clauses imply other accompanying circumstances to the situation described in the main clause. In the journalistic subcorpus, Slovenian translators also emphasised the relationship of manner (15%), condition (11%), concession (7%) and purpose (4%). The category 'other' involves nominal and relative clauses which do not signal any particular logical or circumstantial relationship but only bring additional or parenthetic information to the referent in question. In (6)-(6a), for example, the source gerund is rendered into Slovenian as a non-restrictive relative clause, functioning as a right-detached postmodifier of the nominal phrase *Ta globalizacija* (*This globalisation*):

(6) *En se développant* par la technique et l'économie, cette mondialisation ne fait que favoriser les régressions identitaires. (*Le Monde diplomatique*, January 2008)

[*Developing* by means of technology and economy, this globalisation only promotes regressions of identity.]

(6a) Ta globalizacija, *ki jo razvijata* tehnika in ekonomija, samo spodbuja istovetnostne regresije.

[This globalisation, *that* the technology and economy *are developing*, only promotes regressions of identity.]

In the literary subcorpus, only the relationships of condition (6%) and manner (5%) are worth mentioning, other semantic values being expressed only once (concession, purpose) or twice (reason).

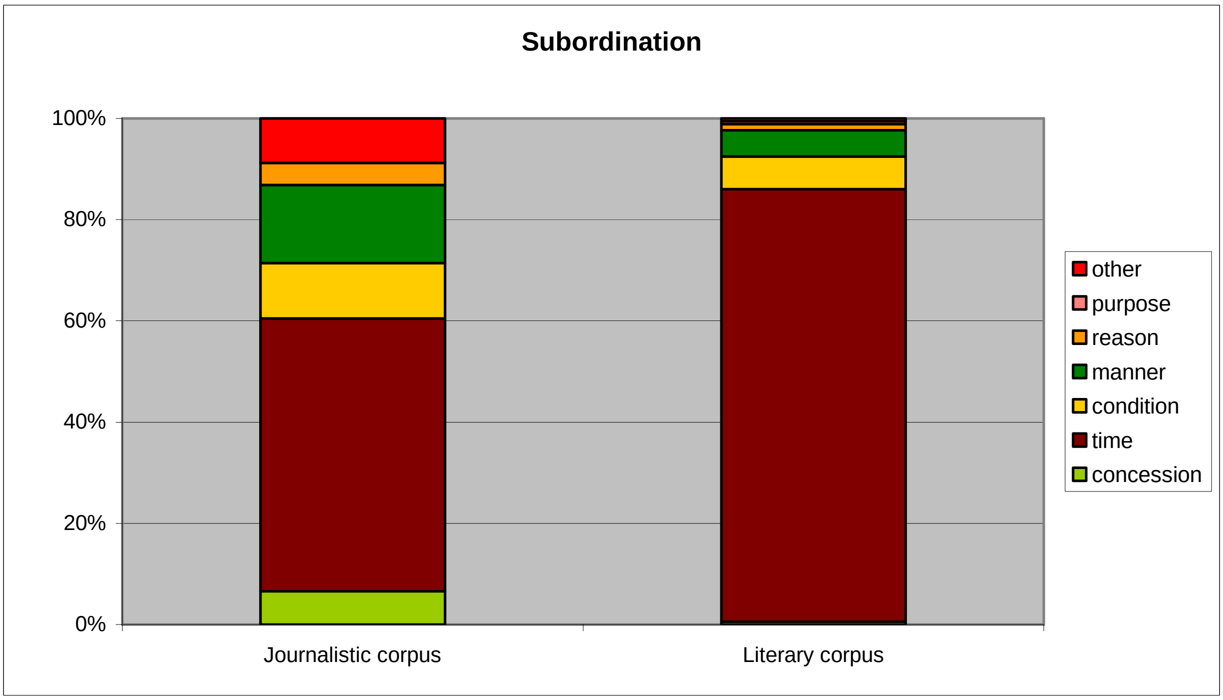

Figure 2. Distribution of semantic values between subordinate and main clauses.

As far as coordination is concerned, only two types are manifested in the Slovenian translations. Conjunctive relationship, mostly signalled by a coordinator *in* (*and*), is overrepresented in both subcorpora (journalistic corpus: 78%, literary corpus: 94%). This relationship is semantically vague in surface structure, simply indicating »that there is some relation between the contents of the linked clauses« (Greenbaum and Quirk 1993: 266). However, the relationship between the clauses can imply different semantic values, which can be inferred from the context. In both subcorpora, coordinators mainly imply a pure addition of the second clause to the first one, though in several instances they also convey a temporal value, indicating a succession of actions or events. The adversative value, on the other hand, signalled by *vendar* (*but*) expressing contrast, appears only twice in the journalistic corpus and once in the literary one.

Finally, contrary to the results for previous categories, the distribution of semantic values expressed by circumstantial adjuncts is quite heterogeneous between the two subcorpora. In the journalistic corpus, the contents of source gerunds are expressed by an adjunct of manner in almost 80% of the examples, other semantic roles appearing only once (place and comparison) or twice (time and reason). In the literary subcorpus, on the other hand, half of the adjuncts express time and approximately one third manner. There are also some adjuncts expressing place (10%), whereas reason and concession appear only once.

4 Conclusion and perspectives

The growing need for language resources enabling different kinds of linguistic research and facilitating langue teaching and learning led us to compile a parallel corpus for the French-Slovenian language combination and thus to address to a certain extent the lack existing in Slovenia in this respect. In the first part of the paper, we tried to present the main design criteria of the corpus and its development process. In sum, the main deficiencies of the corpus are its small size and its limitation to two text types only, as well as its unidirectionality. Nevertheless, it is a valuable resource of authentic data for linguists, translators, foreign-language teachers and students, its added value being a morphosyntactic annotation of all the texts. Recently, a university textbook covering morphosyntactic, semantic and pragmatic aspects of French verbs has been published (see Schlamberger Brezar and Mezeg 2010), wherein the examples in all the exercises were extracted from the parallel corpus presented in this paper. Moreover, new similar projects are already under way. In the future, we intend to enlarge the existing subcorpora of *FraSloK* with new texts and, depending on the availability of sufficient quantities of texts, add to it other text types. Finally, our goal is to make it available to other interested users, therefore we will continue to strive to obtain copyright permission for the remaining French novels.

In the second part of the paper, we showed how *FraSloK* can be used for contrastive and translation studies research. We first presented how complex syntactic structures, such as French sentence-initial gerundial clauses, can be extracted from a parallel corpus, preloaded to *ParaConc*. The Slovenian translations of the extracted examples were then analysed syntactically and semantically in order to find translation strategies that would be useful in the professional as well as pedagogic context.

Even though the analysis was more quantitative than qualitative, it revealed some important findings. Firstly, whereas sentence-initial gerundial clauses appear in peripheral position in French, isolated from the subject of the main clause, in Slovenian it seems more natural and coherent to express their contents with a finite verb form placed within a subordinate and less frequently a coordinate clause, or even with a circumstantial adjunct. Syntactical explicitness is thus evidenced in more than 95% of the Slovenian translations. Secondly, the semantic relationship between a gerundial clause and the main clause, which is implicit in the source language and has to be inferred from the linguistic or extralinguistic context, becomes transparent in the Slovenian translations, particularly within the most represented categories 'Subordination' and 'Circumstantial adjuncts'. Because the relation between coordinate clauses usually remained vague in the translations, semantic explicitness is slightly lower than syntactic, though it still amounts to 85%. On the whole then, the presented results confirm our hypothesis that Slovenian translations of French sentence-initial gerundial clauses from *FraSloK* are syntactically and semantically more explicit than their source counterparts. At least two reasons underlie these findings: grammatical differences between

French and Slovenian and stylistic preferences of the Slovenian language to emphasise syntactic and semantic relations between the clauses.

This case study, however, raises several issues that need to be further explored. First of all, a semantic analysis of the source gerundial clauses would be necessary in order to be able to evaluate logical relationships between the clauses made explicit in the Slovenian texts. Secondly, a future study should focus on the search for criteria or factors influencing the interpretation of a semantic relationship that French sentence-initial gerunds establish with their corresponding main clauses, as this constitutes the main problem for non-francophone students not used to these complex reduced structures. This would enable us to develop methods of decoding such structures, useful in grammar teaching as well as in a professional translation context. Last but not least, an inverse study, based on a Slovenian-French parallel corpus, would also be interesting as it would permit us to examine which structures from Slovenian texts become detached constructions in French, and with what frequency. We could thus test the asymmetry hypothesis (Klaudy 2009; see also Becher 2010) that caused quite a stir during the UCCTS 2010 conference.

References

Barlow, M. (2001). *ParaConc (version 269)*. Houston: Athelstan.

Becher, V. (2010). “Abandoning the notion of ‘translation-inherent’ explicitation: against a dogma of translation studies”. *Across Languages and Cultures*, 11 (1), 1-28.

Benzitoun, C. and S. Caddeo (2005). “Indices linguistiques pour le repérage automatique de certains types d’apposition dans un corpus de presse”. In G. Williams (ed.) *La Linguistique de corpus*. Rennes: Presses Universitaires de Rennes, 307-321.

Blaganje, D. and I. Konte (2005). *Modern English Grammar*. Ljubljana: DZS.

Boch, F., A. Tutin and D. Laurent (2009). "Construction détachée et adjectifs d'affects". In D. Apothéloz, B. Combettes and F. Neveu (eds.) *Les linguistiques du détachement. Actes du colloque international de Nancy (7-9 June 2006)*. Bern: Peter Lang, 99-115.

Combettes, B. (1998). *Les constructions détachées en français*. Paris: Ophrys.

Combettes, B. (2005). "Les constructions détachées comme cadres de discours". *Langue française*, 148, 31-44.

Denis, P. and B. Sagot (2009). "Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort". In *Proceedings of PACLIC*, Hong Kong, China. Available at: http://alpage.inria.fr/~sagot/pub/paclic09tagging.pdf (accessed 14 October 2010)

Erjavec, T., C. Ignat, B. Pouliquen and R. Steinberger (2005). "Massive multi-lingual corpus compilation: Acquis Communautaire and totale". In Z. Vetulani (ed.) *Proceedings of the 2nd Language & Technology Conference (Poznań, 21-23 April)*. Poznań: Wydawnictwo Poznańskie Sp. z.o.o, 32-36.

Forsgren, M. (1991). "Éléments pour une typologie de l'apposition en linguistique française". In D. Kremer (ed.) *Actes du XVIIIe Congrès International de Linguistique et de Philologie Romanes*. Tübingen: Max Niemeyer Verlag, 597-612.

Greenbaum, S. and R. Quirk (1993). *A Student's Grammar of the English Language*. Essex: Longman.

Havu, E. (2002a). "L'interprétation des constructions détachées". *Circulo de Lingüística Aplicada a la Communicación 10*. Available at: http://www.ucm.es/info/circulo/no10/havu.htm (accessed: 14 October 2010)

Havu, E. (2002b). "Sur quels principes l'interprétation des constructions détachées repose-t-elle ?". In D. Hallvard (ed.) *Actes du XVe congrès des romanistes scandinaves (Oslo 12-17*

August 2002), 389-400. Available at: http://www.duo.uio.no/roman/Art/Rf-16-02-2/fra/Havu.pdf. (accessed: 14 October 2010)

Havu, E. (2003). "Comment interpréter les constructions détachées initiales ?". In J. Härmä (ed.) *Le langage des médias : discours éphemères ?*. Paris: Harmattan, 19-38.

Jackiewicz, A., T. Charnois and S. Ferrari (2009): "Jugements d'évaluation et constituants périphériques". In *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles (Senlis, 24-26 June 2009)*. Available at: http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_122.pdf (accessed: 14 October 2010)

Klaudy, K. (2009). "The asymmetry hypothesis in translation research". In R. Dimitriu and M. Shlesinger (eds.) *Translators and Their Readers. In Homage to Eugene A. Nida*. Brussels: Les Éditions du Hazard, 283-302.

Neveu, F. (1996). "La notion d'apposition en linguistique française". *Le Français moderne*, LXIV, 1-27.

Neveu, F. (1998). *Études sur l'apposition. Aspects du détachement nominal et adjectival en français contemporain, dans un corpus de textes de J.-P. Sartre*. Paris: Honoré Champion.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London, New York: Routledge.

Rossi-Gensane, N. (2009). "Register variation in the non-standard use of non-finite forms". In K. Beeching, N. Armstrong and F. Gadet (eds.) *Sociolinguistic variation in contemporary French*. Amsterdam, Philadelphia: John Benjamins, 177-191.

Schlamberger Brezar, M. and A. Mezeg (2010). *La morphosyntaxe et la sémantique du verbe français*. Ljubljana: Filozofska fakulteta.

Schmid, H. (1994). *Treetagger*. University of Stuttgart. Available at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger (accessed: 14 October 2010)

Scott, M. (2007). *Oxford Wordsmith Tools 4.0.* Oxford: Oxford University Press.

Stein, A. (1994). *TreeTagger pour l'ancien français et le français moderne*. University of Stuttgart. Available at: http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html (accessed: 14 October 2010)

Wilmet, M. (1997). *Grammaire critique du français*. Louvain-la-Neuve: Hachette Supérieur, Duculot.

Notes

[1] It is available as part of *Evrokorpus*, consisting of EU texts for several foreign languages in combination with Slovenian. For more details, see http://evrokorpus.gov.si/index.php.

[2] We found it unlikely that individual publishers would be willing to give us permission for more than one book.

[3] That is the directors of the publishing houses and the dean of the Faculty of Arts of Ljubljana.

[4] For example, 'm' was often recognised as 'rn'.

[5] Other commercial automatic alignment tools were at our disposal (*e.g. WinAlign* and *Atril Déjà Vu*) but after testing them, we found them less user-friendly than *ParaConc*.

[6] At this stage, no markup was encoded into the files. However, when Dr Tomaž Erjavec from the Jožef Stefan Institute agreed to tag the corpus, he also kindly converted them into Unicode and marked them up in line with TEI guidelines. All the texts thus include a header and are segmented, as well as lemmatised. At this occasion, I would like to thank Dr Erjavec for the work done.

[7] The examples in square brackets are literal English translations.

[8] The underlying referent of a detached construction can, however, corefer with the object of the main clause (*e.g.* *Vêtu* d'un costume d'aviateur, on **le** vit s'extraire du cockpit… – object in bold (*Le Monde diplomatique*, December 2007) [*Dressed* in a flight suit, we see **him** leaving the cockpit…]). For other non-standard uses, see Rossi-Gensane 2009: 177-190.

[9] The French gerund *en sillonnant* cannot be translated into Slovenian as a gerund (**vozé se*).

[10] Several French linguists use the term *apposition* (*e.g.* Neveu 1996 and 1998, Forsgren 1991) instead of *detached construction* (*construction détachée* in French). However, since it has long been debated and because some linguists (*e.g.* Wilmet 1997) classify among them certain structures quite different from the detached constructions described above, we prefer the term put forward by Bernard Combettes (*i.e. construction détachée* or *detached construction*), which is nowadays becoming increasingly used in French linguistics (*see* Havu 2003, Boch *et al.* 2009, Jackiewicz *et al.* 2009).

[11] Automatic extraction of some types of French detached constructions is, however, used by Boch *et al.* (2009) and Jackiewicz *et al.* (2009).

[12] To take an example, the most productive pattern (<w \w+>En(\W<w \w+>\w+){0,1}(\W)?<w Gds>\w+) searched the examples starting with the preposition *en*, followed by zero or one word and an obligatory present participle.

[13] In Slovenian, a grammatical subject is evident from the inflexion (*je strgal* (*tore off*) → *on* or *he*) and should not be repeated with a personal pronoun, whereas in English and French a personal pronoun has to be explicitly present in surface structure (*he tore off* or *il arrachait*) when mentioned in the sentence for the first time.