# PARALLEL CORPORA AS TEXTS IN A TSP CONTEXT AN ARGUMENTATIVE APPROACH

**Daniela Ionescu**

**Abstract:** The present paper aims at proposing several ways of selecting and typologising specialized texts according to different criteria, such as thematic content, genre criteria, or structural discursive differences. This approach follows the current views on discourse structure and the theory of parallel corpora. The texts that make up a corpus of targeted discourse pieces share several properties but they also serve a different purpose, which is the author's communicative intention to make the information or comments structured in the text relevant in some specific way to readership. It is precisely this relevant markedness of such texts that constitutes the core of the present paper.

## 1. The argumentative dimension

The present paper discusses language as a set of functional resources in use. This involves both looking at texts and consequently, at their constituents, as objects of study in their own right and the study of language structures and functions in order to look for regularities which might help the understanding of language events. In two or several languages the set(s) of texts that make up a given genre and belong to a certain conceptual domain can be analysed, compared and contrasted intra- and interlinguistically. Such an approach proves significant for translators who regularly operate with different corpora and different genres when translating texts. As translators do not receive formal training in any particular LSP (language for special purposes), they must set out conscious learning of the LSP lexical and non-lexical – such as style, punctuation, grammar, register - elements of the language as well as of the LSP conceptual knowledge (Bowker and Pearson 2002) in order to produce acceptable texts for the target audience. In this sense, contrastive studies and corpus-based studies, as that presented below, might be of help for the understanding and writing of genres in target language, in the act and as a result of translation.

In descriptive translation studies, corpus-based approaches have been used to investigate whether and how translations differ from their source texts, or from original texts written in the target language, and how specific languages and genres as well as translators' stylistic preferences affect translations. Beginning in the early nineties (following Baker 1993), work in this area has drawn on and brought together aims and methods from descriptive translation studies and descriptive corpus linguistics (Sinclair 1992). Corpus linguistics, as a methodology which focuses on the identification of recurrent patterns of linguistic behaviour in actual performance data, provides the appropriate tool to test hypotheses about norms and regularities in translated texts. Within this kind of analysis, stress is being laid on those regularities that translate adequately in the target language. Given its emphasis on the target side of translation, a paradigm can be constructed on the basis of comparable target language corpora, i.e. collections of translated texts and of original texts in the same language (Olohan and Baker 2000), as well as on parallel or translational corpora. Corpora are widely used in translation studies classes, where corpus work can prove relevant prior, during or after a translation task (Aston 2000). The main methodological aim for teaching translation through corpora is to evince the salient textual features of the original corpus and then find proper equivalent formulae for the target language corpus, according to the genre characteristics as well.

**2. Working definition of corpora**

Language Corpora are principled collections of texts in electronic format; they are increasingly being used as a resource in linguistics and language related-disciplines and corpus linguistics is now firmly established as a research area and a methodology. One of the many fields where corpora are having a growing impact is translation, both at a descriptive and a practical level. This paper is principally concerned with the use of corpora as resources for the translator and as teaching/learning aids in the context of the translation classroom – an area which should be termed "applied corpus-based studies", terms of the 'map' of translation studies elaborated most notably by Toury (1995).

The use of corpora in language learning contexts was pioneered by Johns, who introduced the principle of concordancing into the foreign language classroom in the '80's. Besides enabling language professionals such as lexicographers and terminologists to produce better reference and learning materials, and allowing language teachers to create classroom activities based on real examples, he showed how corpora could provide learners with direct access to virtually unlimited language data (Johns 1991). This approach is in line with current views on the value of autonomy, motivation and authenticity in language teaching (Zanettin & al. 2003), as well as with current theories of translator education as a process of socialization in a professional community (Kiraly 2000).

**3. Typology and structure of corpora**

What kind of corpora are currently being used in teaching and learning contexts?

Monolingual, bilingual, comparable, parallel, bidirectional, reciprocal, virtual, do-it-yourself (DIY), general (large), specialized (small), and reference (very large) corpora.

Perhaps the most familiar are the monolingual corpora (whether general or specialized, usually in the target language), comparable bilingual corpora (originals in two languages selected according to analogous criteria such as topic and text type), and parallel corpora (=originals in one language and their translations in another).

A matter of terminology: it is not consistent, in that the term "parallel" has been used to indicate both those corpora we refer to as "comparable" (similar originals from two languages) and those which refer to "parallel" (originals and their translations). The latter type has also been called "translational corpora". The ultimate goal of corpora thesaurus is the provide a competent use of text and corpora and text analysis so as to enable students to become better language professionals in a working environment where computational facilities for processing texts have become the rule rather than the exception. Therefore, applied corpus-based translation studies should be the framework for research and pedagogy. Like corpus linguistics (Svartvik 1992), this area has come "of age" and we may consider it particularly central to the way this field of study is progressing (e.g. the growing areas of corpus-based translation studies, computer-aided translation, corpus-based language teaching, etc. Corpus linguistics as a methodology which focuses on the identification of recurrent patterns of linguistic behaviour in actual performance data provides the appropriate tool to test hypotheses about norms and regularities in translated texts.

**4. Definition and terminological/notional delimitations**

The monolingual corpora can provide information about typical 'units of meaning' in the target language or in a specialized subset of it (restricted by topic and/or text type). They provide a deeper insight into 'native-like turns of phrase, appropriate to the communicative situation and in which the target text will be operating. Monolingual corpora in second language learning or language learning for a specific purpose are an example of comparable corpora used extensively in translation research and in the translator training environment.

The comparable bilingual corpora – can provide future translators with a better understanding not only of target but also of source texts, allowing them to compare terminology, phraseology and textual conventions across languages and cultures. While monolingual corpora can be large and general in scope, bilingual corpora are usually limited and specialized. Comparable corpora is a term used by Baker (1995: 234) to refer to "two separate collections of texts in the same language" of which "one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages". Their name – comparable - shows precisely the aim for which they are constituted: to identify and extract those contrastive characteristics, linguistic and extra-linguistic which differentiate them form other types and which contribute to a better understanding and practical use of those structures, features in similar contexts, for the creation and re-creation of new texts. The fact that comparable corpora are monolingual collections of texts distinguishes them from other types of Corpora used in Translation Studies. The information which they contain is likely to yield rich insights into the kind of linguistic features which are typical of translated text, regardless of the language of ST. Little work has so far been carried out with comparable corpora. Hence, a certain degree of fluidity and approximation regarding the approach to this kind of texts included into the comparable corpus or corpora.

The parallel or bilingual corpora may again be general or specialized, they can also be indicative of certain translational strategies that translators generally use – for instance, situationally-constrained expressions or lexical creativity reflected in translation. According to Baker (1995: 230), a type of corpus consists "of original, source language-texts in language A and their translated versions in language B". Parallel corpora can be used in materials writing, translator training and the development of machine translation systems. However, their advantage is that they provide information not on the native patterns of a target language, but on those of specific target texts, and so give insight into the particular translation practices and procedures which have been used by the translator.

The bidirectional parallel corpora or reciprocal corpora are the most comprehensive, in the sense that they provide the opportunity to use two sets of parallel corpora containing two source texts and translations in opposite directions. Thus, textual features can be compared (in the translated versions) with the original texts in both languages, or in several, if a larger comparative project is to be carried out.

## 5. Corpora and translation

The present paper will look more closely at the comparable and parallel collections of texts (corpora), as they can prove to have complementary roles to play in the translation classes. Generally speaking, texts in comparable corpora (CC) are originals, i.e. they have not been translated from another language. They provide evidence of language behaving naturally in a monolingual situation. Parallel corpora (PC), on the other hand, contain texts and their translations. Thus, they contain evidence not only of language produced in a monolingual environment (the source texts), but also of language produced in a bi- or multilingual environment (the translations). Evidence is thus being collected – and this is the rich realm of debates, workshops and essays or extensive translational and terminological projects – that the language used in translation may differ from the language used in the production of an original source text (Laviosa 1997).

Thus, investigations of PC may allow students to see how writers, i.e. translators behave when constrained by the existence of a text composed in another language. Translators have

to act as cultural and linguistic mediators, negotiating their way between languages and between cultures. They have to gauge how much of the material in the source text is directly transferable to the target language, how much of it needs to be adapted or localized in some way, whether any of it can, or indeed should, be omitted. The answer to questions of this nature cannot be found in comparable corpora because these issues never arise in a monolingual text-producing environment. They only arise because of the constraints of a text composed in another language. So, the answers must be sought elsewhere, in PC. By studying PC, particularly aligned PC, translation students have the opportunity to see for themselves what strategies professional translators employ to solve different translation problems. They can learn how info is conveyed, whether any info is lost, adapted or misrepresented in the process. By observing and discussing what has happened, they can begin to devise their own translation strategies.

**5.1 Methodology**

In the present analysis of texts I will include the discussion of the way in which certain discourse features are realized in two different types of texts: scientific, medical research and press texts. Obviously, the aim of this analysis is to show to what extent and in what way these features are present while still differently realized in all or only in some of the texts under grammatical scrutiny. In this attempt, I follow Smith's (1991 and 2003) linguistic guidelines in discourse analysis (and her latest view on Discourse Representation Structures (DRT).

For instance, the entities introduced in the discourse and the aspectual information along with temporal and spatial orientation elements will be comparatively viewed in the three types of texts: scientific, medical and journalese. All observations will hold at passage and paragraph level so as to be closely identified and interpreted in a further analysis of translational approach to the same texts.

Below are the main passages subject to comparative discourse analysis. All these passages have been selected from the introductory part of the articles. The medical excerpt includes practically the whole introductory part which prefaces the article proper, i.e. the abstract format.

Argumentatively, any introductory passage would posit the core idea and focus on it gradually, so as to spotlight the relevant aspect of the issue. In the different type texts, we can notice that despite the apparent discourse similarities (explicit introductory lexical phrases, e.g. "background", "results", "conclusions"(medical/scientific discourse), temporal adverbs ("in 1994"), use of abstract and mass nouns throughout or systematic omission of the determiner in nominal phrases, extensive use of the passive voice, etc., we can also detect a certain number of argumentative, specific or text-bound features realized through grammatical structures, whether syntactic or lexical, or both – in nature.

For instance, the presentational format (background, methods, results, conclusions) is typical of medical reports, articles, clinical studies or trials. Each one of these structural textual elements is a functional unit which has got a certain goal within the entire discourse.

In scientific and medical texts, these units are somewhat rigidly structured according to recognized conventions. In the scientific text below, for instance, the main idea is symbolic representations of the world in early modern humans. The textual structure is hierarchical in that the constituting units are discourse-related, i.e. entities and propositions may consist of situations of Cause and Result, or rhetorical relations such as Elaboration, Evidence, Parallelism, etc. Such relations may also occur in the press texts, where cause-result or

parallelism are viewed as real sources of stylistic effect in the journalese register. In the paragraph below (3), the very first sentence is based on parallelism, while the fourth – is based on Evidence. Parallel structures are meant to ensure continuity (discursive progress, cf. Smith 2003) as it establishes a kind of referential connection between the various items of the sentence or sentences. In the press paragraph (3), continuity is realized epiphorically rather than anaphorically (e.g. *Alaska is famous for big bears, big salmon, big mountains; and, increasingly, big legal trouble for its politicians*).

The obviously different features of each type of text are the lexical items, the terms specific to the domain, and their frequency or constant use in the discourse, with or without explicit reference. The syntactic structures including simple, plain word order structures, repetitive phrases (as in paragraph (1)), compound-complex phrases are typical of medical texts, while the use of mass nouns or bare plural nouns and nouns with no determiner (determiner omission) are common both to medical and scientific texts.

### 5.2 Topic/Focus structure and text typology

Typically, the topic referent of a sentence is in subject position and represents familiar information. The subject of a sentence is prominent positionally and grammatically (Smith 2003: 192). As first element it links directly to what precedes and by extension to the common ground; it is the starting point for the communication of a sentence. This is the aboutness of a sentence, while the focus phrase of a sentence is the speaker's declared contribution to the common ground. The focus of a sentence is its contribution to a discourse, it reflects the speaker's decision as to where the main burden of the message lies (Halliday 1967, in Smith 2003: 202). The features of a focus or the way in which it is structured also depends on the type of text, for instance, the scientific texts organize focus on the basis of the readers' information competence in the domain, while the press texts heavily rely on the readers' background attitude and idiosyncrasies with respect to the issue under debate, whether political, social or economic, or just for entertainment. In he medical-scientific text below, the cardioverter defibrillator therapy is the topic of the whole article, while the survival of patients is the focus pursued by the physicians in their research work and tests. In the press texts (3 and 4) the topics are politicians and Hispanics and blacks living in the US respectively, while the corresponding focus is either evidence of corruption (of politicians) or the accumulating racial conflict between two ethnic populations. In both cases, (paragraphs 3 and 4), the writer builds on the readers' attitude towards these issues, political and social. It is not knowledge or conceptual expertise, but knowledge of the world that determines the prominent topic and emphatic and contrastive focus alike.

Medical:

(1) *Background*

*Implantable cardioverter-defibrillator (ICD) therapy has been shown to improve survival in patients with various heart conditions who are at high risk for ventricular arrhythmias. Whether benefit occurs in patients early after myocardial infarction is unknown.*

*Methods*

*We conducted the Defibrillator in Acute Myocardial Infarction Trial, a randomized, open-label comparison if ICD therapy (in 332 patients) and no ICD therapy (in 342 patients) 6 to 40 days after a myocardial infarction. We enrolled patients who had reduced left ventricular function (left ventricular ejection fraction, 0.35 or less) and impaired cardiac autonomic function (manifested as depressed hear-rate variability or an elevated average 24-hour heart-*

*rate on Holter monitoring). The primary outcome was mortality from any cause. Death from arrhythmia was a predefined secondary outcome.*

*Results […]*

*Conclusions*

*Prophylactic ICD therapy does not reduce overall mortality in high-risk patients who have recently had a myocardial infarction. Although ICD therapy was associated with a reduction in the rate of death due to arrhythmia, that was offset by an increase in the rate of death from non-arrhythmic causes.*

(Abstract of 'Prophylactic use of an implantable cardioverter-defibrillator after acute myocardial infarction', *The New England Journal of Medicine*, December 2004, vol. 351, no. 24)

Scientific:

(2) *In 1994 discovery of France's Grotte Chauvet revolutionized ideas about symbolic expression in early modern humans. The breathtaking drawings of horses, lions and bears that adorned the cave walls were executed with perspective and shading and rivaled the virtuosity of all other known cave art. But when were those drawings made? Early radio carbon dates suggested 32,000 years ago right after a major cold spell hit Europe. This implied that modern humans blossomed under frigid conditions while their Neandertal cousins were going extinct. But improved radiocarbon dating now suggests that the oldest paintings at Chauvet could be at least 36,000 years old. That's smack in the middle of a period of relative warmth and challenges speculation about modern humans' adaptability to a cold climate.*

('Radiocarbon dating's final frontier, *Science*, September 2006)

Journalese:

(3) *Alaska is famous for big bears, big salmon, big mountains; and, increasingly, big legal trouble for its politicians. The state's lone congressman, Don Young, a Republican, is being scrutinised by the FBI for links to an Alaskan company whose two top executives bribed state legislators. Lisa Murkowsky, the state's Republican junior senator, made headlines in mid-July when it was suggested that she had purchased property on the Kenai river without disclosing the transaction, as she was legally bound to, in her annual financial statement.*

('United States/Investigating Alaska', *The Economist*, August 2007)

(4) *Two men will soon stand trial in Los Angeles in a murder case that does not involve white cops, a sportsman or a music producer. As a result, the trial is unlikely to receive minute-by-minute coverage on cable TV. Yet it will reveal as much about the edgy state or race relations in Los Angeles as the cases of Rodney King or O. J. Simpson. Perhaps more so, since it involves the two groups between which there is most tension. The accused men, Ernesto Alvarez and Jonathan Fajardo, are Hispanic. The victim, 14-year-old Cheryl Green – who, prosecutors say, died in a racially motivated attack - was black.*

('The United States/Where black and brown collide', *The Economist*, August 2007)

**5.3 DRS in two types of texts: Medical (1) and Press (4)**

As to the way in which the discourse representation structure is organized in two different texts, the author is generally interested in rendering the most prominent or relevant facet of the issue, that is why the participants in the event and the event proper are presented like a constellation around which the subjective dimension of the text is constructed (also see Smith, 2003). The subjective dimension includes communication, contents of mind and evidentiality,

perception and perspective. In the conclusive part of text (1), the expressions of subjectivity are highly neutral in that they include a plain negative statement which is reinforced by the argument described in the second sentence, through a concessive counterbalance. This last sentence contains no experiential, evaluative or evidential adverb, instead, it builds up a chained assertive perspective on the situation. In fact, no epithets are to be found in this last part of the medical text, nor are they generally frequent in scientific texts either.

In paragraph (4), the DRS would be:

the two men: –x, y; operator: will; e: stand trial; situation: a murder case; negation: white cops, a sportsman, or a music producer (effect: irony);

The communication rule introduces the author who reports the situation: the trial of the accused men Ernesto Alvarez and Jonathan Fajardo for having killed a 14-year old black boy, Cheryl Green. The case will be publicized just because of the accumulated tension between the two ethnic groups. Subjectivity is made apparent especially through the adjunct parts of sentences: *as a result*, *yet*, *perhaps more so*, by epistemic predicates: *is unlikely* or by evaluative adjectives: *edgy state*.

## 6. Conclusions

Text analysis and text collection will necessarily lead to text corpora with a view to finally providing translation corpora consisting of original and translated texts in given languages, whereas parallel corpora will include text collections of authentic texts in the given languages. However, it is the source material that defines the type of texts to be included in the compiled corpus. Computational analysis findings facilitate and refine text analysis based on grammatical, syntactic and pragmatic features. Through this approach, we can provide better knowledge and understanding of text construction and translation, for a multidisciplinary and multidimensional perspective in the creation and variability of professional discourse.

Daniela Ionescu
University of Bucharest
daniones@gmail.com

## References

Aston, G. (2000). *The Learner as Corpus Designer*. University of Bologna: Advanced School of Interpreters and Translators.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target* 7 (2): 223-243.

Bowker, L. and Pearson, J. (2002). *Working with specialized language: A Practical Guide to using corpora*. London: Routledge.

Johns, T. (1991). Should you be persuaded - two samples of data-driven learning materials. In T. Johns and P. King (eds.), *Classroom Concordancing*, 1-13. Birmingham University: ELR.

Kiraly, D. (2000). *A Social Constructivist Approach to Translator Education*. Manchester: St.Jerome.

Laviosa, (1997). How comparable can 'comparable corpora' be?. *Target* 9 (2): 289-319.

Olohan, M. and M. Baker (2000). *Intercultural Faultlines. Research models in Translation Studies 1: Textual and Cognitive Aspects*. Manchester: St. Jerome.

Sinclair, J. M. (1992). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smith, C. (1991). Sentences in texts: A valediction for sentence topic. In C. Georgopoulos and R. Ishihara (eds.), *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda*, 545-564. Dordrecht: Kluwer.

Smith, C. (2003). *Modes of Discourse*. Cambridge: Cambridge University Press.

Svartvik, J. (ed.) (1992) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton de Gruyter.

Toury, G. (1995). Handful of Paragraphs on 'Translation' and 'Norms'. Ms., University of Tel Aviv.

Zanettin, F., Bernardini, S., Stewart, D. (2003). *Corpora in Translator Education*. Manchester: St. Jerome.