Text Structure in a Contrastive and Translational Perspective

On Information Density and Clause Linkage in Italian and Danish

Iørn Korzen
Copenhagen Business School
ik.ibc@cbs.dk

Morten Gylling
Copenhagen Business School
mg.ibc@cbs.dk

This paper argues that both human translators and machine translation systems can greatly benefit from contrastive studies of text structure. Due to the great terminological and definitional confusion regarding structures in texts, the paper first discusses the main viewpoints on these issues and then outlines the two most significant differences between Italian and Danish text structure. One regards the notion of information density: Italian tends to accumulate the same information in shorter text spans and to include a larger number of Elementary Discourse Units in each sentence than Danish. The other regards clause linkage: A higher percentage of Italian clauses is morpho-syntactically and rhetorically subordinated by means of non-finite and nominalised verb forms. Danish text structure, on the other hand, is more informationally linear and characterised by a higher number of finite verbs and topic shifts. These typological differences are transferred into some simple translation rules concerning the number of Elementary Discourse Units per sentence and their textualisation. Each rule is illustrated by a number of examples taken from the parallel part of the Europarl Corpus.

1 Introduction[1]

It has been pointed out that in some aspects Translation Studies (TS) and Contrastive Linguistics (CL) overlap each other. Some scholars talk about "common grounds" between the two, for instance, that they are both *"interested in seeing how 'the same thing' can be said in other ways, although each field uses this information for different ends"* (Chesterman 1998, 39). Another more recent and methodological common ground is the emergence of corpora that serve as empirical bases for TS and CL research (Granger 2003).

However, in spite of such "common grounds", the fact remains that only a limited number of TS scholars have applied CL to their research in order to obtain an awareness of systematic differences between two or more language systems. Vice versa, only a few CL scholars take advantage of TS knowledge of translation norms and strategies. This paper aims to illustrate how TS can benefit from CL findings, as a sort of response to Chesterman's appeal (1998, 6):

> Although these [TS and CL] are neighbouring disciplines, it nevertheless often appears that theoretical developments in one field are overlooked in the other, and that both would benefit from each other's insights.

[1] We thank our colleague Daniel Hardt for his useful comments on an earlier draft of this paper and the Danish Council for Independent Research in Humanities for financial support.

Our point of departure is CL, but where both CL and TS typically confine their investigations to lexical and syntactic levels, we will focus on the textualisation and structure of larger text segments, i.e. segments beyond the boundaries of clauses and sentences.

Structures in texts have been massively investigated from different angles and under different terms: discourse structure, text structure, rhetorical structure, information structure, temporal structure, etc.[2], and particularly between the three first mentioned terms, discourse, text and rhetorical structure, there is great terminological and definitional confusion and overlap. In section 2, we shall therefore briefly examine a few of the most important viewpoints regarding these issues and outline the definitions chosen in our own research. In section 3, we shall describe Italian and Danish text structure on the basis of a relatively large comparable text corpus, the Europarl Corpus, whereas section 4 will be dedicated to the perspectives of our research results for translation, supported by a number of "felicitous" Italian-Danish translations from the parallel Europarl texts.

2 Structures in texts

2.1 Discourse, text and rhetorical structures

Especially discourse (structure) and text (structure) have been subject to different definitions in the literature. For instance, in the *International Encyclopedia of Linguistics*, Chafe notes:

> The term *discourse* is used in somewhat different ways by different scholars, but underlying the differences is a common concern for language beyond the boundaries of isolated sentences. The term *text* is used in similar ways. Both terms may refer to a unit of language larger than the sentence: one may speak of a "discourse" or a "text". (Chafe 2003, 439-440)

Irmer (2011, 43) is a recent example of a scholar who similarly uses the terms interchangeably: *"Generally, a* text *or a* discourse *is a sequence of natural language utterances."* The same viewpoint is found earlier in Stubbs, who adds:

> Sometimes this terminological variation [between text and discourse] signals important conceptual distinctions, but often it does not, and terminological debates are usually of little interest. (Stubbs 1996, 4)

Halliday and Hasan use both terms in their definition of a "text":

> A *text* has texture and this is what distinguishes it from something that is not a *text*. … The texture is provided by the cohesive RELATION [which is set up] where the INTERPRETATION of some element in the *discourse* is dependent on that of another. (Halliday and Hasan 1976, 2-4; our italics)

Similarly, we find in Rijkhoff (2008, 90) both terms, discourse and (co-)text, used for linguistic material, as the scholar affirms*: [D]iscourse in the sense of co-text is a linguistic entity*.

On the other hand, there are scholars according to whom "text" refers to written language and "discourse" to spoken language. For instance, Stubbs (1983, 9) notes

[2] See Hoey (1991) for a discussion of *structure* vs. *organisation* of texts.

that *"One often talks of 'written text' versus 'spoken discourse'"*, and similarly Riazi states:

> The first [approach] is "discourse analysis," which mainly focuses on the structure of naturally occurring spoken language, as found in such "discourses" as conversations, commentaries, and speeches. The second approach is "text analysis", which focuses on the structure of written language, as found in such "texts" as essays and articles, notices, book chapters, and so on. (Riazi 2002, 4)

However, this distinction is rejected by other scholars:

> Discourse … refers to both text and talk, and these not as two separate genres to be compared and contrasted, but rather as overlapping aspects of a single entity. As the object of study, spoken discourse is 'text', much as words spoken in a speech are commonly referred to as the text of the speech. In this sense, 'discourse' and 'text' are synonymous. (Tannen 1982, ix)

> In non-linguistic and non-semiotic circles, text is sometimes used for examples of written language and discourse for the spoken. Nowadays linguists accept that such a distinction based only on medium and channel is simplistic. (Christiansen 2011, 34)

A third group of scholars see discourse structure as the rhetorical organisation of a text (Mann and Thompson 1988), organisation definable as a series of "coherence relations" (Hobbs 1985) between text segments, created in the process of human communication (Brown and Yule 1983, 24-26; Widdowson 1979, 71). Widdowson (2004), who overtly criticises Harris (1952) and Stubbs for conflating the terms "text" and "discourse" (op. cit., 4-5), states more precisely:

> Discourse in this view is the pragmatic process of meaning negotiation. Text is its product. ... The discourse may be prepared, pre-scripted in different degrees. ... But whatever the degree of prescription, the text, the actual language that realizes the interaction, is immediate to it, and is directly processed on line. (op. cit., 8-9).

Seemingly inspired by Halliday and Hasan and their claim (1976, 300): *"discourse … come[s] to life as text"*, Christiansen (2011, 34) similarly states that "*text [is] the form, discourse the content*", and along the same line, Cornish (2009, 99-100) concludes his definition of the two terms in this way:

> Text, then, refers to the connected sequences of signs and signals, under their conventional meanings, produced by the speaker … Discourse, on the other hand, refers to the hierarchically structured, mentally represented product of the sequence of utterance, propositional, illocutionary and indexical acts that the participants are jointly carrying out as the communication unfolds … Text, in normal circumstances of communication, on the other hand is essentially linear, due to the constraints imposed by the production of speech in real time."

Similarly, Ruiz Ruiz (2009, 4) remarks:

> [T]he two concepts [discourse and text] should not be confused or equated. Indeed, every piece of discourse has a textual form or can acquire it; the same text may include different discourses or the same discourse may adopt different textual forms.

In this paper, we shall follow this latter group of scholars and their definitions of

- "discourse" as the process and rhetorical organisation of verbal communication and
- "text" as the (oral or written) product and form.

Both discourse and text can be analysed with regard to their internal relations and structures, but methodology and terminology vary. Before proceeding with the linguistic investigation proper, we shall briefly describe the units between which such relations are created (section 2.2), then define the concept of information structure (section 2.3), and finally present an overall summary of structures in texts (section 2.4).

2.2 Elementary Discourse Units (EDUs)

A text typically consists of more than one clause or sentence. Text segmentation has been treated in various ways in the literature, ranging from a very fine-grained division to a more general one that considers clauses as the minimal discourse unit. Such minimal units have been termed "Elementary Discourse Units", EDUs, by the Rhetorical Structure Theory (henceforth RST), a term that we shall adopt in the following.

In the "classical" RST (Mann and Thompson 1987; Mann, Matthiessen and Thompson 1992; Matthiessen and Thompson 1988 and later work), EDUs are considered to be clauses with the exception of clausal subjects and objects, other clausal complements and restrictive relative clauses. In the "modern" RST (Carlson et al. 2003), EDUs are clauses – including relative clauses – as well as attribution clauses and various phrases with strong discourse cues, such as *because of, in spite of, according to*. In this paper, we shall follow these principles with the exception of not segmenting attribution clauses as distinct EDUs. Normally, as a minimal requirement, EDUs must have a verbal element, but verbless constructions ending with full stops, question or exclamation marks as well as adjectival appositions are also identified as EDUs. An example of EDU identification is reproduced in (1); the sentence comes from the English L1 part of the Europarl Corpus[3] and contains four EDUs. Satellites, i.e. rhetorically subordinated EDUs (see section 2.4), are shown in italics, and the nucleus in bold fonts. Each EDU is delimited by square brackets.

(1) [*Looking at the package of amendments to our Rules of Procedure*]$_1$ [*tabled by the Committee on Constitutional Affairs,*]$_2$ [**I can say on behalf of my group that we will support the thrust**]$_3$ [*of what the Committee on Constitutional Affairs has put forward.*]$_4$ (ep-01-11-12.txt:56)

EDU_3 (the matrix clause), shown in bold fonts, is the nucleus of the sentence, to which EDU_1 (a present participle phrase) is linked as a satellite expressing a circumstantial relation. EDU_2 (a past participle phrase) is linked to EDU_1 expressing an elaboration of *the package* in question. Finally EDU_4 (a relative clause) is linked to the nucleus EDU_4 elaborating on *the thrust*.

---

[3] See section 3.1 for a description of this corpus. The numbers following each Europarl ("ep-") text are read: YY-MM-DD; "txt" indicates SPEAKER ID.

Other scholars use different terms for EDUs, such as "propositions" (e.g. Lehmann 1988), "abstract objects/arguments" (e.g. Prasad et al. 2008), and "discourse segments" (e.g. Irmer 2011, 128), all terms with more or less the same meaning as EDUs.

2.3 Information structure

Information structure is perhaps the linguistic structure that has been defined most uniformly by scholars, starting with Halliday's (1967) "given-new" categorisation based on the Prague School's "communicative dynamism" of the single units of a sentence (see e.g. Vachek 1966). Also subsequent and somewhat similar dichotomies such as "topic-comment" (Hockett 1958), "theme-rheme" (Firbas 1974), and "focus-presupposition" (Krifka 1993) are generally categorised as components of the information structure of a text. Lambrecht (1994, 1) notes, however:

> There has been and still is disagreement and confusion in linguistic theory about the nature of the component of language referred to in this book as INFORMATION STRUCTURE and about the status of this component in the overall system of grammar.

Regarding information density, in this paper we shall focus on the amount of information per text span (words and sentences), thus concentrating on the first of the four definitional elements of the term suggested by Fabricius-Hansen (1996, 529):

> [W]e would probably say that the informational density is higher in A than in B if at least one of the following conditions holds, other things being equal: i. the average amount of discourse information per sentence is higher in A than in B; ...

More precisely, we shall say that the information density is higher in A than in B if both contain the same amount of information but A is shorter than B, and/or A contains more EDUs than B, other things being equal. Among such "other things", a very important issue in a comparison between Italian and Danish is sentence length, as we shall see in section 3.2.

2.4 Structures in texts: An overview

The three kinds of structures in texts dealt with above all involve EDUs and the relations between them, but in different ways, as summed up in Table 1.

Discourse Structure	Text Structure	Information Structure
Content and process of communication:	Form and product of communication:	Information packaging in communication:
- discourse relations (also called coherence or rhetorical relations) between EDUs	- syntactic relations between EDUs	- information density (amount of information per text span, e.g. EDUs per sentence)
- rhetorical co-/subordination of EDUs	- morpho-syntactic co-/subordination of EDUs	- rhetorical and morpho-syntactic co-/subordination of EDUs

Table 1: Three structures in texts

Although rhetorical and morpho-syntactic co-/subordination of EDUs are mentioned under discourse and text structure respectively and may therefore seem redundant

with regard to information structure, they are, however, important aspects also of the "information packaging" of a text, as we shall see in section 3.4.

The taxonomy of discourse relations that we follow in our work (but due to space limitations cannot pursue further in this paper) is taken from RST and consists of coordinated (multinuclear) and subordinated (nucleus-satellite) relations. We have chosen this taxonomy for the reasons also stated by Bateman and Rondhuis (1997, 6), which are particularly relevant in a cross-linguistic context:

> An important claim of this theory is that the same rhetorical relations that hold between spans realised by individual clauses also account for the relationships between larger segments of text … It is this that makes it possible to characterise the structure of a text in terms of a single hierarchy of rhetorical relations inter-connecting the parts drawing on only a relatively small number of relation-types.

Innumerable other taxonomies have been proposed, see e.g. Danlos (2008), Irmer (2011), and Webber and Prasad (2009) for discussions.

3 Text and information structure in the Europarl corpus

At this point we shall confine our investigation to two of the three types of structures in text outlined in Table 1, i.e. the text and information structures, and we shall commence with a brief description of our empirical basis, the Europarl Corpus.

3.1 Corpus details and discussion

The Europarl Corpus is an open source corpus compiled by Koehn (2005) and recently updated[4]. It is a very large multilingual text collection with up to 50 million words per language and with source and target texts covering all 23 official languages of the European Union. The corpus was designed to train and evaluate statistical machine translation, and it is still extensively used for this purpose (see the corpus website for an overview). But as we shall see, it can also serve as empirical basis for other cross-linguistic investigations.

The texts are mainly argumentative – see van Halteren (2008) for a discussion – and consist of speeches made by the members of the European Parliament and other politicians in the years 1996–2011. Around 80 % of the speeches have been tagged with language attributes indicating the native language (L1) of the speaker. This made it possible for us, with the help of a Perl script, to automatically extract all Italian and Danish L1 text from the entire corpus. Since we wanted to perform both quantitative and qualitative analyses, we compiled two subcorpora: a subcorpus 1 with all the Italian and Danish Europarl texts from the period 1996-2003, and a subcorpus 2 with a limited number of quasi-randomly selected Italian and Danish texts from subcorpus 1, totalling some 15,000 words in each language. We used subcorpus 1 to calculate the average sentence length of all Italian and Danish L1 texts, see section 3.2, whereas subcorpus 2 served for more fine-grained and manually performed analyses, see section 3.3 and 4. In order to obtain a balanced and representative subcorpus 2, our requirements for these texts, which were selected manually, concerned variety regarding "chapters" (meeting sessions), dates (so that not all texts were speeches from the same period), speakers (a certain number of different speakers was chosen) and speech length (between 200 and 700 words).

[4] Europarl is available at http://statmt.org/europarl/. In this paper, we use the "v3" (third release) of the Europarl Corpus.

We chose Europarl as our empirical basis because it contains both parallel L1–L2 texts and comparable texts, i.e. L1 texts created in different languages but dealing with similar topics and produced in similar situations and genres for similar targets. Whereas parallel texts are clearly best suited e.g. for improving machine translation, since they permit L1–L2 text alignment and evaluation, comparable texts are generally best suited as the empirical basis for descriptive, possibly typological comparisons. In such cases, parallel texts are inappropriate because the filter of the translator and the translation strategies get in the way, and/or L2 texts may end up with a text structure very similar to that of the L1, as we shall see below. Baroni and Bernardini (2006, 260) refer to this phenomenon as "translationese":

> It is common, when reading translations, to feel that they are written in their own peculiar style. Translation scholars even speak of the language of translation as a separate 'dialect' within a language, which they call *third code* … or *translationese* ... Translationese has been originally described ... as the set of "fingerprints" that one language leaves on another when a text is translated between the two.

In the same vein McEnery et al. (2006, 49) state that

> source and translated texts … alone serve as a poor basis for cross-linguistic contrasts, because translations (i.e. L2 texts) cannot avoid the effect of translationese ... [C]omparable corpora are a useful resource for contrastive studies and translation studies, when used in combination with parallel corpora.

3.2 Sentence length

Differences in text and information structure show themselves in many ways, one of which is the simple sentence length, measured as words per sentence[5]. Of course, many reservations should be made when conducting linguistic measurements in this way, but we find the statistical results cited below convincing enough to be taken into account and used as a first indication of profound typological differences between the two languages analysed[6].

Subcorpus 1	Words	Sentences	Average words/sentence
Italian L1	1,657,592	47,405	34.97
Danish L1	546,425	22,668	24.11
Italian L2	571,115	22,154	25.78
Danish L2	1,845,951	57,574	32.06

Table 2: Sentence length in L1 and L2 Europarl texts.

Table 2 shows the average sentence length of all Italian and Danish L1 texts (subcorpus 1) and of the Italian L2 texts (translated from Danish) and Danish L2 texts (translated from Italian)[7]. Due to the differences of representation in the European

[5] A sentence is defined as a text segment followed by a full stop, a question mark, or an exclamation mark. Colons and semicolons are defined as punctuation marks separating clauses and not sentences.

[6] Also other scholars, such as Fabricius-Hansen (1998) and Teich (2003), use sentence length to measure text complexity in CL studies.

[7] We cannot entirely exclude that some translations from Danish to Italian, or vice versa, may have been translated from another L2 text, e.g. from one of the so-called EU "relay languages" (English, French or German). The Europarl Corpus does not provide any information in this regard, but even

Parliament between the two countries, there are roughly three times as many Italian L1 texts as Danish L1 texts. This, however, has no impact on the average analyses.

As the upper part of Table 2 shows, there is a considerable difference in average sentence length between the Italian L1 and Danish L1 texts, a difference amounting to 10.86 words per sentence, which means that the Italian sentences are almost 50 % longer than the Danish ones, 45.0 % to be exact. However, as the lower part of Table 2 shows, regarding sentence length, the Danish and Italian translators in the European Parliament tend to follow a rather imitative translation strategy. The Danish L2 texts are 33.0 % longer than the Danish L1 texts, while the Italian L2 texts are 35.6 % shorter than the Italian L1 texts. So regarding sentence length these L2 texts are clearly influenced by the L1 text structure, just as predicted by the scholars cited in section 3.1.

The longest Danish (L1) sentence of subcorpus 1 consisted of 146 words, and interestingly enough it had been merged with the following sentence in the Italian L2 text resulting in a 226 word long sentence. Table 3 shows the lengths of the longest sentences in each group, and the overall longest Danish L2 sentence consisting of 282 words is another excellent example of "translationese".

Subcorpus 1	**Italian L1**	**Danish L1**	**Italian L2**	**Danish L2**
Longest sentence (words)	266 ep-97-01-15.txt:97	146 ep-02-12-04.txt:17	226 ep-02-12-04.txt:17	282 ep-97-01-15.txt:97

Table 3: Longest sentence lengths in Europarl texts.

3.3 Information density

At this point we shall return to the concept of "information density", in Table 1 defined as the amount of information per text span, e.g. EDUs per sentence. If we look at some of the numbers of Table 2 that do **not** reflect an imitative translation strategy, we see that the Danish L2 texts have 11.4 % more words and 21.5 % more sentences than the Italian L1 texts. Assuming that the source and target texts contain the same amount of information (one of the most important criteria for EU translations), a very clear "dilution" of information density both on word and sentence level has occurred in the translation process from Italian to Danish. In sections 4.1-4.2, we supply a number of text examples that illustrate this dilution.

Also if we measure information density as the number of EDUs per sentence, there is a clear difference between the Italian and Danish Europarl texts. The count of EDUs textualised in each sentence can be a very time-consuming task, since no parser has been trained to do this convincingly. Therefore, we limited this analysis to subcorpus 2, and the results appear in Table 4:

Subcorpus 2	**Words**	**Sentences**	**EDUs**	**Average EDUs/ sentence**	**Percentage of sentences with five or more EDUs/sentence**
Italian L1	14,708	440	1,473	3.35	21.1 %
Danish L1	14,737	678	1,455	2.15	5.5 %

Table 4: Statistics on Europarl subcorpus 2.

Whereas the number of words and EDUs in the two text groups is roughly the same, the number of sentences varies considerably entailing great differences also in the average number of EDUs per sentence. Regarding sentences containing five

---

if this were the case in some instances, the indicated differences between Italian and Danish would still be valid.

EDUs or more, the Italian percentage is about four times the Danish percentage, as shown in the last column to the right.

We then performed a finer-grained analysis of the Italian and Danish EDU textualisation and found other substantial differences in the distribution of subordinate clauses, cf. Table 5:

Subcorpus 2	a. Subordinate clauses total	b. Relative clauses	c. Other subordinate finite clauses	d. Subordinate non-finite clauses	e. Other subordinate constructions[8]
Italian L1	82.8 %	35.6 %	23.4 %	30.8 %	10.3 %
Danish L1	78.6 %	39.5 %	39.7 %	17.0 %	3.8 %

Table 5: Distribution of EDUs in subordinate clauses in Europarl subcorpus 2.

Subordinate clauses amounted to 82.8 % of all clauses in the Italian texts as opposed to 78.6 % of the clauses in the Danish texts[9]. However, the most significant and interesting differences lie in the distribution of finite vs. non-finite subordinate clauses. The Danish texts contain 79.2 % finite subordinate clauses (columns b+c), but only 20.8 % non-finite clauses and verbless constructions (columns d+e). In the Italian texts, on the other hand, the distribution between finite and non-finite subordinate clauses is more equal: 59.0 % finite clauses (b+c) and 41.1 % non-finite and verbless constructions (d+e). In other words, subordinate non-finite and verbless constructions are twice as frequent in Italian as in Danish. Furthermore (what is not shown separately in Table 5 but included in column d), Italian speakers use the whole range of non-finite verb forms (gerunds, participles, infinitives and nominalisations) much more regularly, whereas Danish speakers generally confine themselves to infinitives (the gerund does not exist in Danish). The majority of the subordinate constructions in (e) are complex adjectival postmodifiers, which in many cases correspond to the Danish relative clauses (column b). We shall investigate this matter in more detail in the following sections, in particular 4.1-4.2.

3.4 EDU linkage

As just stated, morpho-syntactic linkage of EDUs differs greatly between languages and, not least, between language families and groups such as the Romance and Scandinavian. Regarding clause linkage, Lehmann (1988) can still be considered as one of the "classic" and most important papers, whereas syntactic co-/subordination has been investigated and described by many other scholars, e.g. Fabricius-Hansen and Ramm (2008, 2-3) who define these concepts in the following way:

> In what is probably their most widespread application, 'subordination' and 'coordination' – along with their adjectival cognates 'subordinate', 'coordinate', etc. – are syntactic notions denoting relations between parts of a complex syntactic unit. That is, they concern the structure of sentences or clauses and their parts. … As far as the domain of natural discourse and texts is concerned, it is a common observation in various theoretical approaches that entities of this domain too can be organised hierarchically ('subordinating', 'hypotactically') or non-hierarchically ('coordinating', 'paratactically').

[8] These constructions include complex nominal and adjectival postmodifiers (attributives and appositions).

[9] If we included clausal subjects, objects and other complements, the differences between main and subordinate clauses in the two text groups would become much more considerable.

Among the many other cross-linguistic surveys on text structure are e.g. Fabricius-Hansen (1996; 1998) and Ramm and Fabricius-Hansen (2005), who investigate English, German and Norwegian, i.e. three Germanic languages, and Skytte and Korzen (2000), who investigate Italian and Danish. On grammatical shifts e.g. between finite verbs and nominalisations in translation processes between English and German, see Alves et al. (2010), and on information density and explicitness in English-German translations, see also Hansen-Schirra, Neumann and Steiner (2007). As stated in Table 1, and like scholars such as Asher and Vieu (2005, 594), we consider EDU co- vs. subordination (both rhetorical and morpho-syntactic) as part of the "information packaging" of a text, a term suggested by Chafe (1976, 28) and later used, especially in connection with given vs. new entities and definiteness, e.g. by Clark and Haviland (1977), and Vallduví and Engdahl (1996).

In a sequence of EDUs, such as the following:

(2) EDU1: arrive (John, in London); EDU2: go (John, home)

EDU1 can – if interpreted as the rhetorical satellite – be textualised in different ways, as shown in the "Deverbalisation Scale" in Table 6.

	EDU1 textualised as	**Textualisation EDU1 + EDU2**
↓	a. an independent sentence	*John arrived in London.* He went straight home.
	b. a main clause, part of sentence	*John arrived in London* and he went straight home.
	c. a subordinate finite clause	*After John arrived in London,* he went straight home.
	d. a subordinate non-finite clause	*Having arrived in London,* John went straight home.
	e. a nominalisation	*After his arrival in London,* John went straight home.

Table 6: The Deverbalisation Scale

The scale is based on Hopper and Thompson (1984), Lehmann (1988), and Korzen (2007a; 2009), and the deverbalisation of the EDU1 increases from (a/b) to (e) together with its integration into the matrix clause. Whereas the finite verb in a main clause, levels (a/b), has its full (language specific) range of grammatico-semantic values[10] and the clause its full range of pragmatic-illocutionary possibilities, these values are gradually reduced or lost in the textualisations further down the scale. The verb in the subordinate finite clause, level (c), loses its independent tense, mood and illocution; these features will be determined and/or expressed by the matrix clause[11], in the case of for instance tense due to the so-called *consecutio temporum principle*, as in the Italian/English example (3b):

(3) a. So che Leo è arrivato alle 9. 'I know that Leo arrived/has arrived at 9.'

*b. Sapevo che Leo era (*è) arrivato alle 9*. 'I knew that Leo had (*has) arrived at 9.'[12]

The non-finite verb at the (d) level loses all temporal, modal, and aspectual values, and with the exception of the constructions mentioned in footnote 13 below, it cannot render explicit its subject:

---

[10] Hopper and Thompson (1984, 708) here talk about the "prototypical verb function".

[11] Exceptions are appositive relative clauses, which may have an illocution value different from that of the matrix clause: *I brought you these books, **which you will read for the next lesson**!*, assertive (matrix) vs. directive (relative clause) illocutionary acts.

[12] In the Romance languages, the (c) level is divided into two: subordinate clauses in the indicative and in the subjunctive. The latter verbs have lost their aspect distinctions, their ability to assert an event or situation and some tense possibilities. Thus, this level can be considered as more deverbalised than the indicative level. See Korzen (2007a; 2009).

(4) ***John** *having arrived late, John/he went straight home.*

***John** *born into a family of musicians, John/he began studying piano at the age of ten.*

The lack of subject marking of the non-finite constructions generally entails an inherent subject/topic continuity (a topic shift normally requires a finite verb with an explicit subject), which means that the situation or event in question is evaluated and interpreted as related to the on-going topic but less important than the situation or event of the matrix clause, textualised with a finite verb.

The last of the constructions, the nominalisation, (e), is completely integrated in the matrix clause as a second order entity in Lyons' (1977, 442ff) terminology. It has lost all its verbal-morphological characteristics and its valency complements are syntactically reduced to secondary positions or simply left out, as in (5a). An NP such as (5b) will often appear as relatively "heavy" and tend to be avoided.

(5) a. *(The manager evaluated the performance →) The evaluation of the performance… / The manager's evaluation…*

b. *The manager's evaluation of the performance…*

In other words: The further down on the Deverbalisation Scale an EDU is textualised, the fewer grammatico-semantic features are expressed by the verb, i.e. the more "deverbalised" it is, and the more pragmatically and rhetorically subordinated and incorporated in the matrix clause is the EDU. In the case of non-finite and nominalised verbs, levels (d/e), features such as subject, tense, mood, aspect and illocution are entirely interpreted on the basis of the matrix clause[13]. Therefore, the pragmatic and semantic interpretation of non-finite or nominalised structures is entirely dependent on the matrix clause, and such structures express a particularly strong rhetorical backgrounding (explicit satellite status) of the EDU in question, as stated also by Lehmann (1988, 214):

> [A]dvanced hierarchical downgrading of the subordinate clause implies a low syntactic level for it. We will thus be justified if in the following we take advanced downgrading as a sufficient condition for high integration. High integration of the subordinate into the main clause correlates positively with its desententialisation.

See authentic examples of (b)-(e) structures in section 3.5.

3.5 Text (syntactic) structure and discourse (rhetorical) structure

As stated in the previous section, non-finite and nominalised structures explicitly express the satellite status of the EDU in question. Generally – but not necessarily – this is true also of subordinate adverbial clauses, such as the EDU1 clause of the example in Table 6(c), *After John arrived late in London, (he went straight home)*. Exceptions to this rule can be found especially in subordinate temporal clauses, e.g. :

(6) *I was walking on the beach when suddenly I heard a big explosion,*

---

[13] Regarding subject, we here ignore e.g. the so-called "absolute constructions" consisting of a participle or gerund + a subject different from the subject of the main verb, e.g.
***Morto il padre**, Luca partì per Roma '**The father [having] died**, Luca left for Rome'.*
As we saw in (5), in nominalised verb forms the subject may appear as a secondary valency complement:
***The manager**'s evaluation.*

where the first (and matrix) clause is the rhetorical satellite indicating the background scene of the nucleus expressed by the syntactically subordinate circumstantial/temporal clause. Other exceptions are subject, object and subject complement clauses, which are valency constituents of the matrix clause and therefore not textually backgrounded, and appositive relative clauses, which may carry on the story line (the "continuative appositive clauses", cf. Loock 2010, 95) or for other reasons express the most important part of the text sequence, see an example in footnote 11 (and more details in Korzen 2007b and 2009).

On the other hand, the structures at levels (a/b) of the Deverbalisation Scale are in themselves ambiguous as to mono- or multinuclear interpretation. Asher and Lascarides (2003, 165-168) treat fore- and background interpretation of independent sentences in their Segmented Discourse Representation Theory (SDRT), quoting e.g.:

(7) *A burglar broke into Mary's apartment. a) Mary was asleep. b) A police woman visited her the next day.*

The a) continuation is a background sequence, which permits a following pronominal anaphorisation of the NP *a burglar*: ***He*** *stole the silver*. The b) continuation is a foreground sequence, which does not license the same pronominal anaphor. Similar analyses (although not in a SDRT context) are found in Korzen (2000, 486-492) and (2001, 114), where a distinction is made between primary and secondary text topics according to the status of the text segment in which they are located.

However, it is well known that the syndetic coordination with the connective *and* (and its cross-linguistic counterparts), as in Table 6(b), often contains an EDU1 with satellite status. The literature on the function and semantics of *and* is vast and mostly theory and/or language dependent. Important cross-linguistic studies on *and* and counterparts are found e.g. in Ramm and Fabricius-Hansen (2005), Behrens and Fabricius-Hansen (2010) and Skytte (2000, 652-660). Following Txurruka (2000), Asher and Vieu (2005, 598-599) define *and* as an unequivocal coordination marker, a viewpoint which is contrary to our analyses and those of the other scholars just mentioned. In a case like the one cited in Table 6(b), *John arrived late in London and he went straight home*, the EDU1 can very well be seen as a satellite expressing the cause of the EDU2.

To further support this viewpoint, we shall cite a few Italian and Danish examples (similar to the sentences in Table 6 but authentic, and all L1) from a corpus of comparable narrative texts, the so-called "Mr. Bean corpus", consisting of a number of retellings of two Mr. Bean episodes produced by 27 Italian and 18 Danish university students, see Skytte et al. (1999) and Korzen (2007b). The examples below reproduce a scene from the episode "The Library", in which Mr. Bean, sitting in the reading room of a library, has placed a sheet of tracing paper on a manuscript illustration that he wants to copy by hand. But then he happens to sneeze, which causes the tracing paper to fly away with the result of him drawing directly on the manuscript, thereby ruining it. Thus, the sneeze is the cause of the main event from which the following action evolves: the tracing paper that flies away, and the EDU1 indicating the sneeze is the causal satellite. The following examples show textualisations of the EDU1 (indicated with bold italics) at the following levels of the Deverbalisation Scale, (8)-(10): b; (11): c; (12): d; (13): e:

(8) [Danish] ***Mr. Bean kommer til at nyse****, og kalkerpapiret flyver væk uden han opdager det…* (Skytte et al. 1999: DSA9)[14]
[lit.] *'**Mr. Bean happens to sneeze** and the tracing paper flies away without [that] he discovers it'*

[14] Reference indications of the Mr. Bean corpus: D = Danish, I = Italian; S = written, M = oral.

(9) [Danish] ***Pludselig nyser han**, og papiret ryger væk uden at han ser det* (DSA3)
[lit.] *'**Suddenly he sneezes**, and the paper flies away without that he sees it'*

(10) [Italian] *…dopo aver appoggiato una velina su una pagina del libro, **starnutisce fragorosamente** e sporca il libro.* (ISA13)
[lit.] *'after having placed a tracing paper on a page of the book, **he sneezes loudly** and [he] dirties the paper'*[15]

As stated, all three cases are textualisations at level (b) of the Deverbalisation Scale, but in the Italian texts, this structure is rare. Much more often the satellite status is grammaticalised more unambiguously in other ways, i.e. as a subordinate finite (adverbial) clause:

(11) ***poiché starnutisce** il foglio vola via e lui si ritrova a colorare sul libro datogli. (ISA3)*
[lit.] *'**since he sneezes** the paper flies away and he finds himself colouring the book given to him'*

or – especially – as a non-finite clause, (12), or a nominalisation, (13), thus confirming the numbers quoted in column (d) of Table 5 above:

(12) *Poi si mette a ricalcare [...] solo che **starnutendo** il foglio gli vola via (ISA1)*
[lit.] *'Then he starts to copy […] but then **sneezing** the paper [for him] flies away'*

(13) *cerca appunto di-, di di copiare il disegno, solo che eh-, **con uno starnuto** il foglio trasparente gli, gli vola via (IMB8)*[16]
[lit.] *'he tries precisely to copy the illustration, however **with a sneeze** the tracing paper flies away from him'*

However, examples (8)-(10) are cases of syndetic coordination with *and* in which the EDU1 plays the role of satellite, and which therefore contradict Txurruka's (2000) and Asher and Vieu's (2005) conception of *and* as an unequivocal coordination marker. Examples like these (and others) have prompted Korzen (2000, 87) to conclude that at least in Italian and Danish all four of the following combinations are possible:

- rhetorically and syntactically superordinate EDUs
- rhetorically subordinate and syntactically superordinate EDUs
- rhetorically superordinate and syntactically subordinate EDUs
- rhetorically and syntactically subordinate EDUs

3.6 Text and information structure in Italian and Danish: An overview

The cross-linguistic Italian-Danish characteristics outlined in the previous sections are not limited to particular text types or genres. Korzen (2009) quotes a number of surveys that document the exact same situation in six other text types, including the narrative "Mr. Bean corpus" cited in section 3.5. The results all confirm a higher information density and degree of deverbalisation in Italian texts than in comparable Danish texts. Italian sentences tend to be longer and to include more EDUs, of which a higher number is textualised at the lower levels of the Deverbalisation Scale, i.e. backgrounded by means of non-finite and nominalised predicates. This typically leads to EDUs containing fewer words (a finite structure will normally require at least an explicit connective and a subject) and to a multi-layered and hierarchical

[15] Italian is a pro-drop language, and the verb form *sporca* contains the indication of the 3rd person singular.

[16] In the transcriptions of the oral texts, as this one, a comma indicates a short interval and a hyphen the extension of a vowel.

information structure, characterised by a high degree of topic continuity, in which the various events are evaluated with respect to their importance to the on-going topic.

On the other hand, Danish text structure tends to be more informationally linear and characterised by a higher degree of finite verbs and topic shifts. Each sentence holds fewer but longer EDUs, and different events tend to be textualised at the same and higher levels of the Deverbalisation Scale, i.e. more chronologically one after the other and with finite verb forms that permit subject/topic changes.

4 Perspectives for Translation

Concerning the parallel (L2) Europarl texts cited in section 3.2, Table 2, the picture was different. Regarding sentence length, we observed a general tendency towards an imitative translation strategy, i.e. a strategy whereby the target text followed the structure of the source text relatively closely.

In the following sections, we shall advocate a different translation strategy, viz. the functional strategy. This method focuses on the function of the target text with respect to the new addressees, which should be equal to the function of the source text with respect to the original addressees. The functional strategy generally requires, among other things, a particular awareness of the text structure of both source and target language; if the structure of a source text can be considered as "typical" with respect to the source language (and to the particular text type) in question, it should be "typical" also when transformed to the target language. Dealing with translations between a Romance and a Scandinavian language, two of the major text structural differences concern precisely the issues investigated above:

- Information density and sentence length, i.e. the amount for information per text span and of EDUs per sentence;
- EDU linkage, i.e. the textualisation of EDUs, particularly regarding non-finite and nominalised structures.

Very generally speaking, when translating from a Romance to a Scandinavian language, particularly long sentences should be divided into shorter ones, thereby reducing the number of EDUs per sentence, and non-finite and nominalised EDUs should be changed into finite structures, thereby rendering the text structure more linear and increasing the number of words per EDU.

In the following sections, we shall give some specific examples of how this can be done, citing a number of Europarl cases of what we would define as "felicitous translations" from Italian into Danish, "felicitous" in the sense that they respect the cross-linguistic structural differences and thereby contribute to idiomatic and "non-marked" L2 texts. Thus, they are counterexamples to the Europarl L2 tendencies cited in Table 2.

Since we are dealing with text structure, it is often necessary to quote quite lengthy passages, and given that probably not all our readers are familiar with Italian and Danish, we shall have to add an English translation of both source and target passage. So, due to the space limitations of this paper, we shall confine ourselves to relatively few examples and limit the Italian source and Danish target passages to the particular linguistic issue at play (written with bold italics) with a literal English translation of both. To those we shall add a longer co-text of the official English translation of the passage in question in order to clarify the textual content. It is

interesting to see that the official English translations in some cases follow the Italian text structure, in others the Danish structure[17].

4.1 Information density and sentence length

One way of reducing the number of EDUs per sentence is simply to divide long Italian sentences into shorter Danish ones. For instance, syndetic coordinate structures (level b on the Deverbalisation Scale) can be changed into independent sentences (level a) simply by omitting a coordinate connective and changing a comma into a full stop. Translating from Danish into Italian, the reverse manoeuvre can be applied.

(14) [Ital. L1] *…nemmeno in altre lingue europee,* ***ed*** *è sintomatico…* (ep-01-09-04.txt 150)

'…neither in other European languages, and it is symptomatic…'

[Dan. L2] *…heller ikke på de andre europæiske sprog****. Man kan sige*** *at det er symptomatisk…*

'...neither in other European languages. One can say that it is symptomatic…'

[Eng. L2] *…in our common language, the word "governance" does not exist, and it may well be that it does not exist in other European languages either****. It could be said*** *that it is revealing ... that ... the Commission chose ... to adopt a document with an untranslatable title.*

Similarly, a subordinate clause (level c) can be changed into an independent sentence by omitting the connective and adding a full stop. The following is an example of an adversative clause, and in such cases the connective (Ital.: *mentre,* Eng.: *while*) can be changed into an adversative adverbial; in the equivalent independent Danish sentence, the adverb *ellers* 'however' has been used:

(15) [Ital. L1] *…futuri passi* ***mentre, per quanto riguarda il servizio universale, l'esempio svedese dovrebbe rassicurare tutti****…* (ep-00-12-13.txt 20)

'…future steps while, regarding the universal service, the Swedish example should reassure all…'

[Dan. L2] *…de kommende skridt.* ***Når det gælder den universelle tjeneste, burde Sveriges eksempel ellers berolige alle*** *…*

'…the future steps. When it comes to the universal service, the Swedish example should however reassure all…'

[Eng. L2] *In fact, it provides … absolutely no certainty regarding future steps* ***while, as far as the universal service is concerned, the Swedish example should reassure all those who feel that privatisation will mean the end of the postal services.***

Translating from Danish to Italian, the translator should look for a text structuring adverb such as the adversative *ellers* and change it into an equivalent subordinating connective, such as *mentre*.

Even non-finite Romance clauses, level (d) on the Deverbalisation Scale, can be transformed into independent finite sentences in Danish, although this happens more rarely. Ex. (16) is a case of an Italian participle phrase:

[17] Lexically, English can be considered a typological "hybrid language" between the Scandinavian and Romance languages, see e.g. Baron and Herslund (2005). Judging by the examples of the parallel Europarl Corpus, this seems to be true also regarding text structure.

(16) [Ital. L1] *…ci sia stato un piccolo braccio di ferro tra i gruppi,* ***risoltosi nel modo che constatiamo****.* (ep-00-12-13.txt 20)

'…there has been a small tug-of-war between the groups, resolved in the way we can see.'

[Dan. L2] *…der var lidt tovtrækkeri mellem grupperne.* ***Resultatet kender vi****.*

'…there was a small tug-of-war between the groups. The result we know.'

[Eng. L2] *I would like to start by expressing my satisfaction at the fact that this debate is being held today instead of during the January part-session,* ***although this is the result of a minor tussle between the groups.***

Also EDUs without a verbal element, e.g. syntactic appositions consisting of noun, adjective, or prepositional phrases, can be translated into perfectly idiomatic Danish independent sentences. In the following example, this is done by repeating (anaphorising) the noun to which the apposition is linked and using it as the subject of an independent sentence:

(17) [Ital. L1] *…che garantiscano un livello minimo di garanzie e di attenzione verso il mondo degli anziani,* ***uguale in tutti i paesi dell'Unione****…* (ep-02-04-11.txt 43) (see a continuation of this passage in ex. (18))

'…that ensure a minimum level of guarantees and focus on the world of the elderly, equal in all the member states…'

[Dan. L2] *…der sikrer et minimumsniveau af garantier og opmærksomhed over for de ældre.* ***Dette niveau skal være det samme i alle EU-landene,****…*

'…that ensure a minimum level of guarantees and focus on the elderly. This level must be the same in all member states…'

[Eng. L2] *…Europe needs to develop policies ensuring a minimum level of guarantees and focus on the world of the elderly,* ***a level which is the same in all the countries of the Union****….*

In case of a Danish-Italian translation, the translator should here note the repetition of the noun, which together with the modal and copula verb constitutes a segment that is really superfluous in Italian. In general, translating from Italian to Danish should imply "moving upwards" on the Deverbalisation Scale, and translating from Danish to Italian "moving downwards".

4.2 EDU linkage and deverbalisation

When translating from Italian into a Scandinavian language, the translator should be particularly aware of non-finite EDUs and in most cases, also here, seek to "move upwards" on the Deverbalisation Scale. In (18) an infinitive phrase after the preposition *da* 'to', level (d), is syndetically coordinated with the complex adjectival apposition *uguale in tutti i paesi…* cited in (17). Here, *da* + infinitive has the modal sense of *must/should be* + participle, and the Danish translator has followed the same strategy as in (17) by anaphorising the entity described, the *level of guarantees*, and using the explicit anaphor (here a pronoun) as subject of a finite form of the modal verb. Thus, the preposition + infinitive phrase has become a main clause at level (b):

(18) [Ital. L1] *…uguale in tutti i paesi dell'Unione e* ***da proporre come punto d'arrivo****…* (ep-02-04-11.txt 43)

'…equal in all countries of the Union and to propose as a goal…'

[Dan. L2] *…det samme i alle EU-landene,* ***og det skal ligeledes opstilles som****…*

'…the same in all the EU countries, and it shall also be proposed as…'

[Eng. L2] *Europe needs to develop policies ensuring a minimum level of guarantees and focus on the world of the elderly, a level which is the same in all the countries of the Union* ***and must be proposed as a goal for the candidate countries too****.*

Going from Danish to Italian, the translator should again notice the repetition of the (here pronominalised) entity and be aware that a modal structure *must/should* + a passive participle can be rendered with *da* + infinitive in Italian.

The gerund does not exist in Danish, so here translators into Danish are compelled to find other solutions. Very often a Romance finite verb + a gerund will correspond to a particular coordinate Danish (and English) construction, i.e. the serial verb construction, $verb_1$ *and* $verb_2$, where the subject of $verb_2$ is the same as that of $verb_1$ but implicit[18]. Normally $verb_2$ corresponds to the Italian gerund, and if a rhetorical relation can be inferred between the two verbs + complements, the Danish $verb_2$ may be adverbially specified accordingly as in the following example, where Danish *således* 'thus, in this way' expresses consequence or result:

(19) [Ital. L1] *…dei diritti umani in quel paese…,* ***costruendo su tale questione****…* (ep-00-06-14.txt 176)

'…of the human rights in that country, creating on that issue…'

[Dan L2] *…for menneskerettighederne i Tunesien …* ***og således skaber****…*

'…for the human rights in Tunisia and this way creates…'

[Eng. L2] *It is therefore to be hoped that the EU-Tunisia Association Council will assume the responsibility of continuously monitoring the human rights situation in Tunisia…,* ***and that it will create*** *a joint system to monitor the issue which can only bring social improvements to the human rights situation in Tunisia.*

Going from Danish to Italian, the translator should be aware of the special Italian pro-drop situation which means that a finite verb will always contain an indication of the subject's number and person (cf. ex. (10) and footnote 15 above). Therefore, the structure $verb_1$ *and* $verb_2$ is not in itself equivalent in the two languages, and in most cases, the special cohesion of the Danish $verb_1$ *and* $verb_2$ construction should be rendered differently in Italian, for instance by a finite + non-finite verb construction, as in (19), even though such a construction does not render explicit the rhetorical relation between the two verbs + complements – other than specifying the satellite status of the second and non-finite verb + complement.

Instead of by a Danish adverbial specification, like *således* in (19), the rhetorical relation between two verbs can be rendered explicit by a connective of a subordinate adverbial (finite) clause, in (20) *så* 'so (that)': consequence/result. Again, going from a Danish finite clause to an Italian gerund, the explicitness of such a rhetorical relation is lost if the gerund is not particularly specified in some way:

(20) [Ital. L1] *…sostengo la proposta di assumere … gli obiettivi di Lisbona e di Göteborg, …* ***sottraendo l'ammontare di questi investimenti****…* (ep-02-10-21.txt:49)

'…I support the proposal to include the Lisbon and Gothenburg objectives, ***…subtracting the sum of these investments****…'*

[Dan. L2] *…et tillægsmål for stabilitets- og vækstpagten,* ***så udgifterne til disse investeringer bliver trukket fra****…*

[18] On serial verb constructions – also called complex predicates – see e.g. Lehmann (1988, 189), Herslund (2000), Choi (2003) and Aikhenvald (2005).

'…an additional objective of the Stability and Growth Pact, ***so that the sums of these investments are subtracted****…'*

[Eng. L2] *I therefore support the proposal to include the Lisbon and Gothenburg objectives … as an additional objective of the Stability and Growth Pact,* ***subtracting the sum of these investments from the total budgetary deficit of Member States' governments****.*

In many cases, an Italian gerund phrase merely expresses a concomitant (but less important) situation or event, and a frequent Danish translation will be a subordinate clause with the (semantically weak) temporal connective *idet* 'as'.

(21) [Ital. L1]*…consentire all'Unione europea di diventare… l'area più competitiva e più dinamica di una società basata sulla conoscenza, sulla piena occupazione e sullo sviluppo sostenibile,* ***favorendo altresì il loro coordinamento****.* (ep-02-10-21.txt:49)

'…allow the European Union to become the most competitive and dynamic economy based on knowledge, full employment and sustainable development… ***facilitating also the coordination of these investments****.'*

[Dan. L2] *…at gøre EU til verdens mest konkurrencedygtige og dynamiske videnbaserede økonomi, som bygger på fuld beskæftigelse og bæredygtig udvikling,* ***idet man ligeledes fremmer samordningen af disse investeringer****.*

'…to make the European Union the most competitive and dynamic knowledge based economy, which builds on full employment and sustainable development, ***as we also facilitate the coordination of these investments****.'*

[Engl. L2] *…with the objective of making the European Union the most competitive and dynamic economy based on knowledge, full employment and sustainable development in the world…,* ***facilitating the coordination of these investments****.*

Also EDUs textualised as attributive or appositive participle phrases are extremely frequent in Italian, and although participles do exist in Danish, they occur much more seldom. Here, a good strategy is to change the non-finite structure into a finite relative clause, as in the following example, which contains no less than four EDU participle phrases:

(22) [Ital. L1]*…quando questa è ripetitiva su ingredienti e principi attivi* ***conosciuti da anni [known for years]*** *e* ***già immessi in commercio [already put on the market]****, allora il sacrificio di nuovi animali è assolutamente inutile. Ma quando…nell'emendamento* ***sottoscritto da oltre cinquanta parlamentari [supported by more than 50 MEPs]****, si tratta di nuovi cosmetici contenenti ingredienti nuovi, mai* ***testati sperimentalmente prima [never tested before]****…*(ep-01-04-02-txt:42)

[Dan. L2] *Når der er tale om gentagelser af forsøg med ingredienser og aktive bestanddele,* ***der har været kendt i årevis [which have been known for years]****, og* ***som allerede er i handlen [which are already on the market]****, er det absolut unødvendigt at ofre nye dyr. Men når der…i det ændringsforslag,* ***som over 50 parlamentsmedlemmer har underskrevet [that more than 50 MEPs have supported]****, hr. kommissær - er tale om nye kosmetiske midler, der indeholder nye ingredienser,* ***hvis giftighed aldrig er blevet testet på forsøgsdyr før [whose level of poisonousness never have been tested on test animals before]****,…*

[Eng. L2] *I also share the public's concerns regarding animal experimentation; when this is a matter of repeat trials on active ingredients and principles* ***which***

***have been known for years and are already on the market**, then the sacrifice of more animals serves absolutely no purpose. However, Commissioner, when, as I point out in an amendment **supported by over 50 Members of Parliament**, it is a matter of new cosmetics containing new ingredients **which have never been tested in the past** in order to establish their toxicological profile in laboratory animals, then, as a scientist, I am convinced that it is essential to carry out an initial set of experiments on animals…*

In these cases, going from Danish to Italian, the translator should be aware of relative clauses with a transitive verb in the passive voice (or which can be paraphrased with a passive verb), as in (22). In such cases the clause can be changed into a participle phrase[19].

Similarly, other lengthy Italian adjectival attributives and appositions are normally transformed into Danish finite structures, in (17) above an independent sentence. In the following (and very typical) case, we have again a relative clause:

(23) [Ital. L1] *…un'Unione libera e indipendente,* ***portatrice di un progetto di pace****…* (ep-03-04-09.txt:235)

'…a free, independent Union, ***bearer of a project of peace****…'*

[Dan. L2] *…et frit og uafhængigt EU,* ***som er drivkraften bag et projekt****…*

'… a free, independent Union ***which is the driving force behind a project****…*

[Eng. L2] *…the pride felt by each citizen at belonging to both their country and the Union: a free, independent Union* ***which is the author of a project of peace and mutual respect with regard to the rest of the world.***

In such cases, the head of the Italian apposition becomes the complement of a (finite) Danish copula verb. Going from Danish to Italian, the translator should thus be aware of relative clauses of which the relative pronoun functions as subject of a copula verb followed by a complement.

4.3 An overview

We hope that the few text passages in sections 4.1-4.2 will suffice as exemplifications of felicitous "transformations" of information density and EDU linkage from Italian (or another Romance language) into Danish (or another Scandinavian language) and vice versa, and we believe that, in principle, the translation rules cited in relation to each case could be implemented by human translators as well as by machine translation systems.

In a computational context, we believe that a "pre-processing" phase could constitute a compelling method for automatically adapting the text structure of the source language to the text structure of the target language before the actual translation. Scholars such as Collins et al. (2005) have demonstrated how adding knowledge about syntactic structures can significantly improve the performance of existing state-of-the-art statistical machine translation systems, and we see no reason why adding knowledge about text structure should not be able to do likewise.

In summary, the most important text structural amendments in a translational context Italian–Danish (or another Romance-Scandinavian language pair) should be the following:

[19] The same is true of relative clauses with intransitive verbs in the active voice, but due to space limitations, we shall omit such examples in this paper.

- Long sentences, e.g. containing more than four EDUs, are divided into shorter sentences with fewer EDUs; colons and semicolons between finite clauses are changed into full stops.
- Gerund phrases are changed into coordinate finite clauses with the connective *and*. In this way, the often somewhat difficult task of choosing the appropriate adverbial connective is avoided.
- Long appositive and attributive participle phrases are changed into finite relative clauses.

The first modification deals with information density measured as number of EDUs per sentence, the last two ones with information density as well as with clause linkage ensuring an "upward" movement on the Deverbalisation Scale and thus a more finite and paratactic L2 text structure.

Translating from Danish (or another Scandinavian language) into Italian (or another Romance language), a relatively linear text structure should become relatively more hierarchical by means of a "downward" movement on the Deverbalisation Scale. The EDUs moving down the scale should be the rhetorical satellites, for which reason a "perfect" functional translation strategy should include a rhetorical structure analysis, an analysis that probably, at the present state, could not be carried out entirely without human participation.

5 Conclusion

In this paper, we have highlighted some particularly thorny problems linked with text structure in a cross-linguistic and translational perspective. We have focused on a Romance and a Scandinavian language, Italian and Danish respectively, centring our attention on the two text structural issues that to our experience cause the greatest problems in a contrastive and/or translational context, viz. information density and clause linkage.

With the considerable theoretical and terminological confusion regarding text structure – as well as other structures in texts – we found it necessary first to clarify some definitional ambiguities and thereafter to give a fairly in-depth description of the two key issues in the two languages investigated. Serving as our empirical basis, the Europarl Corpus had a number of advantages as a combined comparable and parallel text corpus, permitting in-depth and thorough cross-linguistic L1 comparisons as well as L1–L2 comparisons and statistical counts.

The significant differences between Italian and Danish text structure that we could ascertain in our L1 comparisons, especially regarding sentence length, were not always to be found, at least not to the same extent, in our L1–L2 comparisons. We therefore believe that the two linguistic key disciplines at play here, Contrastive Linguistics and Translation Studies, are not only "neighbouring disciplines" that "would benefit from each other's insights" as Chesterman described them (see the very first paragraphs of this paper), but in fact deeply dependent on each other. Regarding information density and clause linkage, there are no easy solutions; TS, including Machine Translation, must learn about text (and discourse) structure from CL, and similarly CL should widen its horizons by including TS knowledge on translation strategies. Only that way we will be able to create for instance MT systems that can deal more efficiently than what is the case today with the two issues examined here and produce L2 texts that truly resemble L1 texts, also when it comes to the thorny information and text structure.

6 References

Aikhenvald, Alexandra Y. 2005. Serial verb constructions in typological perspective. In *Serial verb constructions: a cross-linguistic typology*, ed. Alexandra Y. Aikhenvald and Robert M.W. Dixon, 1-68. Oxford: Oxford University Press.

Alves, Fabio, Adriana Pagano, Stella Neumann, Erich Steiner, and Silvia Hansen-Schirra. 2010. Translation units and grammatical shifts. Towards an integration of product- and process-based translation research. In *Translation and Cognition*, ed. Gregory M. Shreve and Erik Angelone, 109–142. Amsterdam/Philadelphia: John Benjamins.

Asher, Nicholas, and Alex Lascarides. 2003. *Logics of conversation*. Studies in natural language processing. Cambridge University Press.

Asher, Nicholas, and Laure Vieu. 2005. Subordinating and coordinating discourse relations. In *Lingua*, 115, 591–610.

Bache, Carl, and Niels Davidsen-Nielsen. 1997. *Mastering English: An advanced grammar for non-native and native speakers.* Berlin/New York: Walter de Gruyter.

Baron, Irène, and Michael Herslund. 2005. Langues endocentriques et langues exocentriques. Approche typologique du danois, du français et de l'anglais. In *Le génie de la langue française. Perspectives typologiques et contrastives,* Langue française, 145, ed. Michael Herslund and Irène Baron, 35-53.

Baroni, Marco, and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. In *Literary and Linguistic Computing*, 21, 3, 259–274.

Bateman, John A., and Klaas Jan Rondhuis. 1997. 'Coherence relations': towards a general specification. In *Discourse Processes*, 24, 3-49.

Behrens, Bergljot, and Cathrine Fabricius-Hansen. 2010. The relation accompanying circumstance across languages: Conflict between linguistic expression and discourse subordination? In *Contrasting meaning in languages of the East and West.* Contemporary Studies in Descriptive Linguistics, 14, ed. Dingfang Shu and Ken Turner, 531–552. Bern: Peter Lang.

Brown, Gillian, and George Yule. 1983. *Discourse analysis*. Cambridge: Cambridge University Press.

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current and new directions in discourse and dialogue*, ed. Jan van Kuppevelt and Ronnie W. Smith, 85-112. Dordrecht: Kluwer Academic Publishers.

Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Subject and topic, ed. Charles N. Li, 25–55. New York/San Francisco/London: Academic Press.

Chafe, Wallace L. 2003. Discourse: Overview. In *International Encyclopedia of Linguistics*, ed. William Frawley. New York: Oxford University Press.

Chesterman, Andrew. 1998. *Contrastive Functional Analysis*. Amsterdam/Philadelphia: John Benjamins.

Choi, Seongsook. 2003. Serial verbs and adjunction. In *Proceedings from the first CamLing workshop*. Cambridge: Cambridge University Press.

Christiansen, Thomas. 2011. *Cohesion: A Discourse Perspective*. Bern: Peter Lang.

Clark, Herbert H., and Susan E. Haviland. 1977. Comprehension and the given-new contract. *In Discourse Production and Comprehension*, ed. Roy O. Freedle, 1–40. Hillsdale, N.J.: Erlbau.

Collins, Michael, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (ACL '05), 531-540. Stroudsburg, P.A.: Association for Computational Linguistics.

Cornish, Francis. 2009. Text and discourse as context: discourse anaphora and the FDG Contextual Component. In *Web Papers in Functional Discourse Grammar*, 82, ed. Evelien Keizer and Gerry Wanders, 97-115. Amsterdam: Universiteit van Amsterdam.

Danlos, Laurence. 2008. Strong generative capacity of RST, SDRT and discourse dependency DAGSs. In *Constraints in Discourse*, ed. Anton Benz and Peter Kühnlein, 69-95. Amsterdam/Philadelphia: John Benjamins.

Fabricius-Hansen, Cathrine. 1996. Informational Density - A Problem for Translation and Translation Theory. In *Linguistics*, 34, 521-565.

Fabricius-Hansen, Cathrine. 1998. Information density and translation, with special reference to German - Norwegian - English. In *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, ed. Stig Johansson and Signe Oksefjell, 197-234. Amsterdam: Rodopi.

Fabricius-Hansen, Cathrine, and Wiebke Ramm. 2008. Editors' introduction: Subordination and coordination from different perspectives. In *'Subordination' versus 'Coordination' in Sentence and Text. A cross-linguistic perspective*, ed. Cathrine Fabricius-Hansen and Wiebke Ramm, 1-30. Amsterdam/Philadelphia: John Benjamins.

Firbas, Jan. 1974. Some aspects of the Czechoslovak approach to problems of functional sentence perspective. In *Papers on functional sentence perspective*, ed. Frantisek Daneš, 11-37. The Hague: Mouton.

Granger, Sylviane. 2003. The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies? In *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, ed. Sylviane Granger, Jacques Lerot, and Stephanie Petch-Tyson, 17-30. Amsterdam/New York: Rodopi.

Halliday, Michael A.K. 1967. Notes on transitivity and theme in English part II. In *Journal of Linguistics*, 3, 199-244.

Hasan, Ruqaiya, and Michael A.K. Halliday. 1976. *Cohesion in English*. London: Longman.

Halteren, Hans van. 2008. Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008), 937–944. Manchester, UK.

Hansen-Schirra, Silvia, Stella Neumann, and Erich Steiner. 2007. Cohesive explicitness and explicitation in an English-German translation corpus. In *Languages in Contrast*, 7, 2, 241–265.

Harris, Zellig S. 1952. Discourse analysis. In *Language*, 28, 1–30.

Herslund, Michael. 2000. Le participe présent comme co-verbe. In *La prédication seconde*, ed. Pierre Cadiot and Naoyo Furukawa. Langue française, 127, 86-94.

Hobbs, Jerry R. 1985. *On the coherence and structure of discourse*. Report No. CSLI-85-37. Stanford University: Center for the Study of Language and Information.

Hockett, Charles F. 1958. *A course in modern linguistics.* New York: Macmillan.

Hoey, Michael. 1991. *Patterns of lexis in text*. Oxford: Oxford University Press.

Hopper Paul. J., and Sandra A. Thompson. 1984. The discourse basis for lexical categories in Universal Grammar. In *Language*, 60, 4, 703–752.

Irmer, Matthias. 2011. *Bridging inferences*. Berlin: De Gruyter.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit.

Korzen, Iørn. 2000. Tekstsekvenser / Reference og andre sproglige relationer. In Skytte and Korzen 2000, 65-99 / 161-619.

Korzen, Iørn. 2001. Anafore e relazioni anaforiche. Un approccio pragmatico-cognitivo. In *Lingua nostra LXII*, 107-126.

Korzen, Iørn. 2007a. Linguistic typology, text structure and appositions. In *Langues d'Europe, l'Europe des langues. Croisements linguistiques, Scolia* 22, ed. Iørn Korzen, Marie Lambert and Hélène Vassiliadou, 21-42.

Korzen, Iørn 2007b. Mr. Bean e la linguistica testuale comparativa. Considerazioni tipologico-comparative sulle lingue romanze e germaniche. In *Corpora e linguistica in rete*, ed. Manuel Barbera, Elisa Corino and Cristina Onesti, 209-224. Perugia: Guerra.

Korzen, Iørn. 2009. Struttura testuale e anafora evolutiva: tipologia romanza e tipologia germanica. In *Lingue, culture e testi istituzionali*, ed. Iørn Korzen and Cristina Lavinio, 33-60. Firenze: Franco Cesati.

Krifka, Manfred. 1993. Focus and presupposition in dynamic interpretation. In *Journal of Semantics*, 10, 4, 269-300.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representation of discourse referents*, Cambridge Studies in Linguistics, 71. Cambridge: Cambridge University Press.

Lehmann, Christian. 1988. Towards a typology of clause linkage. In *Clause Combining in Grammar and Discourse*, ed. John Haiman and Sandra A. Thompson, 181–225, Amsterdam/Philadelphia: John Benjamins.

Loock, Rudy. 2010. *Appositive relative clauses in English: Discourse functions and competing structures*, Studies in discourse and grammar series, 22. Amsterdam/Philadelphia: John Benjamins.

Lyons, John. 1977. *Semantics, 1-2*. Cambridge: Cambridge University Press.

Mann, William C., and Sandra A. Thompson. 1987. *Rhetorical Structure Theory. A theory of text organization*. Technical report, ISI/RS-87-190. Los Angeles, C.A.: ISI.

Mann, William C., and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. In *Text* 8, 3, 243-281.

Mann, William C., Christian Matthiessen, and Sandra A. Thompson. 1992. Rhetorical Structure Theory and text analysis. In *Discourse description. Diverse linguistic analyses of a fund-raising text*, ed. William C. Mann and Sandra A. Thompson, 39–78. Amsterdam/Philadelphia: John Benjamins.

Matthiessen, Christian, and Sandra A. Thompson. 1988. The structure of discourse and 'subordination'. In *Clause combining in grammar and discourse*, ed. John Haiman and Sandra A. Thompson, 275–329. Amsterdam/Philadelphia: John Benjamins.

McEnery, Anthony M., Richard Z. Xiao, and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book.* Routledge Applied Linguistics Series. London: Routledge.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. *The Penn Discourse Treebank 2.0*. Technical Report, IRCS-08-02. University of Pennsylvania: Institute for Research in Cognitive Science.

Ramm, Wiebke, and Cathrine Fabricius-Hansen. 2005. *Coordination and discourse-structural salience from a cross-linguistic perspective*. SPRIKreports 30. Oslo: Universitetet i Oslo.

Riazi, Abdolmehdi. 2002. The invisible in translation: The role of text structure. In *The Translation Journal*, 7, 2.
Retrieved from http://accurapid.com/journal/24structure.htm

Rijkhoff, Jan. 2008. Layers, levels and contexts in Functional Discourse Grammar. In *The noun phrase in Functional Discourse Grammar*, ed. Daniel García Velasco and Jan Rijkhoff, 63-116. Berlin/New York: Mouton de Gruyter.

Ruiz Ruiz, Jorge. 2009. Sociological discourse analysis: Methods and Logic. In *Forum: Qualitative Social Research*, 10, 2. Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1298/2882

Skytte, Gunver. 2000. Konnexion og diskursmarkering. In Skytte and Korzen 2000, 621–793.

Skytte, Gunver, and Iørn Korzen. 2000. *Italiensk–dansk sprogbrug i komparativt perspektiv. Reference, konnexion og diskursmarkering, vol. I.-III.*. Copenhagen: Samfundslitteratur.

Skytte, Gunver, Iørn Korzen, Paola Polito, and Erling Strudsholm (eds). 1999. *Tekststrukturering på italiensk og dansk. Resultater af en komparativ undersøgelse / Strutturazione testuale in italiano e danese. Risultati di una indagine comparativa.* Copenhagen: Museum Tusculanum Press.

Stubbs, Michael. 1983. *Discourse analysis*. Oxford: Basil Blackwell.

Stubbs, Michael. 1996. *Text and corpus analysis*. Oxford: Basil Blackwell.

Tannen, Deborah. 1982. Analyzing discourse: text and talk. In *Georgetown University Round Table on Languages and Linguistics*, ed. Deborah Tannen, ix-xii. Washington D.C.: Georgetown University Press.

Teich, Elke. 2003. *Cross-linguistic variation in system and text*. Berlin: Mouton de Gruyter.

Txurruka, Isabel Gómez. 2000. *The semantics of 'and' in discourse*. Technical Report, ILCLI-00-LIC-9. University of the Basque Country: ILCLI.

Vachek, Josef. 1966. *The Linguistic School of Prague*. Bloomingtion: Indiana University Press.

Vallduví, Enric, and Elisabeth Engdahl. 1996. The Linguistic Realization of Information Packaging. In *Linguistics*, 34, 459–519.

Webber, Bonnie, and Rashmi Prasad. 2009. Discourse structure: Swings and roundabouts. In *Structuring information in discourse: the explicit/implicit dimension*, ed. Bergljot Behrens and Cathrine Fabricius-Hansen, 171-190, Oslo Studies in Language 1, 1.

Widdowson, Henry G. 1979. *Explorations in Applied Linguistics*. Oxford: Oxford University Press.

Widdowson, Henry G. 2004. Text, context, pretext. Critical issues in Discourse Analysis. Oxford: Blackwell Publishing.