On Discourse Structure in Italian and Danish

Morten Gylling and Iørn Korzen

Center for Research and Innovation in Translation and Translation Technology
Copenhagen Business School, Dalgas Have 15, 2000 Frederiksberg, Denmark
{mgj.isv | ik.ikk}@cbs.dk

Abstract. This paper examines some typological differences in the discourse structure of Italian and Danish. The results of the study indicate that there are significant differences in information packing in the two languages, especially in their use of deverbalisation. Italian sentences tend to include a larger number of Elementary Discourse Units (EDUs), especially propositions, than Danish. A higher percentage of these is rhetorically backgrounded by means of non-finite and nominalised predicates. Danish text structure, on the other hand, is more informationally linear and characteristic of a higher number of finite verbs and topic shifts. The study also suggests that a more fine-grained classification of non-finite and nominalised EDUs is needed for a complete in-depth analysis of discourse constraints in different language families.

Keywords: Discourse structure, typology, information packing, deverbalisation

1 Introduction

It is well known that discourse structure and discourse constraints are far from identical in the various languages, and if we look at language families, differences of a deeper and typological nature emerge. Hence, studies with a comparative approach can reveal phenomena and characteristics that a monolinguistic approach would not (necessarily) reveal (Herslund and Baron [1]; Korzen [2]). However, there are at present extremely few cross-linguistic textual resources annotated for discourse, in fact, according to Webber et al. [3], they are limited to the ones found in the Copenhagen Dependency Treebanks (CDT). The CDT work with five different Germanic and Romance languages: Danish, English, German, Italian, and Spanish, and annotate them all for four different linguistic layers (apart from part-of-speech): syntax, discourse, anaphora and morphology (Buch-Kromann et al. [4]).

In this paper, we present a few very preliminary results based partly on the work with the CDT and partly on other sources. We shall focus on typological differences in discourse structure between the Scandinavian and Romance languages, represented by Danish and Italian respectively. For the purpose of this paper, we will confine ourselves to two particular diversities, which have to do with text complexity and informational density. In particular, we will focus on sentence length and on the use of deverbalisation, i.e. the way in which propositions can be textually backgrounded by the use of non-finite and nominalised verb forms instead of finite forms.

2 A First Step: Quantitative Analyses

Differences in discourse structure show themselves in many ways, one of which is the simple sentence length, measured as words per sentence[1]. As a first step, and in order to evaluate relatively large amount of data, we compared the complete Danish and Italian Europarl corpus (Koehn [5]). The Europarl texts consist of speeches held by the members of the European Parliament, and most of the speeches (88 %) have been tagged with a LANGUAGE attribute indicating the original language (L1) of the speaker. We then compared the results with those of the texts translated from one of the languages into the other (L2).

Table 1. Sentence length in L1 and L2 Europarl texts.

Original texts (L1)	**Words**	**Sentences**	**Words/sentence**
Italian L1	1,657,592	47,405	34.97
Danish L1	546,425	22,668	24.10
Translated texts (L2)	**Words**	**Sentences**	**Words/sentence**
Italian L2 (texts translated from Danish L1)	571,115	22,154	25.78
Danish L2 (texts translated from Italian L1)	1,845,951	57,574	32.06

Depending on the objectives of a cross-linguistic corpus-based project, either parallel texts (i.e. L1 and L2) or 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) may be best suited as the empirical basis. For projects aimed e.g. at improving machine translation, such as the CDT, parallel texts are suitable because they permit L1–L2 text alignment and evaluation, see Figure 1 below. On the other hand, for projects aimed at descriptive, typological comparisons of discourse structure, the use of parallel texts is ill-suited (McEnery and Wilson [6]; Baroni and Bernardini [7]). The "filter" of the translator and his/her translation strategies "get in the way", and L2 texts risk ending up with a text structure too similar to that of the L1.

As the upper part of Table 1 shows, there is a considerable difference in average sentence length between the Danish L1 and Italian L1 Europarl texts: 10.86 words per sentence or 31.06 %. However, the lower part of Table 1 seems to confirm the problem just mentioned regarding translated L2 texts. As far as sentence length goes,

[1] We are aware of the many reservations to be made when conducting linguistic measurements in this way, but subject to space limitations we cannot go into detail here. However, we feel that the statistical results cited in this section are convincing enough to be taken into account and used as a first indication of profound typological differences between the two languages analysed.

EU translators seem to stick too much to the structure of the L1 text: regarding words per sentence, the Danish L2 texts are 24.82 % longer than the Danish L1 texts, while the Italian L2 texts are 35.64 % shorter compared to the Italian L1 texts. The results clearly show that L2 texts are influenced by the L1 structure when it comes to sentence length.

3 A Second Step: Qualitative Analyses

In order to determine the purpose that the longer Italian sentences serve, we then counted the number of Elementary Discourse Units (EDUs) textualised in each sentence[2] (following Carlson and Marcu [8] who define EDUs as the minimal building blocks of a discourse tree, often textualised by clauses). Here, we discovered a very clear tendency towards a higher number of EDUs in the Italian sentences than in the Danish ones. A statistical count showed that 27.3 % of the Italian sentences contain five or more EDUs. By comparison, only 9.8 % of the Danish sentences contain five or more EDUs.

Many EDUs correspond to propositions, and what may be textualised as one multi-propositional sentence in a Romance language may very well correspond to two or more sentences in a Germanic language. This is the case in the following (parallel) Europarl texts, where one Italian sentence, (1), with one finite verb, a gerund phrase and an infinitive phrase has been translated into Danish, (2), in the form of two coordinated sentences of which the latter consists of two coordinated main clauses[3]:

(1)

(IT)	Signor	Presidente,	*interverrò* [finite verb]	su	INTERREG
[*lit.*]	*Mr*	*President,*	*I will speak*	*about*	*INTERREG*

limitandomi [gerund]	ad	alcuni	aspetti	critici,	anche	per
confining_myself	*to*	*certain*	*aspects*	*critical,*	*also*	*to*

rispettare [infinitive]	ovviamente	i	limiti di tempo	del mio intervento.
respect	*clearly*	*the*	*limits of time*	*of my speech.*

(2)

(DA)	Hr.	formand,	jeg	*vil* [finite verb] gerne	sige	noget	om
[*lit.*]	*Mr*	*President,*	*I*	*will like to*	*say*	*something*	*about*

Interreg.	Jeg	*vil* [finite verb]	nøjes	med	at komme	ind
Interreg.	*I*	*will*	*confine*	*by*	*to touch*	*upon*

[2] This count was carried out on a limited part of the Europarl corpus consisting of 10,000 words in each language. It is our intention to extend our survey to more Italian and Danish Europarl texts shortly.

[3] The official English L2 translation in Europarl sounds: *Mr President, I am taking the floor to speak about INTERREG, but I shall confine myself to a few criticisms, which will clearly also enable me to keep to my speaking time.*

på	visse	kritiske	punkter,	og	det	*er* [finite verb]
on	*certain*	*critical*	*aspects,*	*and*	*that*	*is*

selvfølgelig	også	for	ikke	at	overskride	min taletid.
clearly	*also*	*for*	*not*	*to*	*exceed*	*my speaking time*.

[ep_00-02-14_id=64]

3.1 Finite versus Non-finite Realisation

In cases of parallel (L1-L2) texts, the Copenhagen Dependency Treebanks' DTAG annotation tool and alignment system (Buch-Kromann [9]) can be used to give a precise illustration of the differences:

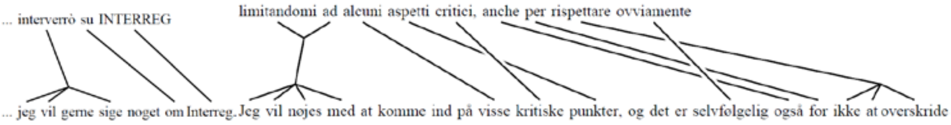
... interverrò su INTERREG
limitandomi ad alcuni aspetti critici, anche per rispettare ovviamente
... jeg vil gerne sige noget om Interreg. Jeg vil nøjes med at komme ind på visse kritiske punkter, og det er selvfølgelig også for ikke at overskride

Figure 1. DTAG alignment of the Italian and Danish example in (1)-(2) (excerpt)

As (1)-(2) and Figure 1 show, the Italian structure with two non-finite verbs corresponds to a Danish structure with two finite verbs. This difference proves to be another distinct characteristic feature of Italian and Romance languages in general on the one side and Danish and Scandinavian languages in general on the other, i.e. a much higher tendency towards hypotaxis and especially non-finite predicate realisation in the former than in the latter. Since non-finite and nominalised verb forms are completely "unmarked" as to grammatico-semantic and pragmatic features such as person[4], tense, mood, aspect and illocution, these values are entirely inherited from – or interpreted on the basis of – the matrix or main clause. Therefore, a non-finite structure is indeed pragmatically and semantically dependent on the main clause, and all non-finite proposition realisations manifest a particular rhetorical backgrounding (as rhetorical satellites, to use the RST terminology) of the proposition in question (Lehmann [10]). With the lack of person marking, the non-finite structures generally express an inherent subject/topic continuity (a topic shift requires a finite verb), which means that the situation or event is evaluated and interpreted as related and less crucial to the on-going topic than the situation or event textualised with a finite predicate.

Inspired by Lehmann [10] and Hopper and Thompson [11], Korzen [12] and later work operate with the deverbalisation scale consisting of five levels shown in Table 2:

[4] In nominalised verb forms the person may appear as a secondary valency complement, e.g. *L'arrivo di* ***John*** *–* ***John****'s arrival.*

Table 2. Deverbalisation scale (Korzen 2007)

0. finite verb in main clause (e.g. John *arrived* late)	↓
1a. finite verb in subclause in the indicative (e.g. I know John *arrived* late)	
1b. finite verb in subclause in the subjunctive (e.g. I hope John *arrived* in time)	
2. non-finite verb (e.g. *Having arrived* late, John missed his train)	
3. nominalised verb (e.g. At John's *arrival*, everybody else left the party)	

The further down on the scale a proposition is textualised, the fewer are the grammatical and pragmatic features expressed by the verb (the more "deverbalised" it is), and the more semantically and rhetorically subordinated and incorporated in the matrix clause is the proposition. Comparing Scandinavian and Romance propositions, there is a very clear tendency for the former to be textualised at levels 0 and 1a (level 1b does not exist as a particular inflected form in Scandinavian), whereas in the Romance languages, all five levels are used much more consistently. In order to show that these differences are not limited to particular text types or genres, such as the (generally argumentative) Europarl texts, we looked into the distribution of finite and non-finite/nominalised verbs in a number of (relatively small) corpora of comparable texts belonging to five different types and genres[5]. The numbers in the tree columns in the centre of Table 3 indicate the percentage of propositions realised with finite, non-finite and nominalised verb forms, respectively:

Table 3. Other corpus-based studies of different text types.

		Finite verbs %	**Non-finite verbs %**	**Nominal. verbs %**	**Words**	**Words/ sentence**
a. Legal texts	IT	43.9	24.2	31.9	3,000	31.6
	DA	56.4	10.2	33.4	1,690	20.1
b. Techni-cal texts	IT	47.5	26.8	25.9	4,883	23.8
	DA	80.7	9.5	9.9	4,974	13.7
c. News-groups	IT	61.1	23.1	15.8	4,193	19.7
	DA	75.8	11.5	12.7	1,826	16.5
d. Web-sites	IT	54	27	19	4,473	24.0
	DA	84	8	8	3,458	12.0
e. Written narratives	IT	52.8	44.2	3.0	4,050	21.7
	DA	88.0	12.0	0.01	4,592	20.9
f. Oral narratives	IT	72.8	27.1	0.1	8,659	–
	DA	93.6	6.4	0	9,077	–

[5] The legal texts are the Italian and Danish acts on insolvency and dissolution of marriage; the technical texts are descriptions of the production of sugar from sugar beets; the arguments of the newsgroups are diets and coffee; the websites are those of two Italian and two Danish chocolate factories, and the written and oral narrative texts are retellings of two Mr. Bean episodes produced by a group of Italian and a group of Danish university students.

The mentioned columns clearly substantiate the claim of a statistically significant difference between Danish and Italian text structure, independently of text type or genre. The columns to the right confirm the tendency mentioned in section 2 of longer sentences in Italian than in Danish, even though the differences vary somewhat from text type/genre to text type/genre.

4 Conclusions and Further Steps

The higher number of EDUs per sentence in Italian texts and the higher percentage of non-finite predicate realisation provide a higher informational density and structural complexity in Italian than in Danish sentences. Italian sentences tend to include more propositions, of which a higher percentage is backgrounded by means of non-finite and nominalised predicates. This results in a multi-layered and hierarchic information structure, characterised by a high degree of topic continuity, in which the various events are evaluated with respect to their importance to the ongoing topic.

On the other hand, Danish text structure is more informationally linear and characterised by a higher degree of topic shifts. Each sentence holds fewer EDUs, and the various events tend to be textualised more chronologically one after the other and with finite verb forms that permit subject changes. Relatively more events are described as having (more or less) the same importance to the particular topic of the given sentence.

In order to arrive at a deeper and more detailed description of the typological differences in Danish and Italian discourse structure, and with a particular focus on argumentative texts, the next steps in our investigation will be, first, to include more data from the Europarl corpus (mainly non-translated texts); secondly, to develop a more fine-grained subdivision of EDUs with distinctions between finite/non-finite/nominalised predicates and between subordinated and coordinated clauses; and thirdly, to insert the findings in a general typological framework that includes other linguistic layers, such as lexicalisation, syntax and anaphora, as well. Each of these steps will include the use of the above mentioned DTAG annotation tool, which facilitates not only the handling of large quantities of data, but also the following quantitative analyses. The results will hopefully provide us with a more precise and detailed knowledge of the typological differences between Scandinavian and Romance discourse structure, differences which are of importance also for syntax (e.g. in the choice of subject type and voice) and for anaphora (e.g. null-forms vs. pronominal forms), phenomena that we will elaborate on in future work.

Acknowledgments. This work was supported by a grant from The Danish Council for Independent Research | Humanities (FKK). We thank the anonymous reviewers for their useful comments.

References

1. Herslund, M., Baron, I.: Language as World View. Endocentric and Exocentric Representations of Reality. In: Copenhagen Studies in Language, vol. 29, pp. 29--42. Samfundslitteratur, Copenhagen (2003)
2. Korzen, I.: Hierarchy vs. Linearity. Some Considerations on the Relation Between Context and Text with Evidence from Italian And Danish. In Baron I. (ed.): Language and Culture. Copenhagen Studies in Language, vol. 29, pp. 97--109, Samfundslitteratur, Copenhagen (2003)
3. Webber, B., Egg, M., Kordoni, V.: Discourse Structure and Language Technology. In: Natural Language Engineering, 1, 1, pp. 1--49 (2010)
4. Buch-Kromann, M. et al.: The Inventory of Linguistic Relations used in the Copenhagen Dependency Treebanks. Copenhagen Business School, Frederiksberg (2010)
5. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit (2005)
6. McEnery, T., Wilson, A.: Corpus Linguistics: an Introduction. 2nd Edition. Edinburgh University Press, Edinburgh (2001)
7. Baroni, M., Bernardini S.: A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. In: Literary and Linguistic Computing, vol. 21, 3, pp. 259--274 (2006)
8. Carlson, L., Marcu, D.: Discourse Tagging Reference Manual. ISI Technical Report, ISI-TR-545 (2001)
9. Buch-Kromann, M.: The DTAG Treebank tool. Technical report, Copenhagen Business School, Frederiksberg (2010)
10. Lehmann, C.: Towards a Typology of Clause Linkage. In: Haiman J., Thompson, S. A. (eds.): Clause Combining in Grammar and Discourse, pp. 181--225. John Benjamins, Amsterdam/Philadelphia (1988)
11. Hopper, P. J., Thompson, S. A.: The Discourse Basis for Lexical Categories in Universal Grammar. In: Language, vol. 60, 4, pp. 703--752 (1984)
12. Korzen, I.: Linguistic Typology, Text Structure and Appositions. In: Korzen, I., Lambert, M., Vassiliadou H.: Langues d'Europe, l'Europe des langues. Croisement Linguistiques. Scolia, vol. 22, pp. 21--42 (2007)