# What Can Contrastive Linguistics Tell Us about Translating Discourse Structure?

**Iørn Korzen, Morten Gylling**

Copenhagen Business School

Dalgas Have 15, DK-2000 Frederiksberg, Denmark

E-mail: ik.ikk@cbs.dk, mgj.isv@cbs.dk

## Abstract

This paper argues that translators can greatly benefit from contrastive studies of discourse structure. Cross-linguistic studies of Italian and Danish point to significant typological differences in information packaging 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 characterised by a higher number of finite verbs and topic shifts. These typological differences are transferred into three simple translation rules concerning 1) the number of EDUs, 2) the rhetorical structure, and 3) the textualisation of rhetorical satellites.

Keywords: discourse structure, information packaging, textualisation, deverbalisation, translation strategies.

## 1. Introduction

Over the last decades, Contrastive Linguistics and Translation Studies have experienced a veritable explosion of interest and attention from scholars in different fields, but the linguistic focus of attention has typically been confined to lexical and syntactic levels. Contrastive studies on discourse structure and intersentential relations, on the other hand, are much less frequent. For instance, there are extremely few cross-linguistic textual resources annotated for discourse. According to Webber, Egg and Kordoni (2010), they are limited to the ones found in the Copenhagen Dependency Treebanks (CDT), which cover five different Germanic and Romance languages: Danish, English, German, Italian, and Spanish. All CDT texts are annotated for four different linguistic layers (apart from part-of-speech): syntax, discourse, anaphora and morphology, see Buch-Kromann et al. (2010).

The research we shall present in this paper is based partly on our work with the CDT and partly on other resources, and we shall focus on two phenomena related to the information and discourse structures of texts, namely on informational density, i.e. the amount of information per sentence, and on text complexity, here defined as the degree of subordination of the text segments that the Rhetorical Structure Theory labels as “rhetorical satellites” (Mann & Thompson, 1987; Mann, Matthiessen & Thompson, 1992; Matthiessen & Thompson, 1988 and later work). Like other scholars, such as Asher and Vieu (2005), we consider these phenomena part of the “information packaging” of a text, a term suggested by Chafe (1976) and later used, especially in connection with given vs. new entities and definiteness, e.g. by Clark and Haviland (1977), Prince (1984) and Vallduvi and Engdahl (1996).

Other cross-linguistic surveys on information packaging have been conducted e.g. by Fabricius-Hansen (1996; 1999), Ramm and Fabricius-Hansen (2005) and Behrens, Solfjeld and Fabricius-Hansen (2010), who investigate English, German and Norwegian, i.e. three Germanic languages. On information density and explicitness in English-German translations, see Hansen-Schirra, Neumann and Steiner (2007). Alves et al. (2010) examine particularly grammatical shifts, e.g. between finite verbs and nominalisations, in the translation process between English and German.

In this paper, we compare two languages of different language families, viz. Danish and Italian, a Scandinavian (Germanic subgroup) and Romance language respectively. Our results regarding Danish confirm the ones obtained by the first mentioned scholars for Norwegian, whereas their findings on English and German are closer to our results concerning Italian. On the other hand, the Italian features presented in the following, are found also in other Romance languages, for which reason we consider it justified to talk about general typological differences between Scandinavian and Romance languages, *ceteris paribus,* with English and German somewhere "in between".

The paper is structured as follows: In section 2, we examine an Italian and Danish corpus of argumentative texts with regard to informational density, measured as the number of words and Elementary Discourse Units (EDUs, cf. Carlson and Marcu, 2001) per sentence. In section 3, we look at text complexity and the textualisation of rhetorical satellites, and in section 4, we formulate our findings as a few relatively simple rules for (human as well as machine) translators that work with Scandinavian and Romance languages.

## 2. Information density

### 2.1. Sentence length

Differences in discourse structure show themselves in many ways, one of which is the simple sentence length, measured as words per sentence[1]. In this context, we used the parallel Europarl corpus, an open source corpus compiled by Koehn (2005). Europarl is a very large multilingual corpus (55 million words) with source and target texts covering all the official languages of the European Union. In fact, the corpus was designed to train and evaluate statistical machine translation, but it can, as we shall see, also be used for other types of cross-linguistic studies. The Europarl texts, which are mainly argumentative (see van Halteren (2008) for a discussion of this), consist of speeches made by the members of the European Parliament from 1996 to 2010, and most of the speeches (88 %) have been tagged with a language attribute indicating the native language (L1) of the speaker. We created a Perl script[2] that extracted all Danish and Italian L1 text from the entire corpus and calculated the average sentence length of all texts. In this context, a sentence is defined as a text segment marked by a full stop, a question mark, or an exclamation mark. We then compared the results with those of the texts translated from one of the two languages into the other (L2). Thus, in Table 1, "Italian L2" texts are translated from Danish into Italian and "Danish L2" texts from Italian into Danish.

| Language | Words | Sentences | Words /sentence |
|---|---|---|---|
| **Italian L1** | 1,657,592 | 47,405 | 34.97 |
| **Danish L1** | 546,425 | 22,668 | 24.10 |
| **Italian L2** | 571,115 | 22,154 | 25.78 |
| **Danish L2** | 1,845,951 | 57,574 | 32.06 |

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

We chose Europarl as the empirical basis for a statistical count 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 for projects aimed e.g. at improving machine translation (such as the previously mentioned CDT) because they permit L1–L2 text alignment and evaluation, comparable texts are generally best suited as the empirical basis for descriptive, typological comparisons like the present one. In such cases, parallel texts are inappropriate because 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. See McEnery and Wilson (2001) and Baroni and Bernardini (2006) for discussions in this regard.

As the upper part of Table 1 shows, there is a considerable difference in average sentence length between the Italian L1 and Danish L1 Europarl texts, a difference amounting to 10.86 words per sentence or 31.06 %. However, the lower part of Table 1 confirms the problem just mentioned regarding translated L2 texts. As far as sentence length goes, EU translators seem to stick very much to the structure of the L1 text: the Danish L2 texts (translated from Italian) are 24.82 % longer than the Danish L1 texts, while the Italian L2 texts (translated from Danish) are 35.64 % shorter compared to the Italian

[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.

[2] We thank our colleague Daniel Hardt for his help in this matter.

L1 texts. When it comes to sentence length, these L2 texts are clearly influenced by the L1 structure.

### 2.2 Elementary Discourse Units

At this point we shall return to the concept of "informational density" and define a little more precisely its application in our project. In order to determine the purpose that the more numerous words in the Italian sentences serve, we then counted the number of Elementary Discourse Units (EDUs) textualised in each sentence, using Carlson and Marcu's (2001) classification. This can be a very time-consuming task, since no parser has been trained to do this convincingly, and we therefore randomly selected a limited part of the Europarl corpus consisting of 7,500 words in each language. We confined ourselves to texts of 200-600 words, and we ended up with a subcorpus in each language consisting of 25 texts of an average length of 300 words each. All texts were manually checked with regard to text type (argumentative), speaker (a certain number of different speakers were required), and date (so that not all text were speeches from the same period).

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 contained five or more EDUs. By comparison, only 9.8 % of the Danish sentences contained five or more EDUs.

We also discovered considerable differences in the number of coordinate vs. subordinate clauses. Finite coordinate clauses amounted to 27.2 % of all clauses in the Danish texts, but only to 17.9 % in the Italian texts. Thus, 82.1 % of the Italian clauses were subordinate as opposed to 72.8 % of the Danish clauses. This may not seem a huge discrepancy, but if we examine in detail the distribution of the subordinate clauses, we encounter considerable differences, cf. Table 2:

| | With connectives | Relative clauses | Attribution | Subordinate non-finite clauses |
|---|---|---|---|---|
| IT | 22.4 % | 40.3 % | 13.1 % | 24.2 % |
| DA | 25.8 % | 40.3 % | 22.5 % | 11.4 % |

Table 2: Distribution of EDUs in subordinate clauses in a Europarl subcorpus

The use of connectives (or "discourse cues" in the RST terminology) and the frequency of relative clauses are more or less equal in the two languages, whereas Danish seems to use attribution more often. In our opinion, this difference should be seen not just as a particular linguistic tendency among Danish parliamentarians, but also as a stylistic feature used to add particular pragmatic values to the argument put forward, a point we shall elaborate in the full version of this paper.

However, the most interesting difference lies in the distribution of non-finite clauses. As Table 2 shows, these occur more than twice as often in the Italian texts as in the Danish ones. Furthermore (not shown in Table 2), Italian uses the whole range of non-finite verb forms (gerund, participles, infinitives and normalisations) much more regularly, whereas Danish mostly confines itself to the use of infinitives (the gerund does not exist in Danish).

## 3. Text complexity

The differences in sentence length seen in Table 1 also have an impact on the distribution of 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 Scandinavian. In a sequence of propositional EDUs, P1 + P2, such as the following:

P1: *arrive* (John, in town); P2: *go* (John, home)

P1 can be textualised in different ways (possibly with added adjuncts or other linguistic material), as shown in the "Deverbalisation Scale" in Table 3[3]:

| | P1 textualised as | Textualisation P1 + P2 |
|---|---|---|
| ↓ | a. an independent sentence | *John arrived late in town.* He went straight home. |
| | b. a main clause, part of sentence | *John arrived late in town* and he went straight home. |
| | c. a subordinate finite clause | *Since John arrived late in town,* he went straight home. |
| | d. a subordinate non-finite clause | *Having arrived late in town,* John went straight home. |
| | e. a nominalisation | *Upon his arrival in town,* John went straight home |

Table 3: Examples of textualisation of EDUs.

[3] The scale is based on Hopper and Thompson (1984), Lehmann (1988), and Korzen (1998; 2007; 2009).

The deverbalisation of P1 increases from (a/b) to (e) together with its integration and absorption into the matrix clause. Whereas the finite verb in a main clause, such as (a/b), has its full (language specific) range of grammatico-semantic values 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 (c) loses its independent tense, mood and illocution; these values will be determined and/or expressed by the matrix clause. The non-finite verb in (d) loses all temporal, modal, and aspectual values and cannot render explicit its subject (see however note 4), and the nominalisation (e) is completely integrated in the matrix clause as a second order entity; its valency complements (here *his*) are syntactically reduced to secondary positions or simply left out.

The further down the scale a proposition is textualised, the fewer grammatico-semantic and pragmatic features are expressed by the verb, i.e. the more the proposition is "deverbalised", and the more it is semantically and rhetorically subordinated and incorporated into the matrix clause. In the case of non-finite and nominalised verbs, (d/e), features such as subject, tense, mood, aspect, and illocution are entirely interpreted on the basis of the matrix clause[4]. Therefore, a non-finite or nominalised structure is entirely pragmatically and semantically dependent on the matrix clause, and such structures express a particularly strong rhetorical backgrounding (or explicit satellite status) of the proposition in question. Furthermore, the lack of subject generally entails an inherent topic continuity (a topic shift typically requires a finite verb with an explicit subject), which means that the situation or event in question is evaluated and interpreted as related and less important to the on-going topic than the situation or event of the matrix clause, textualised with a finite predicate.

Cross-linguistic surveys show that textualisation at the levels (d/e) is much more frequent in the Romance languages than in the Scandinavian ones which show a very clear predilection for finite verbs and textualisation at the levels (a/b/c). These tendencies are not limited to particular text types or genres, such as the (generally argumentative) Europarl texts. Table 4 indicates the percentage of propositions textualised with finite, non-finite, and nominalised verb forms in a number of comparable texts belonging to five different text types and genres. The numbers clearly indicate statistically significant differences between Italian and Danish text structure regarding finite and non-finite verb frequency, independently of text type or genre.

| | | **Verb forms (%)** | | |
|---|---|---|---|---|
| | | **Finite** | **Non-finite** | **Nominalised** |
| **a. Legal texts** | **IT** | 43.9 | 24.2 | 31.9 |
| | **DA** | 56.4 | 10.2 | 33.4 |
| **b. Technical texts** | **IT** | 47.5 | 26.8 | 25.9 |
| | **DA** | 80.7 | 9.5 | 9.9 |
| **c. Newsgroups** | **IT** | 61.1 | 23.1 | 15.8 |
| | **DA** | 75.8 | 11.5 | 12.7 |
| **d. Websites** | **IT** | 54 | 27 | 19 |
| | **DA** | 84 | 8 | 8 |
| **e. Written narratives** | **IT** | 52.8 | 44.2 | 3.0 |
| | **DA** | 88.0 | 12.0 | 0.01 |
| **f. Oral narratives** | **IT** | 72.8 | 27.1 | 0.1 |
| | **DA** | 93.6 | 6.4 | 0 |

Table 4: Verb forms in different text types[5]

As stated above, non-finite and nominalised structures explicitly express the satellite status of the proposition in question. Generally – but not necessarily – this is true also of subordinate adverbial clauses, such as (c) in Table 3. On the other hand, the structures in (a/b) of Table 3 are in themselves ambiguous as to mono- or multinuclear interpretation. However, as is well known, the structure in (b), the syndetic coordination with the connective *and* (and cross-linguistic counterparts), often contains a P1 with satellite status, in Table 3 expressing the *cause* of P2. We shall elaborate also on this issue in the full version of our paper[6].

## 4. Perspectives for translation

The differences described above entail a generally higher

[4] We here ignore the subject of 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 well as the "accusative with infinitive" constructions (*Ho visto* ***Luca arrivare*** – *I saw* ***Luca arrive***). In nominalised verb forms the subject may appear as a secondary valency complement, e.g. *L'arrivo di* ***Luca*** – ***Luca's*** *arrival*.

[5] Precise references will appear in the full version of our paper.

[6] 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).

structural complexity in Italian (and Romance in general) than in Danish (and Scandinavian in general). Romance sentences tend to be longer and to include more propositions, of which a higher number is backgrounded by means of non-finite and nominalised predicates. This results in a multi-layered and hierarchical 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, Scandinavian text structure tends to be more informationally linear and characterised by a higher degree of topic shifts. Each sentence holds fewer EDUs, and different events tend to be textualised more chronologically one after the other and with finite verb forms that permit subject/topic changes.

The results of our study can be transferred into three main rules concerning translations from a Romance to a Scandinavian language or vice versa. The rules regard:

- the number of EDUs per sentence: *ceteris paribus,* there are more EDUs and a higher informational density in Romance than in Scandinavian sentences;
- the textualisation of rhetorical structure: there is a higher tendency in Romance than in Scandinavian to distinguish morpho-syntactically between rhetorical nuclei and satellites;
- the textualisation of rhetorical satellites: there is a tendency to textualise satellites at lower levels of the deverbalisation scale (cf. Table 3) in Romance than in Scandinavian.

Naturally, also phenomena such as e.g. the linguistic register and diamesic dimension (e.g. written vs. spoken text) come into play. The higher the register, the more distinct the mentioned cross-linguistic differences. Oral Italian textualisation and some web variants (such as newgroups, see Table 4) are characterised by a certain structural levelling and are therefore closer to typical Danish textualisation.

## 5. Conclusion

It is well known that a good translation does not (generally, at least) follow the source text word for word. But especially between language families, a good translation does not often follow the source text sentence for sentence, either. Profound typological differences such as those regarding informational density and text complexity must be taken into account, and contrastive studies on discourse structure provide necessary and highly useful linguistic insights for human as well as machine translators.

The results of our study – presented above and in the full version of our paper – will hopefully provide us with more precise and detailed knowledge of typological differences between Romance and Scandinavian 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 develop in future work.

## 6. Acknowledgements

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

## 7. References

F. Alves, A. Pagano, S. Neumann, E. Steiner, and S. Hansen-Schirra (2010): Translation Units and Grammatical Shifts. Towards an Integration of Product- and Process-based Translation Research. In G.M. Shreve and E. Angelone (Eds). Translation and Cognition. John Benjamins, Amsterdam/Philadelphia, pp. 109–142.

N. Asher and L. Vieu (2005): Subordinating and Coordinating Discourse Relations. Lingua 115, pp. 591–610.

M. Baroni and S. 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, pp. 259–274.

B. Behrens and C. Fabricius-Hansen (2010). The Relation Accompanying Circumstance Across Languages: Conflict between Linguistic Expression and Discourse Subordination? In D. Shu and K. Turner (eds.). Contrasting Meaning in Languages of the East and West. Contemporary Studies in Descriptive Linguistics, 14. Oxford et al.: Peter Lang, pp. 531–552.

B. Behrens, K. Solfjeld and C. Fabricius-Hansen (2010): The Relation Accompanying Circumstance Across Languages: Conflict between Linguistic Expression and Discourse Subordination? In D. Shu and K. Turner (Eds.). Contrasting Meaning in Languages of the East and West. Contemporary Studies in Descriptive Linguistics, 14. Oxford et al.: Peter Lang, pp. 531–552.

M. Buch-Kromann et al. (2010): The Inventory of Linguistic Relations used in the Copenhagen Dependency Treebanks. Copenhagen Business School, http://copenhagen-dependency-treebank.googlecode.com/svn/trunk/manual/cdt-manual.pdf.

L. Carlson and D. Marcu (2001): Discourse Tagging Reference Manual. ISI Technical Report, ISI-TR-545.

W.L. Chafe (1976): Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In: Li, Charles N. (Ed.). Subject and Topic. Academic Press, New York/San Francisco/London, pp. 25–55.

H.H. Clark and S.E. Haviland (1977): Comprehension and the Given-new Contract. In Discourse Production and Comprehension, R.O. Freedle (Ed.), Hillsdale, NJ: Erlbau, pp. 1–40.

C. Fabricius-Hansen (1996): Informational Density: a Problem for Translation and Translation Theory. Linguistics 34, pp. 521–565.

C. Fabricius-Hansen (1999): Information Packaging and Translation. Aspects of Translational Sentence Splitting (German - English/Norwegian). In *Sprach-spezifische Aspekte der Informationsverteilung*, M. Doherty (Ed.). Akademie-Verlag, Berlin, pp. 175–213.

H. van Halteren (2008): Source Language Markers in EUROPARL Translations. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, August 2008, pp. 937–944.

S. Hansen-Schirra, S. Neumann and E. Steiner (2007): Cohesive Explicitness and Explicitation in an English-German Translation Corpus. Languages in Contrast 7(2), pp. 241–265.

P. J. Hopper and S. A. Thompson (1984): The Discourse Basis for Lexical Categories in Universal Grammar. Language, 60(4), pp. 703–752.

P. Koehn (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit.

I. Korzen (1998): On the Grammaticalisation of Rhetorical Satellites. A Comparative Study on Italian and Danish. In I. Korzen and M. Herslund (Eds.). Clause Combining and Text Structure. Studies in Language, 22, Copenhagen, pp. 65–86.

I. Korzen (2007): Linguistic Typology, Text Structure and Appositions. In I. Korzen, M. Lambert, and H. Vassiliadou. Langues d'Europe, l'Europe des langues. Croisement Linguistiques. Scolia, 22, pp. 21–42.

I. Korzen (2009): Struttura testuale e anafora evolutiva: tipologia romanza e tipologia germanica. In I. Korzen and C. Lavinio (Eds). Lingue, culture e testi istituzionali. Firenze: Franco Cesati, pp. 33–60.

C. Lehmann (1988): Towards a Typology of Clause Linkage. In J. Haiman and S. A. Thompson (Eds.). Clause Combining in Grammar and Discourse. John Benjamins, Amsterdam/Philadelphia, pp. 181–225.

W.C. Mann, C. Matthiessen and S.A. Thompson (1992): Rhetorical Structure Theory and Text Analysis. In W.C. Mann and S.A. Thompson (Eds). Discourse Description. Diverse Linguistic Analyses of a Fund-raising text. John Benjamins, Amsterdam/Philadelphia, pp. 39–78.

W.C. Mann and S.A. Thompson (1987): Rhetorical Structure Theory. A Theory of Text Organization. ISI, Los Angeles, CA, ISI/RS-87-190, pp. 1–81.

C. Matthiessen and S.A. Thompson (1988): The Structure of Discourse and 'Subordination'. In J. Haiman and S.A. Thompson (Eds). Clause Combining in Grammar and Discourse. John Benjamins, Amsterdam/Philadelphia, pp. 275–329.

T. McEnery and A. Wilson (2001): Corpus Linguistics: an Introduction. 2$^{nd}$ Edition. Edinburgh University Press, Edinburgh.

E.F. Prince (1984): Topicalization and Left-dislocation: a Functional Analysis. In Discourses in Reading and Linguistics, Sheila J. White and Virginia Teller (Eds.). Annals of the New York Academy of Sciences 433, Academy of Sciences, New York, pp. 213–225.

W. Ramm and C. Fabricius-Hansen (2005): Coordination and Discourse-structural Salience from a Cross-linguistic Perspective. SPRIKreports 30.

G.. Skytte (2000): Konnexion og diskursmarkering. In G. Skytte and I. Korzen. Italiensk–dansk sprogbrug i komparativt perspektiv. Reference, konnexion og diskursmarkering, Samfundslitteratur, Copenhagen, pp. 621–793.

E. Vallduví and E. Engdahl (1996): The Linguistic Realization of Information Packaging. Linguistics 34, pp. 459–519.

B. Webber, M. Egg, and V. Kordoni (2010): Discourse Structure and Language Technology. Natural Language Engineering, 1(1), pp. 1–49.

GSCL 2011 Workshop “Contrastive Linguistics – Translation Studies – Machine Translation – what can we learn from each other?”

# An analysis of translational complexity in two text types.

Martha Thunes, University of Bergen.
martha.thunes@lle.uib.no

Key words: automatisation of translation, English-Norwegian parallel text, translational complexity, text types.

## Abstract

This paper is based on the study presented in Thunes (2011), where a selection of English-Norwegian parallel texts have been analysed in order to discuss two primary research questions: firstly, to what extent is it possible to automatise, or compute, the actual translation relation found in the investigated parallel texts, and, secondly, is there a difference in the degree of translational complexity between the two text types, law and fiction, included in the empirical material?

By *automatisation* I here understand the generation of translations with no human intervention, and I assume an approach to machine translation based on linguistic information. In the analysed texts the translations have been produced manually; this is not a study of output produced by machine translation systems, and the automatisation issue is not discussed with reference to any particular translation algorithm or system architecture. Rather, it is related to the assumption that there is a translational relation between the inventories of simple and complex linguistic signs in two languages which is predictable, and hence computable, from information about source and target language systems, and about how the systems correspond. Thus, computable translations are *linguistically predictable*, i.e. predictable from the linguistic information coded in the source text, together with given, general information about the two languages and their interrelations. Further, non-computable translations are correspondences where it is not possible to predict the target expression from the information encoded in the source expression, together with given, general information about SL and TL and their interrelations. Non-computable translations require access to additional information sources, such as various kinds of general or task-specific extra-linguistic information, or task-specific linguistic information from the context surrounding the source expression.

In order to answer the research questions, a measurement of translational complexity is applied to the analysed texts. The degree of *translational complexity* in a given translation task is understood as a factor determined by the types and amounts of information needed to solve the task, as well as by the accessibility of these information sources, and the effort required when they are processed.

For the purpose of measuring the complexity of the relation between a source text unit and its target correspondent, I apply a set of four correspondence types, organised in a hierarchy reflecting divisions between different linguistic levels, along with a gradual increase in the degree of translational complexity. In type 1, the least complex type, the corresponding strings are pragmatically, semantically, and syntactically equivalent, down to the level of the sequence of word forms. In type 2 correspondences, source and target string are pragmatically and semantically

equivalent, and equivalent with respect to syntactic functions, but there is at least one mismatch in the sequence of constituents or in the use of grammatical form words. Within type 3, source and target string are pragmatically and semantically equivalent, but there is at least one structural difference violating syntactic functional equivalence between the strings. In type 4, there is at least one linguistically non-predictable, semantic discrepancy between source and target string, and pragmatic equivalence may, or may not, hold. Thus, the type hierarchy is characterised by an increase with respect to linguistic divergence between source and target string, and by an increase in the need for information and in the amount of effort required to translate, i.e. an increase in the degree of translational complexity. Correspondences of types 1–3 constitute the domain of linguistically predictable, or computable, translations, whereas type 4 correspondences belong to the non-predictable, or non-computable, domain, where semantic equivalence is not fulfilled.

This study applies a strictly product-oriented approach to complexity in translation. The four types of translational correspondences should not be seen as translation methods or strategies, but as descriptions of correspondence relations between given source text units and their existing translations. The empirical analysis of translational correspondences does not aim to study what kinds of knowledge a translator has actually used in order to produce a chosen target expression. Rather, it focusses on the kinds of information about source text expressions that are needed in order to produce the translations.

The correspondence type hierarchy can be seen as a fairly general classification model for translational correspondences. Its main principles were originally defined by Dyvik (1993), and further articulated in Thunes (1998). The approach chosen for the present study is an adapted version of the classification model defined by Thunes (1998). The model is also used as a framework for contrastive language analysis in the studies presented by Hasselgård (1996), Tucunduva (2007), Silva (2008), and Azevedo (in progress).

In the present contribution, the empirical method involves extracting translationally corresponding strings from parallel texts, and assigning one of the types defined by the correspondence hierarchy to each recorded string pair. The finite clause is chosen as the primary unit of analysis, and the main syntactic types among the recorded data are matrix sentences, finite subclauses, and lexical phrases with finite clause(s) as syntactic complement. Since syntactically dependent constructions like finite subclauses occur as translational units, the data include nested correspondences where a superordinate string pair contains one or more embedded string pairs. The assignment of correspondence type to string pairs is an elimination procedure where we start by testing each correspondence for the lowest type and then move upwards in the hierarchy if the test fails. The analysis is thus an evaluation of the degree to which linguistic matching relations hold in each string pair. In cases of nested string pairs, embedded units are treated as opaque items, and the classification of a superordinate correspondence is done independently of the degree of complexity in embedded string pairs. Otherwise, it is a general principle that a string pair is assigned the correspondence type of its most complex non-opaque subpart.

The analysis is applied to running text, omitting no parts of it. Thus, the distribution of the four types of translational correspondence within a set of data provides a measurement of the degree of translational complexity in the parallel texts that the data are extracted from. The extraction and classification of string pairs is done manually as it requires a bilingually competent human analyst. The recorded data cover about 68 000 words, and are compiled from six different text pairs: two of them are law texts; the remaining four are fiction texts. Comparable amounts of text are included for each text type, and both directions of translation are covered.

Since the scope of the investigation is limited, the results do not provide a sufficient basis for generalisations about the degree of translational complexity in the chosen text types and in the language pair English-Norwegian. Concerning the automatisation issue, the complexity measurement across the entire collection of data shows that, in terms of string lengths, as little as 44,8% of all recorded string pairs are classified as computable translational correspondences, i.e. as type 1, 2, or 3, and non-computable string pairs of type 4 constitute a majority (55,2%) of the compiled data. As regards the text type issue, the proportion of computable correspondences is on average 50,2% in the law data, and 39,6% in fiction.

In order to discuss whether it would be fruitful to apply automatic translation to the selected texts, I have considered the workload potentially involved in correcting assumed machine output, and in this respect the difference in restrictedness between the two text types is relevant: law text is strongly norm-governed in a way that fiction text is not. Among the recorded data, I have analysed a set of phenomena that have been identified as recurrent semantic deviations between translationally corresponding units, and this shows that within the non-computable correspondences, the frequency of cases exhibiting only one minimal semantic deviation between source and target string is considerably higher among the data extracted from the law texts than among those recorded from fiction. Such cases can be regarded as minimally non-computable string pairs. Among the law data, as much as 45,7% of the correspondences classified as type 4 are minimally non-computable string pairs, whereas among the fiction data, only 10,5% of the compiled type 4 correspondences are minimal ones. In minimally non-computable correspondences, I assume that only a small effort would be required in order to revise an automatically generated target expression according to the standard of manual translation.

For this reason I tentatively regard the investigated pairs of law texts as representing a text type where tools for automatic translation may be helpful, if the effort required by post-editing is smaller than that of manual translation. This is possibly the case in one of the law text pairs, where 60,9% of the data involve computable translation tasks. In the other pair of law texts the corresponding figure is merely 38,8%, and the potential helpfulness of automatisation would be even more strongly determined by the edit cost. That text might be a task for computer-aided translation, rather than for MT. As regards the investigated fiction texts, it appears likely that post-editing of automatically generated translations would be laborious and not cost effective, even in the case of one text pair showing a relatively low degree of translational complexity. In the analysed pairs of fiction texts, there is a clear tendency that non-computable correspondences exhibit several semantic deviations between the corresponding strings. Hence, I expect that the workload involved in correcting potential machine output would be heavy, and I agree with the common view that the translation of fiction is not a task for MT.

This study is intended to be of relevance to rule-based MT since the chosen analytical framework relies on assumptions about how translations can be computed on the basis of formal descriptions of source and target language systems and their interrelations. However, I assume that the general issue of computability underlying this approach likewise applies to statistical machine translation, which is also dependent on the accessibility of relevant and sufficient information in order to predict correct target expressions from available translational correspondences.

In my view, the framework applied in this study could be used as a diagnostic tool for the feasibility of machine translation in relation to specific text types. That is, by applying the method to limited selections of parallel texts of the same type, it would be possible to estimate to what extent the target text could be generated automatically. If the proportion of assumed computable correspondences would exceed a chosen threshold, it might be worthwhile to tune an MT system for the given language pair to the text type in question. Moreover, in order to estimate the editing distance between potential machine output and a given target text norm, it would be interesting to identify the proportion of minimal type 4 correspondences in a given body of parallel texts. Thus, it would be fruitful to extend the classification model by integrating a fifth correspondence type to be assigned to minimally non-computable string pairs.

**References:**

Azevedo, Flávia. In progress. *Investigating the problem of codifying linguistic knowledge in two translations of Shakespeare's sonnets: a corpus-based study.* Doctoral dissertation. Federal University of Santa Catarina, Florianópolis.

Aijmer, Karin, Bengt Altenberg, and Mats Johansson (eds). 1996. *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4–5 March 1994. Lund Studies in English* 88. Lund: Lund University Press.

Dyvik, Helge. 1993. Text Pair Mapper. Unpublished manuscript. University of Bergen.

Hasselgård, Hilde. 1996. Some methodological issues in a contrastive study of word order in English and Norwegian. In: Aijmer et al. (eds), 1996, 113–126.

Johansson, Stig and Signe Oksefjell (eds). 1998. *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies. Language and Computers: Studies in Practical Linguistics* 24. Amsterdam and Atlanta, GA: Rodopi.

Silva, Norma Andrade da. 2008. *Análise da tradução do item lexical* evidence *para o português com base em um corpus jurídico.* Master's thesis. Federal University of Santa Catarina, Florianópolis.

Thunes, Martha. 1998. Classifying translational correspondences. In: Johansson and Oksefjell (eds), 1998, 25–50.

Thunes, Martha. 2011. *Complexity in Translation. An English-Norwegian Study of Two Text Types.* Doctoral dissertation. University of Bergen.

Tucunduva, Camila de Andrade. 2007. *Translating completeness: a corpus-based approach.* Master's thesis. Federal University of Santa Catarina, Florianópolis.