

Abstract

The present study discusses two primary research questions. Firstly, we have tried to investigate to what extent it is possible to compute the actual translation relation found in a selection of English-Norwegian parallel texts. By this we understand the generation of translations with no human intervention, and we assume an approach to machine translation (MT) based on linguistic knowledge. In order to answer this question, a measurement of translational complexity is applied to the parallel texts. Secondly, we have tried to find out if there is a difference in the degree of translational complexity between the two text types, law and fiction, included in the empirical material.

The study is a strictly product-oriented approach to complexity in translation: it disregards aspects related to translation methods, and to the cognitive processes behind translation. What we have analysed are intersubjectively available relations between source texts and existing translations. The degree of translational complexity in a given translation task is determined by the types and amounts of information needed to solve it, 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, we 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, 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. The correspondence type hierarchy, ranging from 1 to 4, is characterised by an increase with respect to linguistic divergence between source and target string, an increase in the need for information and in the amount of effort required to translate, and a decrease in the extent to which there exist implications between relations of source-target equivalence at different linguistic levels.

We assume 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 predictable from the linguistic information coded in the source text, together with given, general information about the two languages and their interrelations. Further, we regard non-computable translations to be 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 our approach, 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.

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 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 complexity measurements of this study are meant to show to what extent we assume that an ideal, rule-based MT system could simulate the given translations, and for this reason the finite clause is chosen as the primary unit of analysis.

The work of extracting and classifying translational correspondences is done manually as it requires a bilingually competent human analyst. In the present study,

the recorded data cover about 68 000 words. They are compiled from six different text pairs: two of them are law texts, and 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, we cannot, on the basis of our analysis, generalise about the degree of translational complexity in the chosen text types and in the language pair English-Norwegian. Calculated in terms of string lengths, the complexity measurement across the entire collection of data shows that 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. On average, the proportion of computable correspondences is 50,2% in the law data, and 39,6% in fiction.

In relation to the question whether it would be fruitful to apply automatic translation to the selected texts, we have considered the workload potentially involved in correcting machine output, and in this respect the difference in restrictedness between the two text types is relevant. 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. For this reason we 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 automatisations 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 is our view 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. Hence, we concur with the common view that the translation of fiction is not a task for MT.