PART I
INTRODUCTION

1 Overview and background

1.1 The study in a nutshell

How complex is the translational relation between two languages, and to what extent may we expect that translation between that pair of languages can be done automatically? These topics constitute one of our primary research questions, and the present study attempts to answer this with reference to the language pair English-Norwegian, and by investigating two specific text types. In order to study the translational relation between two languages, it is necessary to examine its manifestations, and we have thus chosen an empirical approach where we analyse selected extracts of parallel texts as these constitute parts of the extension of the translational relation. By 'parallel text' we understand an original text paired with its translation into another language, and we have investigated human-translated texts since we regard the product of the bilingually competent human translator as a "gold standard" for translation. The extent to which our study can answer the questions raised initially is of course limited to the scope of our empirical analysis. That is, our results apply only to that part of the translational relation between English and Norwegian which is covered by the selected parallel texts. Furthermore, it is not our ambition to find out to what extent it is possible to achieve automatic translation in general; that is an issue far too wide for us.

In this project the translational relation is treated as a theoretical primitive, not to be defined in terms of other concepts. As will be explained in 2.3.1, we distinguish between the translational relation between two language systems and the translational relation between textual tokens of those languages.

The present study applies a method where translationally corresponding text units are classified according to a measure of the complexity of the relation between source

and target expression. In our analysis the basic unit of translation is the finite clause. The complexity measure is based on assumptions concerning a translator's need for information when producing the given target text, and this need for information is analysed in terms of how much information is needed, what types of information this involves, and the effort required in order to access and process them. We assume a scale of translational complexity, and on this scale we have identified four main types of translational correspondence. When a pair of translational units is analysed, it is assigned one of these four types, as a classification of the complexity of the translational relation between the two units. The four correspondence types are organised in a hierarchy, reflecting an increase in translational complexity. Thunes (1998) presents a pilot investigation of these matters, and the method of analysis applied in that study has been adopted, with some modifications, for the project reported on here.

The classification of correspondences involves no evaluation of translational quality as, for instance, in terms of the model by House (1997). Among the empirical data there are occasional instances of unsuccessful translations, but translational quality is by itself no element in the classification of correspondences. Moreover, our notion of translational complexity, being based on information sources for translation, is in principle independent of grammatical complexity, and of factors that may influence the ease or difficulty with which the translator comprehends the source text.[1] Translational complexity is also distinct from the notion of linguistic complexity, which will be discussed in 3.2.3.

In the present study the question of automatisation is directly linked with the notion of computability. We assume that automatic translation between two languages may be achieved to the extent that it is possible to compute the translational relation between those languages. We will discuss this with reference to our categorisation of translational correspondences, and in the light of the empirical investigation we will tentatively draw a borderline for the possibility of automatisation, a line to be drawn on the complexity scale that we apply to the

[1] Grammatical complexity in relation to translation is discussed by Izquierdo and Borillo (2000).

translational relation. Although the results of our analysis are most directly relatable to rule-based machine translation, we assume that the general issue of computability addressed here 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.[2]

Of importance to automatic translation is the issue of text type, and two different text types, narrative fiction and law text, are represented in the analysed text material. The motivation behind this is to investigate whether the degree of translational complexity differs between the two text types, and this is another primary research question. It is an established view that the possibilities for automatising translation are better with respect to texts dealing with restricted semantic domains than with unrestricted texts (cf. 1.4.2.3). The chosen fiction texts represent unrestricted text types, whereas the law texts instantiate restricted text types. We do not intend to decide whether the subject areas dealt with in the selected law texts are true examples of restricted semantic domains, nor to find out whether those laws can rightly be said to be written in sublanguages of English and Norwegian. Our aim will be to focus on the difference in restrictedness between the two text types, and to discuss its impact on translational complexity.

1.2 Information typology

The present study is neither a cognitive nor a psycho-linguistic investigation of translation, and we do not investigate the procedure of human translation. Our approach is to analyse the product of translation, since we assume that an empirical investigation of parallel texts, as instantiations of the translational relation, may serve as a basis for studying translation competence. Thus, our investigation concerns external, intersubjectively available objects: pairs of source and target texts (cf. 2.2.4).

One important topic in the present investigation is the information that is accessible through the competence of translators, and we assume that analysing a

[2] Cf. the presentation of non-linguistic approaches to machine translation in 1.4.2.5.

translation in relation to its original may reveal the types of information included in translators' competence, as well as other types of information accessed by a translator in order to produce a specific target text. Process-oriented translation studies (cf. 1.4.1.3) have tried to develop cognitive models of what is referred to as *translation competence*.[3] That topic will not be pursued, but for the purposes of our study we may sketch a simple and intuitive conception of translation competence as a combination of the following:

(i) Competence in the source language (SL) as well as in the target language (TL), and knowledge of how these two language systems are interrelated.
(ii) Necessary background knowledge of various kinds.
(iii) The ability to assign an interpretation to the SL text by merging the information encoded in the text itself with the information present in the textual context and in the utterance situation.
(iv) The ability to construct a translation which will receive an interpretation in the TL context and utterance situation which is as close as possible to the interpretation of the original, given its purpose.

The various kinds of information that are accessible through translation competence are part of the information needed to produce a specific translation from a given SL expression. The present work aims to describe a typology of information sources for translation, and in this respect, the following main types provide a starting point:

(a) Purely linguistic information, some of which is encoded in the SL expression, and some of which is inherent in a translator's bilingual competence and knowledge of interrelations between source and target language systems.
(b) Pragmatic information from the textual context and the utterance situation of the source expression.

[3] Hurtado Albir and Alves (2009: 63–68) present an overview of different translation competence models; cf. 2.4.2. We discuss the knowledge of translators in 2.4.1.5.

(c) Various kinds of extra-linguistic background information.

In addition to these categories we apply a distinction between general and task-specific information sources. The general sources include information about source and target language systems and their interrelations, as well as information about the world (cf. (i) and (ii) above). These information types are given, and hence easily accessible, in any case of translation. The task-specific sources cover information about a particular piece of source text and the concrete task of translating it into a given target language.

The typology of information sources for translation is presented in 2.4.2 with subsections. Since we describe translational complexity in terms of the amounts and types of information needed to produce a given target expression, the information typology is developed for the purpose of analysing the degree of translational complexity in correspondences between expressions of two languages. In relation to the various information sources for translation, we will in chapters 2 and 3 consider two questions that are decisive for the complexity of translational correspondences: to what extent can the different kinds of information be represented in a finite way, and what is the amount of effort required in order to access and process them?

1.3 The correspondence type hierarchy

As mentioned in 1.1, our scale of translational complexity is captured by a hierarchy of four main types of translational correspondence. The origins of this hierarchy is found in Helge Dyvik's work on an experimental machine translation system, documented in Dyvik (1990, 1995). The four correspondence types will here be briefly presented in order to illustrate how this hierarchy is linked with a translator's need for information when producing a specific target text. We will refer to instances of correspondence types as *(translational) correspondences* or, alternatively, as *string pairs*, i.e. translationally related pairs of word strings. Our notion of 'translational correspondence' is in accord with that of Johansson (2007: 23), who uses the term *correspondences* about "the set of forms in the source text which are found to correspond to particular words or constructions in the target text." Furthermore, we

will use the term *correspondent* to refer to either of the units that constitute a translational correspondence. Hence, this term is neutral between original and translation. Moreover, *correspondent* may refer to entire units of translation, as well as to subparts of them.[4]

1.3.1 Four types of translational correspondence

In this section we present and illustrate the four correspondence types with reference to the finite clause, since it is, as pointed out in 1.1, the basic unit of translation in this study.[5]

The least complex correspondence type is labelled *type 1* and comprises cases of word-by-word translations where source and target string are identical with respect to the sequence of word forms. Cf. string pair (1):

(1a) Hun har vært en skjønnhet. (BV)[6]
'She has been a beauty.'
(1b) She has been a beauty,

Type 2 correspondences are somewhat more complex, since source and target string are not matched word by word, but every lexical word in the source expression has a target correspondent of the same lexical category and with the same syntactic function as the source word. Otherwise, there may be differences between source and target string with respect to the sequence of constituents and/or the use of grammatical form words; cf. string pairs (2) and (3):

(2a) Dessuten virket hun overlegen. (BV)
'Also looked she haughty.'
(2b) She also looked haughty.

[4] The notion of 'translational correspondence' is further discussed in 4.3.1.
[5] Our units of analysis are defined in 4.3.2.
[6] *BV* refers to the author Bjørg Vik; see the list of primary sources. When examples of translational correspondences are given, the source text is always given under (a) and the target text under (b). Punctuation is reproduced as given in the primary text.

(3a) Leiligheten var ufattelig rotete. (BV)
'Flat.DEF was unbelievably untidy.'[7]
(3b) The flat was unbelievably untidy.

In (2) source and target string differ with respect to constituent sequence: (2a) has a fronted adverbial (*dessuten*), followed by the verb *virket*, and then by the subject *hun*, whereas in (2b) the subject *she* is in the initial position, followed by the adverbial *also*, and then by the verb *looked*.[8] In example (3) the English definite article *the* in the translation is not matched by any word form in the source sentence.

In *type 3* correspondences, translational complexity is still higher as they involve greater structural discrepancies between source and target than correspondences of type 2 do: there is at least one structural difference violating syntactic functional equivalence between the strings, but there is no mismatch between original and translation on the semantic level; cf. string pair (4):

(4a) Hildegun himlet lidende mot taket og svarte med uforskammet høflighet: (BV)
'Hildegun rolled-eyes suffering towards ceiling.DEF and answered with brazen politeness'
(4b) Hildegun rolled her eyes in suffering towards the ceiling and answered with brazen politeness.

There are two main reasons why string pair (4) cannot be assigned a type lower than 3. Firstly, the Norwegian intransitive verb phrase *himlet* corresponds with the English expression *rolled her eyes*, which consists of a transitive verb phrase and a noun phrase (NP) functioning as direct object. But these expressions correspond semantically: the Norwegian verb *himle* ('roll one's eyes') describes the activity of rolling the eyes of the agent, and since this information is inherent in the lexical meaning of *himle*, the existence of the referent of the English NP *her eyes* is implied by the Norwegian verb phrase.[9] Secondly, the adverb phrase *lidende* ('suffering') in (4a) is

[7] The label *DEF* will be used as a shorthand for the grammatical feature *definite*.
[8] (2a) illustrates subject-verb inversion in Norwegian. The example is also discussed in 3.3.3.1.
[9] The mismatch between the verb phrases *himlet* and *rolled her eyes* may be described as a conflational divergence; cf. 1.4.2.3.

of a different syntactic category than the preposition phrase *in suffering* in (4b), and the English preposition *in* is not matched by any lexical unit in (4a). But the two expressions *lidende* and *in suffering* correspond semantically: both phrases modify the action described by the verb phrases *himlet* and *rolled her eyes*, and the verbs *lide* and *suffer* are denotationally equivalent.[10]

Finally, in *type 4* correspondences complexity is even higher: in such cases there are discrepancies between original and translation not only on the structural level, but also on the semantic; cf. string pair (5):

(5a) Her kunne de snakke sammen uten å bli ropt inn for å gå i melke-butikken eller til bakeren. (BV)
'Here could they talk together without to be called in for to go in milk-shop.DEF or to baker.DEF'

(5b) They could talk here without being called in to go and buy milk or bread.

In (5) there is a semantic difference between the corresponding expressions *for å gå i melkebutikken eller til bakeren* ('to go to the milk shop or to the baker') and *to go and buy milk or bread*. The italicised expressions do not denote the same activities, although we may infer from background information about the world that both activities may have the same result, i.e. the purchase of milk or bread.

A central aspect of the correspondence type hierarchy is the increase in the degree of translational complexity from type 1 upwards. A parallel to this increase in complexity is found in Vinay and Darbelnet's (1995) set of seven translation procedures, which are presented "in increasing order of difficulty", ranging from the simplest method of translation to the most complex.[11] Although this is an interesting similarity, the present correspondence type hierarchy is not related to Vinay and Darbelnet's classification of methods. Our type hierarchy is designed for the purpose of analysing existing correspondences between source and target texts, and must not be associated with the notion of translation procedures.

[10] Denotational equivalence between expressions of different languages is discussed in 6.3.2.

[11] The quotation is taken from Venuti (2000: 92), where an overview of the seven procedures is presented. Pages 31–42 of Vinay and Darbelnet (1995) are reprinted in Venuti (2000: 84–93).

We have applied the method to one language pair only, English-Norwegian, but in principle it is a language-pair independent approach. However, occurrences of the lower correspondence types require a certain degree of structural relatedness within a given language pair: if SL and TL are structurally unrelated, the lowest types may not be found. On the other hand, in the case of languages that are very closely related, such as Norwegian, Danish, and Swedish, the most complex types may be rare.

The basic principles of the correspondence type hierarchy were originally described by Dyvik (1993), and the hierarchy is further developed in Thunes (1998), where the notion of translational complexity is discussed in relation to information sources needed in translation. Another contribution made by the latter is that subcategories of the main correspondence types 3 and 4 have been identified and explored. A further development of the correspondence type hierarchy is here discussed in chapter 3, where the information processing structure of individual translation tasks is related to each correspondence type. Chapter 4 provides a new discussion of criteria for the identification of analysis units, and for the assignment of correspondence type to string pairs.[12]

In our analysis we assume that a translator's need for information is greater in translational correspondences of the higher types than in those of the lower types. If we consider a human translator, this may not seem so obvious: a bilingual person will simply produce a target text without paying much attention to the amount of information he or she uses when doing so, perhaps with the exception of those cases where the translator really needs to think twice, and possibly check with reference works etc., to create a target text. The increase in a translator's need for information from correspondence type 1 to 4 is easier to grasp if we imagine giving the translation tasks to an automatic translation system, and the discussion will be related to the PONS system (Dyvik 1990, 1995) since its design is the main source of inspiration for the correspondence type hierarchy.

[12] Cf. 1.5 for more information on how the content of this thesis is organised.

1.3.2 The background for the correspondence type hierarchy

The PONS machine translation system is endowed with information about source and target language systems and their interrelations; this may be seen as a model of the translator's bilingual competence. The first step of the translation task is to analyse the input, a procedure which is comparable to a translator's reading and understanding of the source sentence. The analysis provides the system with information about the syntactic structure of the input text, which is then compared with information about source and target language interrelations. Through this comparison, the PONS system is able to choose between three different modes of translation, according to the complexity of the translation task. In practice, the system identifies cases where the syntactic structure of the source text is matched by the target language and exploits this match for the purpose of target text generation.

If the entire structure of the input text has a match in the TL grammar, the system will translate word by word, thus producing a type 1 correspondence. In such cases generation of the target sentence requires information about the word order and syntactic structure of the source sentence, and about the translationally corresponding TL word forms.

In other cases the PONS system may find that the source sentence structure is matched by the target grammar except for at least one difference with respect to constituent sequence and/or the presence of grammatical form words. The system may then be said to translate constituent by constituent, and will produce a type 2 correspondence. In such cases the generation of the target sentence requires information about the syntax of the input text, about the syntax of the structurally deviating parts of the target text, and about the translationally corresponding TL word forms. In this way translation requires a greater amount of information than in type 1 correspondences.

In cases where the PONS system finds that with respect to the function and/or category of at least one lexical word, the syntactic structure of the source sentence cannot be matched by the target language, the system will produce a full semantic analysis of the input, and use a semantic representation of the source sentence as the basis for target text generation. The result will be a type 3 correspondence, and

generation of the output sentence requires semantic information about the input text together with structural and lexical information about the target language. Cases of types 2 and 3 have in common that solving the translation task requires information about how the target text will deviate structurally from the source text. But since type 3 correspondences exhibit greater structural discrepancies between source and target than type 2 correspondences do, the translation task requires a more thorough linguistic analysis than in the case of type 2, and hence the need for information is greater.

With respect to type 4 correspondences, we assume that they are not included in the set of translations that could be computed by the PONS system, since they are cases where purely linguistic information is insufficient, and the translation task requires additional information sources, such as extra-linguistic background information and discourse information derived from a wider linguistic context.

In our study the distribution of the four correspondence types within a body of parallel texts is meant to serve as an estimate of its degree of translational complexity, and this estimate may be seen as an indication of to what extent automatic translation is feasible within the investigated texts. That is, the complexity measurement may indicate how far it is possible to simulate human translation for the specific language pair, text types, and translational choices as instantiated by the analysed parallel texts. We will later argue that the limit of automatisation is defined by the limit of linguistic predictability in the translational relation, and it follows from the organisation of the correspondence type hierarchy that the distinction between the linguistically predictable and the non-predictable is drawn between types 3 and 4.[13] It should be emphasised that in the present project the question of automatisation is discussed without reference to the architecture of any particular machine translation system, although the analytical framework is inspired by the PONS design.

[13] Linguistic predictability in the translational relation is defined in 2.3.2. For details on correspondence types 3 and 4, see chapter 3.

1.3.3 Related contributions

Hasselgård (1996) employs a slightly modified version of the correspondence type hierarchy as defined by Dyvik (1993). In Hasselgård (1996) the method is used for classifying correspondences between translationally aligned sentences in a small-scale investigation of word-order differences between English and Norwegian. Adapted versions of the correspondence type hierarchy as presented in Thunes (1998) are used by Tucunduva (2007), Silva (2008), and Azevedo (in progress), all of which are studies where the model is applied for the purpose of analysing and describing translational correspondences in parallel texts. These contributions are concerned with the language pair English-Portuguese, and they study various types of text.[14]

A related approach is provided by Merkel (1999), who combines translation studies, natural language processing, and corpus linguistics in a study where the main theme is correspondence relations in parallel corpora. His contribution includes a model for describing various kinds of structural and semantic correspondences between translationally aligned sentences in a Swedish-English parallel corpus. The aim of the analysis is to find out to what extent the translations exhibit changes in structure, function, and content in comparison to the originals, and this, in turn, is done to investigate differences between text types and translation methods.[15]

Another approach is found in Macken (2010), who presents research on automatic alignment of translational correspondences below sentence level, i.e. words, phrases and chunks. This is relevant to the present study since the data compiled in our investigation also include a large number of correspondences involving sub-sentential units.[16] In Macken's project different alignment tools have been tested against a manually aligned Dutch-English reference corpus. Her presentation of various categories of sub-sentential translational correspondences contains many similarities to the correspondence type hierarchy as described in Thunes (1998), in particular

[14] I am indebted to Marco Antonio Esteves da Rocha, of the Federal University of Santa Catarina, for information on the studies presented in Tucunduva (2007), Silva (2008), and Azevedo (in progress).
[15] Cf. chapters 10, 11, and 12 in Merkel (1999).
[16] Cf. the presentation of extraction criteria in 4.3.2.

regarding the kinds of linguistic properties that are shared (or not shared) by translationally matched units (cf. Macken 2010: 33–36).

1.4 Relevant fields of research

The present study draws on insights from several disciplines: general and computational linguistics, translation studies, and corpus linguistics, to mention some. 1.4 with subsections will present a selection of topics from a few relevant fields, i.e. translation studies, machine translation, and parallel corpus linguistics. Since a key issue in our investigation is the division between linguistically predictable and non-predictable translations, and since this is related to the limit of automatisation, the discussion will give more weight to machine translation than to the other disciplines.

1.4.1 Translation studies

The very notion of 'translation' has so far not been commented on. The present study is limited to written translation, and by 'translation' we will understand the act of transferring a text from one language into another. Koller (1992: 81, referring to Wienold 1980) points out that translation belongs to a group of several kinds of text reproduction, all kinds involving an original text and a new version of it. In addition to translation, examples of such activities are popularisation, the writing of abstracts, and creating children's versions of literary works. The latter activities have in common with translation that they may be performed across languages, but translation differs from them in (at least) one important way, as translation does not allow any of the differences between original and version typically found in the other kinds of text-reproduction. Still, it is not unproblematic to define 'translation' whether by delimiting the concept of translation or by specifying its set of necessary and sufficient properties.

There is, however, an intuitive concept of translation, one that has intersubjective validity. Halverson (2000) shows that 'translation' is a prototypical concept: firstly, the concept displays "graded membership" in the sense that certain types of

translation seem to be more central members of the category than others, and, secondly, the concept has "fuzzy boundaries" in the sense that there are gradual slides, and not discrete leaps, from 'translation' to related concepts. In agreement with the prototypical view of translation we regard the following characteristics as *central* to the concept of interlingual translation:

(i) Taking into account differences between source and target language systems, the translated version will as far as possible convey the same meaning as the source text.
(ii) The sender of a translated text is identical to the sender of its original.[17]
(iii) Taking into account cultural differences between the source and target language communities, the recipient group of the translation is as parallel as possible to that of the original in the source language community.
(iv) The communicative function of the target text is as parallel as possible to that of the source text.

In relation to this list of characteristics, at least two reservations can be mentioned. Firstly, it follows from a prototypical view of translation that not all of the properties (i)–(iv) must be present in everything that can qualify as 'translation'. Secondly, we do not imply that if these four properties are present in a translation, it will necessarily be a fully satisfactory version of the original.

Although the study of translation may be traced back to antiquity, it is only after the Second World War that the field has become a substantial area of research. During this time translation researchers have tried to form theories explaining translational phenomena, and they have constructed models of the relationship between originals and translations, as well as models of the translation process. Theoretical frameworks like those of general linguistics and contrastive language analysis have been applied in order to define translation models. The heterogeneity of the field is illustrated by the fact that it is difficult to find a single cover-term for all

[17] However, in the view of Koller (1979, 1992), where translation is described as a bilingual communication process, the translator is regarded as the sender of the target text; cf. 1.4.1.1.

its branches. *Translatology*, *translation theory*, *translation studies*, or the German *Übersetzungswissenschaft* — none of these expressions can serve as a fully neutral label in the sense that all translation scientists would accept it as a cover-term.[18]

As stated in 1.2, our approach is to analyse the product of translation. There are basic differences between studying, respectively, the product and the process of translation. We may directly observe the translation product as a text available to our perception, whereas the translation process is not as easily observable. Special elicitation techniques are required to examine the mental processes behind the production of the target language text. Hence, the distinction between product- and process-oriented approaches is important when describing the field of translation.

Chesterman (2005) provides a critical review of the terms and concepts that have been used over the years in various studies of the changes that may occur when a source text is translated into a target text (cf. 6.2.1). In this connection he discusses the opposition between product and process orientations, and he observes that many translation researchers are not entirely "clear about whether the focus is on processes themselves or the results of processes" (2005: 19). To illustrate his point he gives several examples from various contributions, and presents a possible explanation for the confusion: many of the terms used to describe translational changes often have a linguistic form that is "ambiguous between a process reading and a result reading" (2005: 20).[19] It would require a larger study of the field to support this position, but the main points argued by Chesterman (2005: 17–22) seem indisputable: lack of terminological stringency across the field works against conceptual clarity, and it is necessary to start by defining the concepts in order to improve the terminology of translation studies.

In our view, the difference between product and process orientations can be perceived as a continuum rather than as a dichotomy. In 1.4.1.1–3 we will present a selection of approaches illustrating this. At one extreme there are models describing the product of translation in a declarative way, thus focussing on the relation between

---

[18] For this piece of information the author is indebted to Dagmar Čejka. However, according to Baker (1993: 234), *translation studies* is the most common term, and we will mainly use this expression when referring to the field.

[19] Chesterman's examples of such terms are *compression*, *omission*, and *compensation* (2005: 20).

original and translation. If such descriptions are truly declarative, they specify sets of relations holding at the same time between certain entities, and they may be interpreted as declarations of static facts about the entities involved. At the other extreme there are procedural models describing the translation process. A procedural approach implies that the object of study is described in terms of a set of operations that will produce that object, and hence the description is of a dynamic kind. In positions between the declarative and the procedural there are models describing the product of translation partly by paying attention to the steps leading from source to target text, and there are models describing the translation process, but to some extent in terms of the relation between source and target text.

Sections 1.4.1.1–3 are not intended as a full overview of the various directions within translation studies, nor as a historical outline. Our aim is to present a few contributions chosen as representatives of certain positions within the field, and in chapter 2 we return to the division between product and process orientation. For surveys of different theoretical approaches in translation studies, as well as information on the historical development of this area of research, see e.g. Venuti (2000), Gentzler (2001), and Munday (2008, 2009). Kittel et al. (2004, 2007) provide a more detailed reference work on translation studies, and Baker (2010) presents a state-of-the-art view of the field. Moreover, chapter 4 in Munday (2008) gives an overview of product- and process-oriented approaches, respectively.

1.4.1.1 Product-oriented approaches to translation

Among the topics of interest to product-oriented studies of translation there are phenomena such as particular features of translated texts, and relations between source texts and their translations. In such studies it is relevant to probe the texts by means of different linguistic analyses, i.e. analyses concerned with domains like syntax, semantics, discourse, textual macrostructure, and stylistics.

Starting at the end of the continuum mentioned, where we find clearly product-oriented approaches, we may discuss Werner Koller's explication of the concept of 'translational equivalence'. His work is representative of the so-called "equivalence tradition", one of the linguistically oriented approaches within studies of translation.

According to Koller (1992: 81, 215; 1995: 196), 'translation' is defined by means of 'translational equivalence': we have a proper instance of translation when there exists an equivalence relation between an original in the source language and a translated version in a target language.[20] His definition of translation is provided with a description of a set of different frames of reference under which translational equivalence may hold (1992: 214–216; 1995: 196–197). In that manner he decomposes the relation into five different equivalence types: denotational, connotative, text-normative, pragmatic, and formal-aesthetic equivalence (1992: 216). Each such type specifies properties with respect to which the source and target texts should be equivalent. Denotational equivalence pertains to the extra-linguistic state of affairs described by the source text, whereas connotative equivalence deals with the connotations conveyed by the expressions used in original and translation respectively, especially through choice of words, level of style, the use of particular sociolects or dialects, and the like. Text-normative equivalence is determined by text type-specific norms of language use, and formal-aesthetic equivalence by the formal aspects of source and target text. Finally, pragmatic equivalence pertains to the communicative function of the texts, to the recipient of the translation, and to her/his capacity of understanding the translated message.

The concept of 'translational equivalence' has been much debated, and Koller's view of it is not the only one. In general, 'equivalence' is always equivalence with respect to a set of given properties and is not in itself a gradable concept. Hence, problematic aspects of the notion of 'translational equivalence' arise from the fact that cultural differences, and differences with respect to grammatical and lexical structure between source and target language, often makes it impossible to achieve translational equivalence with respect to all desirable properties. In practice, then, the translation task is to create a target version that is equivalent to the original with respect to as many as possible of relevant properties, and the selection of relevant properties will depend on the purpose and communicative function of the source text.

[20] Translational equivalence, in the sense used in translation studies, is not an *equivalence relation* in the terms of formal logic.

Or in the words of Juliane House: "… the translator has to set up a hierarchy of demands on equivalence that [he] wants to follow" (1997: 26).

Koller focuses on the result of the translation process in relation to its starting point, and his view is thus directed towards phenomena which are available to inter-subjective investigation. He has also addressed the translation process, but, as noted by Krings (1986: 9), Koller (1979: 112) regards its investigation to be a task for psycholinguistics. Elsewhere he has presented translation as a bilingual process of communication: first, the source text is communicated from the original sender to the translator in the role of recipient; second, the translator transfers the source text to the target language, and, third, the target text is communicated from the translator, as a secondary sender, to the final recipient (1979: 123–125; 1992: 106–107).[21] However, Koller does not present this as a model of the translation process, but as an account of aspects of the translation situation.

Another important contribution among the product-oriented approaches is the work of Gideon Toury (1995) on norms in translation. In relation to the task of studying the norms that govern translation, he states explicitly that the norms themselves are not available for observation; it is only the products of norm-governed translation behaviour that can be studied in order to detect the norms (1995: 65).[22] However, Toury's work is not as purely product-oriented as Koller's account of translational equivalence. Since norms control the work of translators, they exist during the translation process, and the study of norms aims at revealing how they influence the production of target texts. Toury (1995: 88) describes this study as "an attempt to gradually reconstruct both translation decisions and the constraints under which they were made." In Toury's approach there are several points of relevance for the present investigation, but due to the elements of process orientation, it will not be discussed further here.

[21] Bhatia (1997: 204) also takes the view of the translator as a secondary sender, at least implicitly, when stating that translation "is an attempt to communicate someone else's message through another language."
[22] In 2.2.1 we will discuss the principled difference between behaviour and the products of behaviour.

1.4.1.2 An intermediate position

Koller's view of translation as a communicative process may lead over to other approaches intermediate to the extreme positions of product and process orientation. An example of these is Juliane House's model of translation quality assessment, as laid down in House (1997). House may be said to belong to the functionalist tradition within translation studies, in which the communicative purpose of translation is a central notion. Her model is based on pragmatic theories, and on cross-cultural studies of the language pair German-English.

House assumes that translation quality assessment requires a theory of translation, and that different theories will yield different views of translation quality and of its evaluation (1997: 1). In her theory the equivalence concept is central, and she holds equivalence, as a relation between source and target text, to be the fundamental criterion for translation quality evaluation (1997: 25, 29). Her equivalence concept pertains to the preservation of meaning, and she views it as a functional and communicative notion. With respect to translational equivalence she distinguishes three aspects of meaning: semantic, pragmatic, and textual meaning (1997: 30–31).

Another central ingredient of House's theory of translation is her distinction between overt and covert translation (1997: 29, 66–70). In the case of overt translation the product is presented to the target language recipient as nothing but a translation, and the original links to the source language culture are preserved. A typical example is translated foreign language literature. In the case of covert translations, for instance translated user instructions, the target text appears as an original text, so that the function of the translation in the target language community corresponds to that of the original in the source language community. In order to achieve this, covert translations are subject to what House describes as "cultural filtering", i.e. a process in which the translator must "transmute the original such that the function it has in its original and situational environment is re-created in the target linguaculture" (1997: 163).

House's method for translation quality assessment (1997: 36–45) involves three steps: First, the source text is subject to a detailed linguistic and pragmatic analysis in order to detect its function, or "textual profile", and the source text profile will be the

norm for the assessment of quality in the translation. Second, the same kind of profile analysis is applied to the target text, and, third, the textual profile of the translation is compared to that of the original, in order to evaluate the degree of match. A high degree of match will be the mark of good translation quality.

House's concern with translation quality assessment is by nature product-oriented, as it is impossible to evaluate translation quality without analysing the result of the translation process. However, her work also reveals a concern with the translation process: to some extent the cultural filter gives an account of what goes on during translation, or at least of certain consequences of the process. Moreover, her distinction between overt and covert translations is tied to the issue of translation strategy, as the two types of translation represent different tasks: in overt translation the translator must make as few alterations as possible, whereas in covert translation the translator must erase, or adapt, all traces of the source language culture or community (1997: 164).[23]

1.4.1.3 Process-oriented approaches to translation

In a process-oriented study of translation focus is directed towards the translator's activity during translation. Since this activity primarily takes place in the translator's brain, it is not sufficient to analyse the translation situation and the result of the translation process in relation to the source text. In order to discover the inside workings of this instance of a black box it is necessary to use the methods of psychology.

However, translation research offers examples of theorists who have created models of the translation process even if they have not carried out psycholinguistic studies of it. One of them is Eugene A. Nida, whose contributions from the 1950ies onwards were of great value to the development of modern translation studies. Nida

[23] Deliberately, we have so far not defined the notion of 'translation strategy', as it is not part of our object of study, but occasional references to it are inevitable when we discuss translation and its product. We will merely apply an intuitive understanding of the concept, and use the expression *translation strategy*, or *translation method*, to refer to the set of actions chosen, either deliberately or not, by the translator during the creation of the target text. See Palumbo (2009: 131–133) for a discussion of the notion, including an overview of relevant references. Within the field of machine translation, a special meaning is attributed to 'translation strategy'; cf. 1.4.2.4.

was strongly interested in translation activity, and his research was based on wide experience with Bible translation. His works were also rooted in descriptive and theoretical linguistics, as well as in anthropology.[24]

In Nida's model the translation process consists of three main stages: analysis, transfer, and restructuring (Nida 1975: 80–95). The analysis stage identifies relations of meaning and reference, as well as the connotative values of the source text. Thus, analysis yields a disambiguated version of the source text, which can be transferred to the target language at a level "deeper" than that of surface structure. It is Nida's opinion that transfer takes place at a level where languages exhibit a greater degree of similarity than at the surface. The transfer stage he describes as a process of redistribution, operating on structures of semantic features representing the source text, and this process will most likely modify the source text meaning. The process of restructuring is to a large extent determined by the target language system, and it involves both formal and functional aspects, the latter requiring that the translation is made equivalent to the original with respect to communicative effect.

Although Nida's model is a procedural description, it captures linguistic effects of the translation process rather than the nature of the process itself. Nida was, however, aware of the psychological aspects of translating, but at the time the field of psychology did not offer adequate methods for probing the cognitive activities of a translator at work. Nida carried out this research while behaviourism still held a strong position, and according to the behaviourist paradigm the processes inside our brain could not be investigated through truly scientific methods, since they could not be observed directly (see Lörscher 1991: 67). The behaviourists had thus renounced the method of introspection, which had been applied during the late 19th and early 20th century as a tool for the investigation of mental activity.

After the exit of behaviourist views, there has been a revival of the use of introspection in psychological research. The methodology aims at externalising internal data, thus making them available to intersubjective investigation, and the

[24] See the "Introduction" to Nida (1975).

means to this end is verbal reporting.[25] In the 1980ies the elicitation technique named Think-Aloud Protocols (TAPs) came into use among researchers concerned with the mental processes involved in the act of translating. The use of Think-Aloud Protocols is based on a psychological model in which human cognition, including translation, is understood as information processing, and a cognitive process is "seen as a sequence of internal states successively transformed by a series of information-processing steps" (Lörscher 1991: 71).[26] Moreover, the model assumes that we are able to monitor our own cognitive processes, and hence the act of "thinking aloud" will provide access to the steps of information processing. In TAP studies the informant, in this case a translator, is typically asked to report, unselectively, everything that goes through her/his mind when performing the translation task, i.e., literally, to think aloud, while the reporting is audio- or video-taped. Other actions, such as note-making and consulting reference works, are also documented. TAP studies involve substantial criticism of previous models of the translation process. E.g., Krings (1986: 8) is of the opinion that those models do not describe what he deems to be the real facts of the translation process. Rather, he views them as attempts at analysing the translation process in terms of categories external to the process, such as the categories of linguistic analysis.

Within process-oriented translation studies, Krings (1986) is worthy of attention. Jääskeläinen (1999: 40) describes it as the "first extensive published TAP study", and according to Palumbo (2009: 92), it is generally seen as the "beginning of the process-oriented research tradition in translation studies." On the basis of his empirical data Krings makes certain generalisations on the global course of a translation task (1986: 178–187). He splits the process into three phases, pre-processing, main processing and post-processing.[27] Moreover, he finds it necessary to distinguish between translation *from* the translator's first language (L1→L2) and translation *into* her/his first language (L2→L1), the reason being that he finds more

[25] On verbal reporting see Ericsson and Simon (1984, 1993), or Krings (1986: 63–64). Lörscher (1991: 69–76) presents an overview of the development of introspective methods in modern research on cognition, and in particular on language learning and translation.

[26] For information on the TAP method, see also Toury (1995: 234–238), Jääskeläinen (1999, 2000), and Jakobsen (2003).

[27] In Krings' words: "Vorlauf", "Hauptlauf", "Nachlauf".

similarities between the informants' strategy choices in translation from L2 into L1 than in translation from L1 into L2.[28] Also, in L2-to-L1 translation there are two main types of translation problems, i.e. reception problems and production problems, whereas in L1-to-L2 translation production problems are dominant and reception problems nearly absent.

We may look briefly at the three phases in Krings' model. Pre-processing basically involves reading through the source text. Some of the informants omit this phase in L2-to-L1 translation. Otherwise during this phase, there is generally great variation with respect to the efforts put into identifying, and possibly solving, translational problems. During the main processing phase all subjects perform the bulk of the work required by the translation task. At this stage there is more variation with respect to strategy choices in L2-to-L1 translation than in L1-to-L2 translation. In the latter case all subjects translate sentence by sentence, in sequence. Finally, the post-processing phase, if not omitted, involves correcting and completing the target text, typically in the way of proof-reading.

It is interesting to compare Krings' model with earlier models of the translation process. The earlier models typically comprise either two or three different stages in the process. In general, two-phase models contain an analysis stage and a reconstruction stage, and three-phase models comprise analysis, transfer, and synthesis.[29] There is, however, no isomorphy between Krings' model and earlier three-phase models. Although it may not be evident from our brief presentation of Krings' work, it is a fact that in each of the three phases he has identified there may occur elements of analysis, transfer, as well as synthesis, depending on the translator's strategy. Moreover, Krings' study shows that some translators do not perform any pre- or post-processing. On the other hand, in the earlier models of translating the three stages of analysis, transfer, and synthesis are discrete, and none of them are dispensable.

Above all, TAP studies have shown that there is great variation among translators with respect to translation strategies. Another interesting finding is the distinction

[28] In the case of Krings (1986), the informants' L1 is German and their L2 is French.
[29] Cf. Wilss (1977: 95f, 1978: 15f), cited by Krings (1986: 6).

between processes performed automatically by the translator and processes requiring conscious decision-making (cf. Jääskeläinen and Tirkkonen-Condit 1991). Although the method of verbal reporting has clearly been helpful, and TAPs represented a breakthrough in translation studies, there are also shortcomings in these techniques. Hurtado Albir and Alves (2009), who provide a comprehensive overview of process-oriented research on translation, mention several weak points (2009: 69): The major problem is that TAP studies document the informants' subjective view of their own activity, and not necessarily the correct facts about it. Moreover, the method is intruding in that the subjects are aware of being observed, and perform verbalisation along with translation. Also, TAPs do not reveal unconscious or automatic processes. In more recent years the methodological trend has been to combine verbal reporting with other techniques (cf. Hurtado Albir and Alves 2009: 70–71). These may include traditional ones like interviews and questionnaires, and more modern ones, such as measuring brain activity, and logging the keystrokes and eye movements of translators at work. Hurtado Albir and Alves (2009: 72–73) conclude that the empirical methods of process-oriented translation research still need refinement. As methods improve, interesting discoveries about the cognitive aspects of translation are sure to be made.

1.4.2 Machine translation

We will understand *machine translation* (MT), or *automatic translation*, as the use of a computer program to translate text in one natural language into another. Thus, the notion of machine translation does not include computerised bilingual dictionaries, since they apply to the translation of single words, possibly including multi-word expressions. On the other hand, it does include systems able to translate spoken language (speech-to-speech translation), but the present discussion of MT will primarily be limited to the translation of written text.

Jurafsky and Martin (2009: 898) divides the field into classic and modern machine translation, an opposition reflecting the important distinction between *rule-based MT* and *statistical MT*. In the former approach the translation procedure relies on information about source and target language and their interrelations, whereas in

the latter approach translations are computed on the basis of statistical information about existing correspondences in large bodies of parallel texts.[30] As indicated in 1.1, the results of our product-oriented study are in principle also relatable to statistical MT, but the following presentation will focus on the classic, rule-based approaches, since our principal interest, in relation to automatic translation, lies in the question of how far it is possible to simulate human translation by processing linguistic sources of information.

Machine translation started as a research field; commercial applications gradually appeared, and MT has grown into a quite heterogeneous field with a great variety of applications. Several authors have presented overviews of the field, and their different contributions show that machine translation systems can be described and categorised in various ways, depending on which aspect of the field the description is focussed on.[31] Some of these aspects will be presented in 1.4.2.2–5, while the remainder of this section will discuss the division between experimental and commercial MT systems, which may answer questions like: who builds MT systems, and where are they used?

Experimental translation systems are typically developed within research institutions, and for the purpose of investigating pure research issues, such as the testing of formalisms for computational language descriptions. Although the development of an MT system normally requires a team of researchers working together, experimental systems may be the work of one or only a few researchers. Such systems may also be used for educational purposes, especially in university courses on computational linguistics. Normally, experimental MT systems are limited with respect to the coverage of the grammars and vocabularies of the languages they are applied to. The PONS system, discussed in 1.3.2, is an example of an experimental MT system; it may be described as a development environment where the user creates his or her own lexicons and grammars for source and target language, thus

[30] The dichotomy between rule-based and statistical MT is also mentioned in 1.4.2.1, and it is further discussed in 1.4.2.5.

[31] See for instance Hutchins (1986), Lehrberger and Bourbeau (1988), Hutchins and Somers (1992), Dorr et al. (1998), Nirenburg et al. (2003). Chapter 25 in Jurafsky and Martin (2009) provides a more recent introduction to machine translation. Other possible information sources for updates on the field are the journal *Machine Translation* and proceedings from the conference series Machine Translation Summit.

experiencing how the encoding of linguistic information will enable the system to translate.

Commercial systems are developed for the purpose of reducing the amount of work needed by professional, human translators. Typically, they are developed by teams where different specialists, such as computational linguists, programmers, lexicographers, and terminologists, work on different modules that together constitute a translation tool. The overall motivation behind the design of the system will be cost effectiveness: a net profit must be the outcome when the expenses of development, which can be substantial, are measured against the eventual benefits from saving translators' work hours, and possibly also from selling the tool to other users. Thus, with respect to system design, operational efficiency will be more important than matters such as the soundness of theoretical assumptions underlying language descriptions encoded in the system. A prerequisite for the usefulness of a commercial system is that grammar and lexicon modules cover the vocabulary and set of constructions found in the texts to which the system is applied, and this normally means that such information modules are large and expensive to build. It is also common that commercial systems are designed for text types special to restricted, technical domains, since technical texts tend to exhibit a controlled vocabulary and limited inventory of sentence types, which means that such MT systems will not necessarily need broad-coverage grammars and lexicons. Typically, commercial MT systems have been developed by, or for, large multinational enterprises, of which IBM is a well-known example, and for the purpose of translating technical documentation. Some commercial systems have been available for decades, with new and improved versions appearing now and then.

It may seem as if experimental and commercial MT systems have belonged to separate camps with no mutual interests, but that is not true. There are many examples of system developers with experience from research institutions who have joined in the construction of commercial systems, and issues like efficiency, cost effectiveness, and broad coverage are clearly not uninteresting to developers working in the research sector, although they may not be the dominating research aims. Moreover, the German Verbmobil project (Wahlster 2000) is an example of coopera-

tion between research and science: in a large and prestigious project academic and commercial interests joined forces to develop a system for the translation of spontaneous speech.

In addition to experimental and commercial systems, in recent years certain MT applications have become available to everyone with access to the Internet. These tools are incorporated in search engines, so that if an information request identifies a document in a foreign language, the system can offer an automatic translation of that document. This will typically be a translation of low quality, but it may be sufficient for the user to decide whether it is worthwhile making further efforts to access the information contained in that document.

1.4.2.1 A brief historical overview

The earliest attempts at constructing mechanical systems for automatic translation were made in the first half of the 20th century (Hutchins 1986: 22), but with no success. After the Second World War the advent of modern computer technology paved the road for new attempts, and in the 1950ies machine translation was among "the first non-numerical applications of computers" (Hutchins 1986: 16). In the early years the major sources of motivation and funding behind MT development was found among military and intelligence authorities, notably in the United States and the Soviet Union. It was the era of the Cold War, and in many nations intelligence agencies were busy collecting information about enemy countries, so that there was a great demand for translating text produced in the languages of those states. During the war, computers had been used for coding and decoding military messages, and it is not surprising that in this context automatic translation was seen as a promising tool. MT activities were not only initiated in the US and Soviet Union, but also in Japan and certain Western European countries, as well as in Canada from the late 1960ies.

Early work on machine translation was strongly inspired by information theory, in the US especially by the work of the information theorist Warren Weaver, who argued that translation basically involved decoding the source language text into target language symbols (Weaver 1949). At the time, similar conceptions of

translation were also harboured by several translation researchers: to generalise, translation was seen as decoding the source text message and recoding it in the target language.[32] In the first generation of MT systems the encoding of linguistic information was based on shallow language descriptions. Roughly, the first systems could be seen as implementations of bilingual dictionaries with certain reordering rules for accommodating structural differences between SL and TL. The lack of linguistic sophistication in the early systems is understandable: theoretical linguistics did not yet offer linguistic models suitable for computational implementation, and the capacity of available computer technology put narrow limits on the amount of language information that could be encoded, and on how it could be done. Still, there were great expectations with respect to what would be achieved.

In the 1960ies the optimism vanished since there were still no really successful results of machine translation development. Even if computer technology was continually improving, there had been no substantial breakthrough, and MT researchers came to realise that certain fundamental problems related to linguistic issues had to be solved before better MT systems could be built. It became a widespread view that since natural languages are in so many ways ambiguous, it would be an unreasonable goal to achieve fully automatic, high quality translation of unrestricted text. As early as in 1960 the influential researcher Yehoshua Bar-Hillel explained why: “A human translator, in order to arrive at his high quality output, is often obliged to make intelligent use of extra-linguistic knowledge which sometimes has to be of considerable breadth and depth. Without this knowledge he would often be in no position to resolve semantical ambiguities. At present no way of constructing machines with such a knowledge is known, nor of writing programs which will ensure intelligent use of this knowledge.”[33]

Then, in 1966 the famous ALPAC report appeared. It was presented by an evaluation committee appointed by the US state agencies that were the main sponsors of MT activities. The report brought MT into disrepute, and efficiently drained away research funding in the United States as it concluded that the field had so far been a

[32] Cf. the discussion in Koller (1992: 89–92) of early models of translation.
[33] The quotation is taken from Nirenburg et al. (2003: 62), where Bar-Hillel (1960) is reprinted.

failure, that there remained too many unsolved fundamental problems, and that human translation would anyway be more cost effective than developing automatic translation. After the ALPAC report US research environments turned their focus to artificial intelligence and fundamental issues in computational linguistics. In other countries the change was not so acute; work on MT development continued although it was not carried out on such a large scale as had been the case in the United States.

Even if perfection was not achieved, workable MT systems did appear on the market during the 1960ies, and the fact that they were actually used shows that there clearly was a need for MT as a supplement to human translation, even if it involved a considerable amount of revision by translators. An important market was the translation of technical documentation in industry.

In the late 1970ies the pessimism that spread during the sixties was slowly giving way to renewed, but careful, optimism. In 1977 the Canadian MT system METEO® was completed for the purpose of translating weather forecasts between English and French. The system was a success and in operation for about two decades. This achievement strengthened the view that machine translation was suited for texts with a controlled vocabulary and a limited set of possible syntactic constructions. Moreover, it fuelled new interest in MT development, and during the 1980ies research activities were increasing in a range of countries. Achievements made since the 1960ies in several fields of science now offered far better conditions for creating automatic translation. Computer hardware had improved greatly; new programming techniques had been developed, and formalisms more suitable to computationally implementable language descriptions had been developed within linguistics.

Thus, by the beginning of the 1990ies a range of different MT projects had appeared in many countries across the world, and, in comparison to early machine translation, systems were now of a quite different quality with respect to computational efficiency as well as sophistication in the treatment of linguistic phenomena. Also, research efforts were not any longer limited to languages with large numbers of speakers (like English, Russian, French, Japanese, etc.), but MT development was also carried out for small languages, such as those in Scandinavia. Moreover, multilinguality had become an important design issue: multilingual MT systems are

not limited to one language pair, but are constructed for translating between several languages, and should easily facilitate the inclusion of new language pairs. Hence, modularity was an important design issue, so that linguistic information was to a larger extent than before kept separate from the actual translation procedure in the systems. This was another way in which MT had come to differ from the earliest systems, where translation procedures generally were strongly dependent on the linguistic differences between specific pairs of languages.

In 1993 Sergei Nirenburg pointed out that machine translation had "recaptured its place as the single most important application of computational linguistics and natural language processing" (1993: v). Since then research funds have come from national governments as well as from commercial interests, and MT has retained an important, although today not dominating, position within the larger field of language technology. Here MT has had to compete over research grants with other activities like voice recognition, speech synthesis, word sense disambiguation, and the building of language resources.

Statistical approaches to machine translation emerged in the early 1990ies. While commercial systems were still rule-based, MT conferences during that decade became dominated by the discussion of statistical methods and the evaluation of their performance. Gradually, research efforts were directed mainly towards statistical MT, as it appeared to be highly promising. However, after 2000 there has been a growing awareness in the field that further improvement of performance requires that the statistical methods are augmented with some processing of linguistic information, an approach often described as *hybrid* (cf. Dorr et al. 1998: 35).

1.4.2.2 Degree of automation

One important aspect of rule-based machine translation systems has been degree of automation. Some MT systems have been fully automatic, whereas others have required interaction with a human user. E.g., Hutchins (1986: 19), and Sager (1994: 290) classify systems according to a scale ranging from fully automatic translation to human translation with no machine aids. In fully automatic translation (or *batch* systems) the user only needs to enter the source text and wait for the system to output

a translation. In interactive systems some kind of intervention is required from the human user during the translation process. This could amount to resolving linguistic ambiguities in the source text, or entering target words for certain SL words whose translations are unknown to the system, or also selecting the most appropriate target text when the system produces alternative translations. The operation of such interactive systems can be described as *human-aided* machine translation.

Other important kinds of human intervention in translation tools are known as *pre-* and *post-editing*, respectively. Pre-editing involves preparing the input so that the MT system is able to compute a translation given the linguistic information encoded in the system. The pre-editor must remove from the source text syntactic structures and lexical items which are not covered by the language descriptions of the system. Pre-editing may also involve inserting syntactic labels in the source text so that the system will be able to resolve linguistic ambiguities.

Post-editing of the output of an MT system means that a human who is competent in both SL and TL revises the target text according to demands on translation quality. This is really the same task as revising a draft version of a "manual" translation, but, as noted by King (1986: 6), there is great variation between human and machine translation with respect to the amount of post-editing needed and the types of errors made. When a considerable amount of post-editing is required, the phenomenon at hand may be described as *machine-aided* translation rather than as MT proper. Post-editing is still a current topic in machine translation, and the amount of necessary post-editing of the output has always been an important criterion in the evaluation of the performance of MT systems.

In relation to the degree of automation, there is perhaps one kind of tool used in machine-aided translation that is particularly relevant, i.e. the *translation memory* (TM).[34] This is defined by Palumbo (2009: 127–128) as "[a]n electronic database containing translated texts stored together with their originals," and the texts "are normally segmented into units one sentence long." Clearly, as Merkel (1999: 43) has observed, translation memory tools are particularly useful for maintaining consisten-

[34] Chapter 8 in Macken (2010) provides a survey of translation memory systems, and reports on an evaluation of the performance of two available TM systems.

cy in the translation of types of text with repetitive language, such as technical texts. The latter point is relevant to the dimension of text type, which will be introduced in chapter 4.

1.4.2.3 Challenges for automatic translation

As the history of machine translation shows, automatic translation is a greater challenge than merely decoding the source text and recoding it in target language symbols. Dorr et al. (1998: 4–12) have presented the challenges involved in MT building along two different dimensions, described as operational and linguistic considerations, respectively. Our primary focus will be on the latter kind, and the discussion in this section relates mainly to rule-based MT.

Among the *operational considerations* of machine translation, Dorr et al. (1998: 10) include "extension of the MT system to handle new domains and languages; handling a wide range of text styles; maintenance of a system once it has been developed; integration with other user software; and evaluation metrics for testing the effectiveness of the system." Operational considerations in MT building are of greater relevance to implementation issues than to the linguistic aspects of automatic translation. Hence, we will give more attention to the latter topic than to the former, since our interest lies with the question of automatisation independently of the architecture of any particular MT system.

However, among the operational issues there is some relevance to the present project in the topic of extending a system to new domains and languages. That is, the challenge can be said to be not only to extend, but to build, altogether, those information modules that will serve as lexicons and grammars for source and target languages in an MT system. Without such information sources the system cannot translate.[35] Another prerequisite for successful translation is that those information modules cover the lexical inventory and set of linguistic structures found in the input texts at hand. Realistic requirements in operative MT systems are lexicons with tens of thousands of entries, and grammars with hundreds of rules. MT system builders must

[35] Cf. our discussion of information sources for translation in 1.2, and in 2.4 with subsections.

collect this information from somewhere, and another prerequisite is a grammar formalism for the representation of lexical entries and grammar rules. Normally, creating such linguistic information modules involves a lot of manual work since it is impossible to convert traditional dictionaries and grammars into computational ones without major adaptations.

The following quotation indicates what a great challenge it is to build linguistic information modules for an MT system: "Providing the linguistic knowledge for an entire language is truly a staggering task. In fact, no single human language has yet been fully described in a form usable by computers" (Grishman and Kittredge 1986: ix). Now, about 25 years later, this is still true. One possible way of meeting the challenge is to tune an MT system for texts from a restricted semantic *domain*, and by this we normally understand a certain technical field, such as a specific trade, a branch of industry, a field of science, etc. The group of speakers associated with a restricted domain typically share some domain-specific knowledge which is not part of the common knowledge of the speakers of the entire language community.[36] Furthermore, in such domains only subsets of alternative meanings of certain ambiguous words will be probable, and texts dealing with restricted domains will normally share certain linguistic characteristics. More specifically, discourse related to a restricted semantic domain typically employs a limited set of preferred linguistic constructions, and a set of technical *terms*, whose meanings are unambiguous.

Such discourse can be tied to the concept of a *sublanguage*, a notion which was originally given a mathematical definition by Zellig Harris (1968).[37] Here, leaving the mathematical properties aside, we will emphasise the fact that a sublanguage is a well-defined subset of a given language. The meanings of its expressions are a subset of the meanings expressed by the general language, and it is regarded as a more manageable task to describe the grammar and lexicon of the sublanguage than of the general language. Thus, if an MT system is designed for a restricted semantic domain, it is not necessary to build lexicons and grammars for entire languages, as it is sufficient to cover the given sublanguages of SL and TL. It may be necessary to

[36] Cf. Kittredge (1987: 59).
[37] For information on this, see Kittredge and Lehrberger (1982: 1), and Kittredge (1987: 59–60).

describe constructions and lexical items which do not belong to the general languages, since they belong only to the source and target sublanguages, but that will be a limited task. The effect of tuning an MT system to a specific domain and sublanguage is to avoid many of the problems involved in achieving automatic translation of general text, problems we will mention in connection with linguistic challenges for MT. The disadvantage is that extending the system to other domains demands that new sublanguage lexicons and grammars must be created.

Linguistic challenges for machine translation are referred to by Dorr et al. (1998: 4–10) as *linguistic considerations* in MT development, and like Dorr et al. (1998: 4) we will divide them into problems related to source text analysis, to target text generation, and to the mapping between source and target language. Our main focus will be on types of analysis problems because identifying the correct interpretation of the input is crucially important to successful machine translation.

Analysis problems in automatic translation are, above all, caused by ambiguity in natural language expressions, i.e. the fact that more than one possible interpretation may be associated with a word, phrase, or sentence. One possible way of sorting the types of ambiguity that cause analysis problems is to divide them into lexical, structural, and referential ambiguity (cf. Thunes 1994: 4–6). In general language use ambiguity phenomenona are extremely frequent, whereas in sublanguage texts their incidence is lower, as indicated above. Ambiguity phenomena indeed highlight the difference between the human translator's ability to interpret a source text and the way in which an MT system is able to understand input text. The types of ambiguity that cause trouble in automatic translation are normally resolved effortlessly by humans, because we continuously make use of contextual and extra-linguistic information when reading a text. Thus, if a word, phrase, or sentence has more than one possible interpretation, we filter out all improbable alternatives to the intended interpretation by means of information surrounding the ambiguous expression. An MT system, on the other hand, normally works sentence by sentence and must rely on the information that is linguistically coded in the given input sentence, and the analysing system will try all possible readings of ambiguous expressions, and their combinations. This may yield a large number of possible interpretations, and in MT

systems it is difficult to simulate a significant amount of the kind of inferences used by the human translator, mostly subconsciously, when improbable interpretations are filtered out.

Lexical ambiguity covers phenomena like homonymy, homography, and polysemy. Here we shall not go into great detail, but mention a classic example of homonymy: the English noun *bank* has at least two meanings: 'river bank' and 'financial institution', respectively. Given a sentence like *They camped by the bank of the river*, a human reader with general world knowledge would never consider the second meaning of *bank*, but for an MT system it is a challenge to identify the intended meaning of the ambiguous noun *bank* in order to choose a correct target language equivalent. This is a problem especially since it is extremely rare that the translations of homonymous source words are homonyms, too. A possible way of handling this is to encode, in the lexical information associated with *bank*, the semantic conditions governing the proper use of the different meanings, and to do so in a principled way is a challenge for the designer of the lexicon of the MT system.

Lexical ambiguity frequently involves cases where a lexical item is ambiguous with respect to syntactic category, such as the English word form *increase*, which can be either a verb or a noun, thus constituting a pair of homographs. In automatic translation, such categorial ambiguity can be resolved by parsing the local syntactic context: e.g., if an article like *an*, or *the*, immediately precedes the word form *increase*, then the analysing system will be able to choose the noun reading. Lexical ambiguity is a kind of analysis problem that researchers have tried to amend by integrating automatic word sense disambiguation (WSD) in MT systems. In simplified terms, WSD methods work by estimating the probability of a given sense in relation to other words occurring in the context of the ambiguous word, thus exploiting the fact that different senses of a word tend to be used in different types of contexts.[38] However, Ide and Wilks (2006: 54) observe that WSD tools do not seem to improve the performance of MT systems substantially. One reason may be that although quite successful WSD tools have been developed, an even higher degree of

[38] For an introduction to WSD, see chapter 20 in Jurafsky and Martin (2009).

accuracy is required, since disambiguation errors during analysis can have quite damaging effects (cf. Ide and Wilks 2006: 65). Another reason may be that in systems where categorial ambiguities are anyway resolved by syntactic parsing of the input the usefulness of separate WSD modules is probably limited (cf. Ide and Wilks 2006: 55–56).

Structural ambiguity can be described as the phenomenon where an expression has more than one possible interpretation because the expression can be partitioned into phrases in more than one way. A standard example for illustrating structural ambiguity is (6):

(6) I saw the man with the binoculars.

(6) can be interpreted as the statement that the referent of *I* either saw a man by means of a pair of binoculars, or saw a man who was carrying binoculars. Choosing the intended interpretation requires extra information from the context in which the expression is uttered. For a human recipient it is trivial to access and use such information; for the analysis procedure in an MT system it is not, especially if the system works sentence by sentence and is unable to retain information from the linguistic context preceding each input sentence.

Such structural ambiguity is not necessarily a translational problem: if the target language is ambiguous in the same way, then the ambiguity must not be resolved before translating. (7) is a Norwegian translation of (6), and the possible syntactic analyses and interpretations of (7) are an exact parallel to those of (6):

(7) Jeg så mannen med kikkerten.
'I saw man.DEF with the binoculars.'

There is a fair degree of structural relatedness between English and Norwegian, which in this case helps the translation task. If the target language is Japanese, which is a structurally unrelated language, it is necessary to resolve the source sentence ambiguity because the two interpretations require different translations. The first reading of (6), 'I saw the man by means of the binoculars', can be translated as (8):

(8) Watasi wa booenkyoo de otoko o mita.
'I-TOPIC binoculars-INSTRUMENT man-OBJECT saw.'

In (8) the particle *de* marks the noun *booenkyoo* as an instrument in the described situation. The second reading of (6), 'I saw the man who was carrying the binoculars', can be translated as (9):

(9) Watasi wa booenkyoo o motte iru otoko o mita.
'I-TOPIC binoculars-OBJECT carrying was man-OBJECT saw'

In (9) the particle *o* marks the noun *booenkyoo* as an object of the verbal phrase *motte iru* ('was carrying').

Referential ambiguity occurs in cases where it is possible to assign more than one referent to an anaphoric pronoun. Example (10) may illustrate this:

(10) There is a ship on the harbour, and it is crowded with tourists.

In (10) there are two possible antecedents for the pronoun *it*: *a ship* and *the harbour*. Again, translation may require that the intended interpretation is found if the two different alternatives must be translated in different ways. That would be the case when translating (10) into Norwegian, as in (11) or (12), where the use of italics indicates the possible binding relations between antecedent noun phrase and anaphoric pronoun:

(11) Det ligger *et skip* på havnen, og *det* er fullt av turister.
'It lies a ship on harbour.DEF, and it (i.e. the ship) is full of tourists.'

(12) Det ligger et skip på *havnen*, og *den* er full av turister.
'It lies a ship on harbour.DEF, and it (i.e. the harbour) is full of tourists.'

In (11) the neuter gender of the noun *skip* ('ship') requires the neuter form of the anaphor *det*, while in (12) the masculine form of the anaphor *den* agrees with the masculine gender of the noun *havn* ('harbour').

The examples used to illustrate structural and referential ambiguity show that when these phenomena occur, the amount of information that is encoded in the linguistic expression itself is insufficient in order to choose one interpretation rather than another. For automatic translation it is a true challenge that information from a wider linguistic context, or even from background world knowledge, is necessary to resolve the ambiguities.[39]

Having discussed analysis problems for MT, we will look at *generation problems*, and concentrate on two main categories: first, problems created by lack of isomorphy between lexical distinctions in source and target language, and, second, problems arising when the target language obligatorily expresses grammatical distinctions absent in the source language. These are not the only kinds of problems for generation in MT, but the ones we would like to focus on.[40]

Dorr et al. (1998: 7) refers to the first type as the *lexical selection problem* in target text generation. It is a well-known fact that different languages carve up reality in different ways, and this has the consequence that lexical items in one language only rarely correspond one-to-one with lexical items in other languages.[41] Thus, the challenge for machine translation is that finding the correct target language equivalent for a given source word frequently involves making a choice within a set of possible candidates. E.g. the English verb *know* corresponds translationally with various Norwegian verbs, depending on the linguistic context. Appropriate translations of *know* in the sense used in *Do you know French?* are the verbs *kunne* and *beherske*, whereas in the case of *Do you know what time it is? know* corresponds with the Norwegian verb *vite*. Hence, we may say that *know* is translationally ambiguous. The semantic conditions governing these translational choices are fairly subtle and nontrivial to represent in a format usable in an MT system, and extra-linguistic information about the world may be needed to identify the appropriate target word in a given context. General language words, such as *know*, are normally polysemous, or semantically vague, and hence may cover various senses and have

[39] Cf. comments on *the resolution problem* in 2.4.2.2.
[40] In 3.3.2.2 the second type of generation problem is illustrated by morphological differences between English and Norwegian present tense verbs.
[41] Cf. the discussion in 6.3.2 of denotational equivalence between lexemes of different languages.

several possible translations.[42] Clearly, it is easier to manage the generation task if the input text is written in a sublanguage with a high frequency of technical terms. Typically, technical terms correspond one-to-one with terms in the target language, since it is a characteristic property of terms that they have been designed to be unambiguous.

The second type of generation problems is caused by a fact once formulated by Roman Jakobson: "Languages differ essentially in what they *must* convey and not in what they *may* convey" (1959: 236). Several grammatical categories, of which tense, number, and gender are typical examples but do not constitute an exhaustive list, are obligatorily expressed in certain languages while being absent in other languages. That is, the semantic distinctions expressed by these grammatical categories may be drawn in the other languages, too, but then by other means than grammatical markers. For instance, in English finite verb forms express either past or present tense,[43] while in certain East- and South-East Asian languages, e.g. Vietnamese, there is no tense-marking verbal morphology. When translating from Vietnamese into English, it is a problem to pick appropriate tense markers on finite verbs in the target text if the source text contains no explicitly expressed information to settle the choice. In practice, there will be contextual cues which a human translator will be able to interpret easily, but in automatic translation such information is normally not accessible. In such cases the challenge for MT lies in the fact that the amount of information that is linguistically expressed in the source sentence is insufficient for the generation of the target sentence.

Finally among linguistic considerations in MT development we want to mention *mapping problems*, i.e. problems related to the mapping between source and target language. This is a topic area where many researchers from, roughly, the 1980ies onwards, have tried out a multitude of sophisticated approaches for describing

---

[42] Insofar as automatic translation relies on successful word sense disambiguation, it is a harder problem to keep polysemous senses apart than to distinguish homographs with semantically unrelated meanings and which may even occur in separate domains. The reason is that there is a greater degree of overlap between the types of contexts that senses related through polysemy occur in than between those of homographs. Cf. Ide and Wilks (2006) on a discussion of what level of sense distinctions it is fruitful to aim at in natural language processing.

[43] Exceptions are imperative and subjunctive verb forms, which are marked with respect to the category of mood.

various kinds of linguistic phenomena that occur in the cross-linguistic setting. Interesting work has been done especially with reference to phenomena involving differences in predicate-argument structure between source and target text. Dorr et al. (1998: 8–9) discuss five different classes of such phenomena, among which we want to illustrate two types.

First, in the case of "thematic divergence" a verbal argument realised in one language as a syntactic subject corresponds translationally with an argument realised as a syntactic object in another language.[44] A simple illustration of the phenomenon is the English sentence *Writing pleases me* translated into Norwegian as *Jeg liker skriving* ('I like writing').

Second, there is the phenomenon referred to as "head-switching divergence", where lexical material realised as a main verb (i.e. a syntactic head) in one language corresponds translationally with lexical material realised as a subordinated verb in another language. A much used example of this is the correspondence between the German sentence *Peter schwimmt gern* ('Peter swims with-pleasure') and the English sentence *Peter likes to swim*.

In addition, Dorr et al. (1998: 9) mention structural, categorial, and conflational divergence as types of mapping problems. Structural divergence means that an argument has different syntactic realisations in source and target text, respectively. Categorial divergence covers cases where a given source word corresponds translationally with a target word of a different syntactic category, and in the case of conflational divergence a pair of translationally corresponding verbs differ with respect to the number of arguments that must be overtly expressed.[45]

The various kinds of mapping problems are easily solved by the human translator provided that he or she has sufficient knowledge about the relationship between source and target language. For the MT system developer the challenge is to identify and describe the divergence phenomena, and encode such descriptions in the linguistic components of the translation system. This can be implemented in a separate

---

[44] This has often been referred to as *argument switching*, which concerns divergences in the mapping of semantic arguments onto syntactic functions.

[45] The translational correspondence between the verb phrases *himlet* and *rolled her eyes* in example (4) in 1.3.1 is an example of conflational divergence.

component, a transfer module, which contains information about mapping relations between SL and TL.[46] In some cases source-target divergences of the types mentioned are associated with individual predicate-argument structures expressed by specific lexical items. If a certain type of divergence phenomenon pertains to several lexical items, then it is desirable to find a uniform description of the whole class of instances as this contributes to economy in the information modules. Moreover, an important question is whether the specific mapping relations apply whenever certain predicates are expressed in the source text. With respect to the English verb *please* (cf. above), it is not necessarily translated into the Norwegian verb *like*. The predicate expressed by *please* corresponds semantically with the predicate expressed by the Norwegian verb *behage*, and in that case there will be no head switching divergence as *please* and *behage* have isomorphic predicate-argument structures. *Behage* is, however, somewhat more archaic than the Norwegian verb *like*, and would not be an appropriate translation in any context. Then the problem for automatic translation is how to identify, in the source text, the conditions governing the choice between different possible mappings between SL and TL. To handle such challenges MT systems need to make correct choices between rather fine-grained sense distinctions. Citing Edmonds and Hirst (2002), Ide and Wilks (2006: 65) indicate that this can be achieved by integrating "additional knowledge and/or reasoning", which they regard as a task for computational lexicography and artificial intelligence, and not for word sense disambiguation.

From the perspective of theoretical linguistics, it is in itself an appealing task to account for such divergence phenomena through adequate grammatical descriptions, but in the context of machine translation, system developers will have to consider whether such efforts of grammar development are worthwhile. They are probably not if a given system is designed for a text type where the mapping problems are infrequent.

[46] Cf. the presentation of MT systems architectures in 1.4.2.4.

1.4.2.4 MT system architectures

In the presentation of machine translation we have several times referred to procedures and information modules, understood as components of MT systems. This section will briefly look at different types of MT system architectures, and we shall see that differences with respect to translation strategy are reflected by different ways of structuring the linguistic information encoded in an MT system. In this context the notion of 'translation strategy' covers the set of principles underlying the design of the translation procedure in an MT system, and it is commonly used for the purpose of classifying systems. There is a basic division between systems using *direct* strategies, and those using *indirect* strategies, and within the latter group a further distinction is made.

In direct MT systems translation is basically done by mapping the words in the input text directly onto words in the target language. The earliest systems, so-called *first generation systems*, used direct strategies, and, as already pointed out in 1.4.2.1, those systems could be seen as implementations of bilingual dictionaries with certain reordering rules for accommodating structural differences between SL and TL. Hence, in direct systems the encoding of linguistic information, as well as the implementation of translation procedures, were strongly dependent on the specific language pair, and the direction of translation, that each system was designed for. It has frequently been said that in direct systems the source text was analysed in terms of the target language, so that the target text could be generated directly from the result of the analysis.

In indirect MT systems translation is done by means of some sort of intermediate representation produced by a linguistic analysis of the input text. Such systems appeared as a response to the apparent failure of the direct technique, and are by some referred to as *second generation systems* (cf. Hutchins and Somers 1992: 71–72). Within indirect MT systems a distinction gradually evolved between the *transfer* strategy on the one hand and the *interlingua* strategy on the other.[47]

[47] Traditionally, the perhaps most common approach in MT system typologies has been the tripartite division into direct, transfer, and interlingua systems; cf. Hutchins and Somers (1992: 71–76), Dorr et al. (1998: 12–18).

Transfer-based MT systems are characterised by three separate stages in the translation process: analysis, transfer, and generation.[48] The first stage is a linguistic analysis of the input: by means of a grammar and lexicon describing the source language the system produces a representation of the meaning and structure of the source sentence. During the transfer stage this representation is changed so that it can eventually serve as the basis for target text generation. Necessary changes involve finding TL equivalents of the lexical items in the source text and transforming the input structure wherever it does not conform with the structural requirements of the TL grammar. Then, during the generation stage the information contained in the transformed representation of the input is used, together with information contained in the target language descriptions, to produce TL word forms and to arrange them according to correct TL word order.

The basic difference between interlingua systems and the transfer-based ones is that the transfer stage is dispensed with in interlingua systems. This can be done because the analysis stage "translates" the input text into an *interlingua expression* from which the target text may be generated. In the context of machine translation, an *interlingua* is a level of representation, in principle of a language-neutral kind, and in practice at least neutral between source and target language. The basic idea is that through linguistic analysis the information contained in the source text will be explicitly expressed in the format of an interlingua. Thus, the interlingua representation of the source text, together with target language descriptions, contains sufficient information for the system to produce an output sentence. In theory, an interlingual MT system does not need any bilingual information modules — not even a bilingual lexicon, provided that each monolingual lexicon is mapped onto the interlingua. Examples of interlinguas that have been used in MT systems are artificial logical languages, sets of (presumably) universal semantic primitives, and the artificial language Esperanto (cf. Hutchins 1986: 55). The PONS system (Dyvik

[48] Here we have omitted the initial stage of tokenisation, which involves reading the input text and identifying its word forms. This stage is, however, not peculiar to transfer systems, but necessary in any kind of automatic translation where the input text is syntactically parsed.

1990, 1995), when translating in mode 3 (cf. 1.3.2), uses situation schemata as an interlingua.[49]

The division between transfer and interlingua systems may be seen as a gradual one rather than as a discrete one. In a transfer system, the amount of work needed during the transfer stage depends on the depth of the linguistic analysis of the source text. If the analysis creates a sufficiently detailed, and sufficiently language-neutral, representation of the input, then it may contain enough information to serve as a basis for the generation of the output.

An important difference between direct and indirect MT systems is that in the latter type it is possible to keep linguistic information separate from the translation procedure, which makes it far easier to extend a system to new language pairs.[50] As pointed out in 1.4.2.3, it is a demanding task to build linguistic resources for MT systems, and it is an advantage if such information modules, once they have been compiled, may be reused. In this respect interlingua systems appear more attractive than transfer systems, since the interlingua strategy does not require any language-pair dependent components. Transfer systems, on the other hand, need bilingual lexicons as well as sets of transfer rules, and the latter may be not only language-pair specific, but also dependent on the direction of translation.

On the other hand, the interlingua strategy is not necessarily the most attractive approach to automatic translation, given the degree of complexity in the translation task. Interlingual translation requires a deep analysis of the input text, and this is computationally demanding. But actual translation does not always require great efforts. If there is a sufficient degree of structural similarity between source and target language, then it is sometimes possible to translate word-by-word, or almost word-by-word. Thus, there are cases where the direct translation strategy would be sufficient; those are included among what we have described as type 1 correspondences.[51] With respect to type 2 correspondences, the transfer strategy seems appro-

---

[49] The PONS situation schemata are based on Situation Semantics; cf. Barwise and Perry (1983), Fenstad et al. (1987).

[50] Cf. the remarks on modularity in 1.4.2.1.

[51] Cf. the brief introduction to the type hierarchy in 1.3.1. Quantitative results concerning the distribution of the four correspondence types within the analysed data are presented in chapter 5.

priate: at the transfer stage the structure of the source text is changed according to the TL grammar. The types of source-target divergences found in type 2 correspondences pertain to surface syntactic structure, which means that translation can be done by transfer at a "shallow" linguistic level. Moreover, as direct systems have been able to accommodate certain word order differences between SL and TL, it is possible that also type 2 correspondences could be handled by the direct strategy. Then, in cases where the translation task is more complex than in correspondences of types 1 and 2, transfer must take place at a deeper level, and it may be necessary to do a full semantic analysis of the source text in order to reveal sufficient information for target text generation. The experimental PONS system combines, in a sense, all three translation strategies — direct, transfer, and interlingua. The system demonstrates that deep analysis and interlingual translation is necessary only in certain cases, and that an interesting challenge is to find those instances of translation where either the direct strategy or shallow transfer is sufficient to produce an appropriate translation.

1.4.2.5 Linguistic vs. non-linguistic approaches

As mentioned in 1.4.2.1, a division emerged in the early 1990ies between linguistics based and non-linguistics based approaches to machine translation. This division applies to a dimension independent of that of translation strategy; it pertains to what kinds of information resources an MT system is equipped with, and in what ways those resources are designed.

Ever from the early days of machine translation and until about 1990 there was a general view that to achieve automatic translation it was necessary to use linguistic information, i.e. information about source and target language and about how SL and TL are interrelated. Such information sources can be seen as a parallel to the bilingual competence of a human translator (cf. 1.2 and 2.4.2). Until about 1990 the established view among MT researchers was not only that MT systems needed linguistic information, but also that such information should be given in language descriptions designed according to principles of linguistic theory. A great variety of approaches of this kind have been investigated, and they are presented as *linguistic-based research paradigms* by Dorr et al. (1998: 19–30).

It indeed caught some attention when researchers had implemented methods for automatic translation that did not use linguistic information. From about 1990 onwards several techniques of this kind appeared; they are presented as *non-linguistic-based paradigms* by Dorr et al. (1998: 30–35), and they cover what is referred to as *statistical MT* in 1.4.2. Non-linguistic translation systems have in common that they depend, either for their development or for their functioning, on the existence of large parallel corpora. That is, non-linguistic MT techniques use large parallel corpora as repositories of information about the translational relation between two languages. Another important prerequisite for the workability of these approaches is the development of efficient algorithms for the automatic alignment of words or word sequences.[52] Word alignment applies to translationally parallel texts of two different languages, and it involves identifying links between translationally corresponding word forms in the two texts. By using the information contained in such links it is possible to find recurring translational correspondences. To put it simply, non-linguistic MT systems compute translations on the basis of which translational patterns that are frequent in the parallel corpus used by the system. The key to identifying a target equivalent *b* for a given source expression *a* is the probability that *a* corresponds with *b* based on the actual correspondences in the parallel corpus.

An important reason why non-linguistic approaches have been developed is that even though linguistic methods have reached a high level of sophistication, there are large development costs involved when building linguistic-based MT systems, and it is not easy to combine computational efficiency and broad coverage in grammars and lexicons. On this background it is appealing to investigate what may be achieved by doing without linguistic information modules and by applying pure computer science to parallel corpora. Clearly, there are certain linguistic phenomena that are too complex to be handled by non-linguistic techniques (e.g. long-distance dependencies; cf. Dorr et al. 1998: 35), and now the trend is to integrate the two approaches in so-called hybrid MT design, so that the strengths of both techniques may be combined.

---

[52] An important contribution in this respect is Gale and Church (1993).

The question of automatisation which is implicit in the present study of translational complexity is not neutral in relation to the division between linguistic and non-linguistic approaches to MT. Our investigation relies on several assumptions regarding the types of information needed to produce a translation, and these assumptions have consequences for where and how we draw the limit of computability.[53] Although we have previously indicated that the results of our product-oriented study are in principle also relatable to statistical MT, it is the linguistic-based approaches that we see as relevant to our discussion of comput-ability.

1.4.2.6 The scope of machine translation

After a history of more than 50 years there seems to be general agreement that MT will not replace human translation. It seems unrealistic that automatic systems will reach a level of perfection where they produce high quality translations of unre-stricted text without any kind of human intervention. Thus, we cannot expect that post-editing of machine translation output will be dispensed with. On the other hand, MT systems have been applied for decades as translation tools, and this is because they have been useful, within their limitations. For years now it has been common to talk about *the translation industry*, and that expression indicates, firstly, how large the demand for translation is, in particular of the non-literary kind, and, secondly, that automatised tools are needed in order to meet that demand.

Thus, practice shows that, given certain conditions, computerised translation can be a very helpful tool for reducing the workload for human translators. For one thing, if there is a high degree of structural relatedness between source and target language, then the challenges involved in MT design are reduced. Moreover, researchers and developers have experienced that successful systems can be designed for so-called sublanguage texts. Examples could be maintenance manuals and similar kinds of technical documents, which are characterised by relatively precise and unambiguous language, often repetitive, and dominated by a limited set of syntactic constructions.

[53] This will be discussed in chapter 2.

Such texts are not attractive to human translators, and the task of translating them rather resembles what computers are particularly good at: to repeat tedious computations, and to do so with precision.

Although fully automatic high quality translation probably will remain an unattainable ideal, it is still the notion of fully automatic translation which is of relevance to the present project: when discussing to what extent it would be possible to simulate human translation as instantiated by the investigated parallel texts, we assume that the translation task is to be solved without any human intervention. This must be seen as a framework for posing research questions, and not as a norm for practical systems.

1.4.3 Parallel corpus linguistics

As our investigation of translational complexity applies to parallel corpus data, it is appropriate to pay some attention to the field of parallel corpus linguistics. And, as mentioned in 1.4.2.5, the availability and use of parallel corpora has also become highly important to machine translation research. The label *parallel corpus linguistics* is taken from Borin (2002), who identifies the field as a subpart of the larger domain of *corpus linguistics*.

1.4.3.1 Corpus linguistics

This field is defined as follows by McEnery and Wilson (2001: 2): "Corpus linguistics is not a branch of linguistics in the same sense as syntax, semantics, sociolinguistics, and so on. … Corpus linguistics in contrast is a methodology rather than an aspect of language requiring explanation or description."[54]

In recent years this methodology has come to be regarded as an indispensible part of linguistic research, and, basically, it involves providing empirical resources in the shape of machine-readable and searchable corpora, together with systematic methods for using the corpora in order to investigate specific linguistic phenomena. Clearly, it

[54] For an overview of the field see, in addition to McEnery and Wilson (2001), Sampson and McCarthy (2004), McEnery et al. (2006), Renouf and Kehoe (2006, 2009).

is impossible to do linguistic research without testing theories against examples of actual language use. Earlier there used to be some antagonism between linguists who advocated corpus-based studies and those who claimed that corpus data would always be incomplete and were inferior to what might be gained from studying the intuitions of individual language users.[55] Over the years, large corpus resources have become available for many languages, and computational linguists have developed efficient tools for identifying and processing linguistic data in large corpora. Thus, there is now a general trend that investigations of linguistic phenomena are carried out, preferably, with the use of corpus data, since corpora are important repositories of information about language use. There is, however, always the possibility that even in a large corpus a certain linguistic phenomenon might have no manifestations; in such cases the problem is to interpret the absence of occurrences: it is accidental or a consequence of aspects of the language system? Still, such cases do not reduce the value of the data that are found.

The Latin word *corpus* means 'body', and as stated by McEnery and Wilson (2001: 29), any body of text is in principle a corpus. However, "… the notion of a **corpus** as the basis for a form of empirical linguistics differs in several fundamental ways from the examination of particular texts" (2001: 29). More specifically, the building of corpora as used in modern corpus linguistics is normally subject to certain demands, of which McEnery and Wilson (2001: 29–32) discuss four kinds. Firstly, a corpus for linguistic research should be representative in the sense that it must, as far as possible, cover a whole variety of a language. Hence, it will be unsatisfactory to include texts of for instance only one type, or texts produced by only one author, or by authors of only one sex. Secondly, a corpus is normally of finite size: once it has been compiled according to a certain plan, new texts are not added.[56] An example of a fairly large, finite corpus is the British National Corpus with about 100 000 000 running words. Thirdly, it has now become a standard requirement in corpus building that such resources are machine-readable. Otherwise, computerised research tools

[55] For a discussion of this, see chapter 1 in McEnery and Wilson (2001).

[56] There are some exceptions, in particular corpora where new texts are continually added in order to keep the corpus up-to-date on current language use; cf. McEnery and Wilson (2001: 30–31).

cannot be used. Fourthly, once a representative, finite corpus has been compiled and made available for a research community, it is in a sense unavoidable that it will be attributed the status of a standard reference. Because such resources are valuable repositories of linguistic data, and may be kept constant, they are excellent test beds for varying approaches to the description of linguistic phenomena. On the background of these four requirements, or characteristics, McEnery and Wilson (2001: 32) present a prototypical definition of a corpus in modern linguistics: "… a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration."

There is an important division between *annotated* and *unannotated* corpora. Unannotated corpora contain "raw" text, i.e. plain text with nothing added, whereas in annotated corpora labels signifying various types of linguistic information have been attached to specific word forms. Examples of such information types are parts of speech and syntactic functions. Corpus annotation may be done manually or by software. The field of natural language processing now offers a range of different applications for automatic linguistic analysis, among which corpus annotation programs are an important subclass. As pointed out by McEnery and Wilson (2001: 32), a significant difference between annotated and unannotated corpora is that in the case of the former the added labels make explicit linguistic information that is only implicit in unannotated text, and hence annotation increases the value of a corpus. However, it may also add some "noise": if the annotator, whether a human or a computer program, makes any wrong analyses, then errors are included in the corpus.

The present investigation is carried out using data taken from parallel texts, and as will be described in chapter 4, the result of our analysis is a manually annotated corpus of translationally corresponding strings extracted from running texts. Still, our empirical analysis has not been done with reference to corpora in the sense given above, and hence we shall not go deeply into the field of corpus linguistics.

1.4.3.2 The added value of parallel corpora

For language researchers working under a cross-linguistic perspective parallel corpora are an invaluable resource. Borin (2002: 1) applies the label of *parallel*

corpus linguistics to research on parallel corpora, and he states that the "prototypical kind" of parallel corpora "is that which consists of original texts in one language, together with their translations into another language" (2002: 1). This is in contrast to the phenomenon of *comparable corpora*, which are collections of original texts in different languages, but of the same, or similar, text type, so that the texts are functionally comparable (cf. Borin 2002: 3). Comparable corpora fall outside the focus of our interest, since they do not contain translational correspondences of the same kind as parallel corpora do, but they are clearly of great value to contrastive linguistic studies. Johansson (2007: 9) makes the point that the term *parallel corpora* has unfortunately been used to cover comparable corpora as well as parallel corpora in the prototypical sense given by Borin. To solve this problem Johansson refers to parallel corpora as *translation corpora* in order to keep them distinct from comparable corpora, and he adds the multilingual dimension by defining translation corpora as containing "original texts and their translations into one or more other languages" (2007: 9).

In the previous section we discussed the usefulness of corpora for linguistic research, and it is not difficult to see what is the added value of parallel corpora. A representative parallel corpus may of course provide empirical data for monolingual studies,[57] but primarily it serves as a repository of information about the translational relation between the source and target language texts included in it. We have already seen that large parallel corpora have been used to develop MT systems operating without linguistic information modules (cf. 1.4.2.5), and the great utility of parallel corpora in research on translation, manual as well as automatic, is obvious. In addition to (machine) translation research, Borin (2002: 1) mentions other examples of areas where parallel corpora have been put to use: translation training, language teaching, bilingual lexicography, and contrastive and typological linguistics. For the latter kind of studies, multilingual parallel corpora are especially useful.

[57] That is, preferably with reference to the original texts. It is generally agreed that target texts normally exhibit certain linguistic properties specific to translations.

With respect to the present project, it could not have been carried out without access to parallel texts.[58] Approximately one half of the empirical data are collected from texts included in the English-Norwegian Parallel Corpus (ENPC), documented in Johansson (1998, 2007), and Johansson et al. (1999/2002). The ENPC is described by Johansson (2007: 11) as "a bidirectional translation corpus consisting of original English texts and their translations into Norwegian, and Norwegian original texts and their translations into English." It includes fiction as well as general, non-fiction texts and has a total of approximately 2,6 million words (cf. Johansson 2007: 13). An important feature of the ENPC is that it is *sentence aligned*, which means that each sentence in the corpus is linked to a translationally corresponding sentence (if found) in the parallel text (cf. Johansson 2007: 14–16). Thus, the ENPC is also an example of an annotated corpus, and it provides a goldmine of empirical data for contrastive linguistic research.

A strong field of modern contrastive language studies has evolved along with the development of corpus-based methods for linguistic research. The value of contrastive studies is obvious: they provide information about systematic differences between specific language systems, and about the effects of those differences as manifested in parallel corpus data. Both kinds of information are highly useful in many other fields, such as translation, language teaching, and translator training.[59] We may quote Johansson (2007: 1) on the great value of modern text corpora, and in particular of multilingual corpora, as repositories of representative data about language use: by exploring such resources "[w]e can see how languages differ, what they share and — perhaps eventually — what characterises language in general."

1.5 Organisation

This thesis consists of five main parts, among which the present chapter constitutes the first one. The purpose of this chapter has been to state our research questions, to

[58] Cf. the list of primary sources.

[59] Describing the large field of contrastive linguistic research falls outside of the scope of this work. Concerning the language pair English-Norwegian, Johansson (2007) is a representative study within the field: it presents corpus-based contrastive investigations of a range of linguistic phenomena, and also provides a multilingual perspective by including German and Swedish in some analyses.

introduce our framework, and to present some important topics of disciplines which are relevant to this study.

Part II includes chapters 2 and 3, and covers the theoretical and analytical foundations of our investigation. In chapter 2 we argue for a product-oriented approach to the study of translation, before explaining principles for drawing the limit of computability, or linguistic predictability, in the translational relation. Then, the basic notions of information, knowledge, and informational content are discussed, and we present our typology of information sources for translation. Chapter 3 opens with an informal presentation of the information-theoretic concepts of computability, complexity, and related notions. Then we present some approaches to the description of linguistic complexity, and describe our own notion of translational complexity, as well as its relation to computability. The remainder of chapter 3 is a detailed description of the correspondence type hierarchy. The four types are presented as translation tasks in order to capture the information requirements of each type, and to relate the notion of translational complexity to the amount and types of information needed for solving a translation task, including necessary processing effort.

Part III contains chapter 4, which describes our empirical investigation. The chapter starts by presenting the analysed parallel texts, as well as the concerns lying behind the selection of texts. Further, the syntactic criteria for identifying units of analysis are presented and illustrated, before we discuss the principles governing the classification of extracted string pairs in terms of translational complexity. Also, chapter 4 describes several practical aspects of the recording of translational correspondences.

Part IV covers chapters 5 and 6, which present the results of our analysis, and discuss them in relation to the initial research questions. Chapter 5 focuses on the analysed pairs of texts, and we present the complexity measurements across all recorded data, as well for each direction of translation, for each text type, and for the individual text pairs. Text-typological differences revealed by the analysis constitute a central topic in the discussion of the results. Chapter 6 presents certain phenomena which are recurrent among the recorded data, and which involve some kind of semantic deviation between translationally corresponding units. These are sorted into

a set of subtypes within the main correspondence types. The discussion of the semantic subtypes shows how the line is drawn between, respectively, computable and non-computable translation, and it illustrates certain phenomena that are not included in the domain of linguistically predictable correspondences of the language pair English-Norwegian.

Part V consists of chapter 7, where certain conclusions are drawn. These are centred around three topics: our framework, the method, and the results of the study. Also, we indicate a possible extension of our analytical approach.