THE UNIVERSITY *of* EDINBURGH

Incorporating Pronoun Function into Statistical Machine Translation

Liane Guillou

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2016

Abstract

Pronouns are used frequently in language, and perform a range of functions. Some pronouns are used to express coreference, and others are not. Languages and genres differ in how and when they use pronouns and this poses a problem for Statistical Machine Translation (SMT) systems (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Novák, 2011; Guillou, 2012; Weiner, 2014; Hardmeier, 2014). Attention to date has focussed on coreferential (anaphoric) pronouns with NP antecedents, which when translated from English into a language with grammatical gender, must agree with the translation of the head of the antecedent. Despite growing attention to this problem, little progress has been made, and little attention has been given to other pronouns.

The central claim of this thesis is that pronouns performing different functions in text should be handled differently by SMT systems and when evaluating pronoun translation. This motivates the introduction of a new framework to categorise pronouns according to their function: Anaphoric/cataphoric reference, event reference, extra-textual reference, pleonastic, addressee reference, speaker reference, generic reference, or *other* function. Labelling pronouns according to their function also helps to resolve instances of *functional ambiguity* arising from the same pronoun in the source language having multiple functions, each with different translation requirements in the target language. The categorisation framework is used in corpus annotation, corpus analysis, SMT system development and evaluation.

I have directed the annotation and conducted analyses of a parallel corpus of English-German texts called **ParCor** (Guillou et al., 2014), in which pronouns are manually annotated according to their function. This provides a first step toward understanding the problems that SMT systems face when translating pronouns. In the thesis, I show how analysis of manual translation can prove useful in identifying and understanding systematic differences in pronoun use between two languages and can help inform the design of SMT systems. In particular, the analysis revealed that the German translations in ParCor contain more anaphoric and pleonastic pronouns than their English originals, reflecting differences in pronoun use. This raises a particular problem for the evaluation of pronoun translation. Automatic evaluation methods that rely on reference translations to assess pronoun translation, will not be able to provide an adequate evaluation when the

reference translation departs from the original source-language text. I also show how analysis of the output of state-of-the-art SMT systems can reveal how well current systems perform in translating different types of pronouns and indicate where future efforts would be best directed. The analysis revealed that biases in the training data, for example arising from the use of "it" and "es" as both anaphoric and pleonastic pronouns in both English and German, is a problem that SMT systems must overcome. SMT systems also need to disambiguate the function of those pronouns with ambiguous surface forms so that each pronoun may be translated in an appropriate way.

To demonstrate the value of this work, I have developed an automated post-editing system in which automated tools are used to construct ParCor-style annotations over the source-language pronouns. The annotations are then used to resolve functional ambiguity for the pronoun "it" with separate rules applied to the output of a baseline SMT system for anaphoric vs. non-anaphoric instances. The system was submitted to the DiscoMT 2015 shared task on pronoun translation for English-French. As with all other participating systems, the automatic post-editing system failed to beat a simple phrase-based baseline. A detailed analysis, including an oracle experiment in which manual annotation replaces the automated tools, was conducted to discover the causes of poor system performance. The analysis revealed that the design of the rules and their strict application to the SMT output are the biggest factors in the failure of the system.

The lack of automatic evaluation metrics for pronoun translation is a limiting factor in SMT system development. To alleviate this problem, Christian Hardmeier and I have developed a testing regimen called **PROTEST** comprising (1) a hand-selected set of pronoun tokens categorised according to the different problems that SMT systems face and (2) an automated evaluation script. Pronoun translations can then be automatically compared against a reference translation, with mismatches referred for manual evaluation. The automatic evaluation was applied to the output of systems submitted to the DiscoMT 2015 shared task on pronoun translation. This again highlighted the weakness of the post-editing system, which performs poorly due to its focus on producing *gendered* pronoun translations, and its inability to distinguish between pleonastic and event reference pronouns.

Lay Summary

Pronouns are used frequently in language, to serve different functions. These differ somewhat from language to language, and this poses a problem for Statistical Machine Translation (SMT) systems. Focussing on cases where a pronoun refers to an entity introduced earlier in the text with a noun phrase (its *antecedent*) – where the pronoun serves an anaphoric function – when translating from English into a language with grammatical gender such as French or German, the anaphoric pronoun must agree with the translation of its antecedent. Despite growing attention to this problem, little progress has been made, and little attention has been given to pronouns serving other functions.

The central claim of this thesis is that pronouns performing different functions in text should be handled differently by SMT systems and when evaluating pronoun translation. This motivates the introduction of a new framework to categorise pronouns according to their function. Labelling pronouns according to their function also helps to resolve instances of *functional ambiguity* arising from the same pronoun in the source language having multiple functions, each with different translation requirements in the target language. The categorisation framework is used in corpus annotation, corpus analysis, SMT system development and evaluation.

I have directed the annotation and conducted analyses of a collection of parallel English-German texts called **ParCor**, in which pronouns are manually annotated according to their function. This provides a first step toward understanding the problems that SMT systems face when translating pronouns. An analysis of manual translation revealed differences in the use of anaphoric and pleonastic ("dummy") pronouns between English and German. An analysis of the output of state-of-the-art SMT systems highlighted (1) the need to disambiguate the function of those pronouns with ambiguous surface forms so that each pronoun may be translated in an appropriate way and (2) the need to further sub-categorise anaphoric pronouns according to the translation requirements of the target language.

To demonstrate the value of this work, I have developed an automated post-editing system which aims to identify pronouns that have been translated incorrectly and correct them, using rules that differ for anaphoric and for non-anaphoric pronouns. The system was submitted to the DiscoMT 2015 shared

task on pronoun translation for English-French. As with all other participating systems, my system failed to beat a simple baseline. A detailed analysis revealed that the design of the rules and their strict application to the SMT output are the biggest factors in the failure of the system.

The lack of automatic evaluation metrics for pronoun translation is a limiting factor in SMT system development. To alleviate this problem, Christian Hardmeier and I have developed a testing regimen called **PROTEST** comprising (1) a hand-selected set of pronoun tokens categorised according to the different problems that SMT systems face and (2) an automated evaluation script. Pronoun translations can then be automatically compared against a human-authored reference translation, with mismatches referred for manual evaluation. The automatic evaluation was applied to the output of systems submitted to the DiscoMT 2015 shared task on pronoun translation. This again highlighted the weakness of the post-editing system, particularly with respect to the translation of non-anaphoric pronouns.

Acknowledgements

I have received valuable advice and support from many people during my PhD studies. First and foremost I wish to thank my supervisor, Professor Bonnie Webber, without whom this thesis would not have been possible. Really. Thank you for everything that you have done for me, from patiently explaining linguistic concepts to helping me see the "big picture" to proofreading the many drafts of papers, presentations and of course, this thesis. With your enthusiasm and endless encouragement, you have undoubtedly been the best teacher I could have ever asked for. Thanks are also owed to Professor Philipp Koehn, my second supervisor, whose classes I looked forward to with great excitement as a Masters student. It was an honour and a privilege to learn about Statistical Machine Translation for the person who (quite literally) wrote the book on it. Thank you for the support and opportunities that you have given me.

I would also like to thank my examiners, Andrei Popescu-Belis and Adam Lopez for their insightful comments and recommendations, and for an enjoyable viva.

I have benefitted enormously from the knowledge and support of both past and present members of the StatMT Group at the University of Edinburgh: Lexi Birch, Nikolay Bogoychev, Christian Buck, Federico Fancellu, Ulrich Germann, Barry Haddow, Eva Hasler, Kenneth Heafield, Hieu Hoang, Matthias Huck, Miles Osborne, David Matthews, Maria Nadejde, Hervé Saint-Amand, Nathan Schneider, Rico Sennrich, Dominikus Wetzel and Philip Williams. Thank you for asking the difficult questions at my annual reviews and practice presentations, for the many useful suggestions, and for your company.

During the course of my studies I have engaged in many interesting conversations about my work, but the following people stand out. Christian Hardmeier has proven to be a great ally and collaborator, without whom my path would have seemed a lot darker. Ekaterina Lapshinova-Koltunski and Marine Carpuat have taken a great interest in discourse in SMT and have provided numerous valuable insights and suggestions.

Susanne Tauber has provided the much-needed German language expertise that I have relied on for this work, performing a dual role as both my German tutor and primary German annotator for the ParCor corpus and manual evaluation tasks. Thanks also to Sam Gibbon and Aaron Smith, the primary English

annotators for the ParCor corpus, and to Petra Strom and Daniel Lawrence for providing the secondary annotations. Thanks to Felix Suessenbach and (again to) Sam Gibbon, my colleagues in Psychology, for their patient explanations of statistical methods and for assisting in the significance testing over the ParCor corpus.

Special thanks are owed to those people who have undertaken this journey with me. To my friends at Edinburgh Aikido club, thank you for many enjoyable hours on the mats, and for the welcome distraction from my work. To Okojo-san, who never failed to brighten my day. To Andreea Radulescu and Clare Llewellyn, thank you for being the voices of reason and for keeping me on track through the tougher times. To Wolodja Wentland, thank you for all of the help and support that you have given me; from providing German language assistance and late-night "tech support" to your unfaltering emotional support upon which I have heavily relied. Lastly, thank you to Anna Guillou, my mother, teacher and friend, who taught me the value of education and has supported me in everything that I have chosen to do.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Liane Guillou*)

For Anna Guillou

Contents

Chapter 1

Introduction

1.1 Pronouns in SMT

Pronouns and noun phrases (NPs) belong to the wider class of *referring expressions* which are used to identify entities and events in discourse. Pronouns are used frequently in language, and perform a range of functions. Some pronouns are used to establish a *coreference* link, where two or more expressions refer to the same thing, while others are not.

Pronominal coreference is a form of *reduced coreference* in which pronouns are used in place of full referring expressions such as noun phrases, NPs, verbs, verb phrases, or entire clauses or sentences in a text. Languages differ in how and when they use pronominal coreference and this continues to challenge Statistical Machine Translation (SMT) systems (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Novák, 2011; Guillou, 2012; Weiner, 2014; Hardmeier, 2014).

Non-coreferential uses of pronouns include pleonastic pronouns, comparative anaphors ("other", "another"), and instances where the type of the pronoun and the NP on which it anaphorically depends are different. Consider Ex. 1.1 from Webber (1988) in which "They" refers to dachshunds in general via anaphoric reference to the specific dachshund down the block, but in a context in which the referent must be generic.

(1.1) The dachshund down the block bit me yesterday. **They**'re really vicious beasts.

1.1.1 The Anaphoric Problem

One of the most well researched problems concerning the translation of pronominal coreference is the translation of *anaphoric* pronouns into languages with grammatical gender. *Anaphora* is a form of pronominal coreference in which a pronoun is used to refer to an entity previously introduced in the discourse (its *antecedent*). For example, in Ex. 1.2, the *anaphoric* pronoun "it" co-refers with "umbrella":

(1.2) I have an **umbrella**. **It** is red.

(1.3) Ich habe einen **Regenschirm**. **Er** ist rot. ✓

(1.4) Ich habe einen **Regenschirm**. **Sie** ist rot.

(1.5) Ich habe einen **Regenschirm**. **Es** ist rot.

An SMT system will typically translate each sentence of the English source text in isolation. That is, the second sentence of Ex. 1.2 will be translated without knowledge of the first sentence, and in particular without the knowledge of the pronoun's antecedent. When translating into German an SMT system may therefore choose to translate "it" as "er" (masculine, Ex. 1.3), "sie" (feminine, Ex. 1.4) or "es" (neuter, Ex. 1.5), all of which are third-person singular pronouns corresponding to the English "it". Given that "umbrella" is translated as "Regenschirm" (masculine, singular), "it" should be translated as "er" (masculine, singular) to ensure that the pronoun and antecedent agree in terms of number and grammatical gender. This pronoun-antecedent agreement is a requirement in German, French and many other languages, although the agreement features may vary between languages.

The problem of translating anaphoric pronouns is not limited to the *inter-sentential* scenario of the above example, in which the pronoun appears in a different sentence from its antecedent. SMT systems also struggle to accurately translate *intra-sentential* pronouns, which appear in the same sentence as their antecedent. If both pronoun and antecedent appear in the same translation unit, or some other local context is sufficient to disambiguate the pronoun, a correct translation may result. Or the pronoun may simply be translated correctly by chance. However, these conditions cannot be guaranteed and in many cases the pronoun will be translated incorrectly.

1.1.2 Functional Ambiguity

To further compound the problem of pronoun translation, some pronouns exhibit *functional ambiguity* – that is, a single pronoun may perform multiple functions. For example the English pronoun "it" may be used as an anaphoric, pleonastic or event reference pronoun. These different pronoun *functional types* have different translation requirements in German and other languages. A *pleonastic*, or *dummy* pronoun is syntactically required but does not refer to anything and as such, there are no agreement constraints. For example, the "It" in Ex. 1.6 does not refer to anything in the previous sentence and its correct translation into German requires the pleonastic pronoun "Es".

(1.6) I have an umbrella. **It** is raining.

(1.7) Ich habe einen Regenschirm. **Es** regnet.

As with English "it", "es" may serve as both an anaphoric pronoun (neuter, singular) or as a pleonastic pronoun. If pleonastic pronouns are common in both the source and target-language texts, alignments of "it" and "es" will be common, more so than alignments between "it" and "er" (masculine) or "sie" (feminine). As SMT systems are built by extracting co-occurrence frequencies from parallel texts, such frequent alignments can introduce a bias toward translating anaphoric "it" as "es" in English-German SMT.

Event reference pronouns may refer to verbs, verb phrases, or entire clauses or sentences in a text. For example, the "It" in Ex. 1.8 refers to the invasion of Y by X. Again there are no agreement constraints for the translation of "it". The correct translation of "it" in German requires the event reference pronoun "dies":

(1.8) X invaded Y. **It** resulted in war.

(1.9) X besetzte Y. **Dies** führte zu Krieg.

The translation of "it" into German therefore requires disambiguation. That is, in order to provide a correct translation in German, we must first know whether an instance of "it" is anaphoric, pleonastic or event reference. For anaphoric "it" we also need to know what the pronoun's antecedent is. Note that in some unambiguous cases it may not be necessary to know the pronoun's antecedent in order to produce a correct translation. For example there are fewer *translation*

options in German for the English pronoun "she" ("sie") than for anaphoric "it" ("er", "sie' and "es"). In the case of event reference pronouns, the same pronoun may be used to refer to different events, independent of the properties of the event that is being referred to. This is true of both English and German.

Similar problems of ambiguity exist for other pronouns. For example, "they" may be used as an anaphoric pronoun, or as a *generic* pronoun as in Ex. 1.10. In German the generic pronoun is "man" (Ex. 1.11).

(1.10) **They** say it always rains in Scotland.

(1.11) **Man** sagt, dass es in Schottland immer regnet.
(Lit: **They** say, that it always rains in Scotland)

Another example is the second-person personal pronoun "you" which can be used as a deictic or a generic pronoun in English. When used as a *deictic* pronoun, "you" refers to a specific person or group of people, and should be translated in German as "Sie/Ihnen". When used as a *generic* pronoun, as in Ex. 1.12 where "you" doesn't refer to a specific person or group but rather to people in general, the correct translation in German is "man" (Ex. 1.13).

(1.12) In England, if **you** own a house **you** have to pay taxes.

(1.13) In England muss **man**, wenn **man** ein Haus besitzt, Steuern bezahlen.
(Lit: In England **you** must, if **you** own a house, pay taxes)

Again, to ensure correct translation of these ambiguous source-language pronouns, especially those for which many *translation options* exist, it may be necessary to first disambiguate their function. This need for source-language disambiguation is not restricted to the English-German pair. The effects of functional ambiguity should be considered for all language pairs. For example, if the source language is German, it would be wise to disambiguate feminine singular vs. plural uses of "sie" and if the source language is French we need to distinguish between anaphoric (masculine, singular) and pleonastic uses of "il".

1.2 Variation of Pronoun Use

In the examples in Section 1.1.1 variation in the reference translation arises from choosing different translations of the English antecedent head and selecting a

pronoun with the appropriate gender. For other pronoun functions variation arises from the use of different pronouns which capture the same meaning and may therefore be used interchangeably. Consider the following examples:

(1.14) [**You/One**] should always tell the truth.

(1.15) I got the hiccups when I drank Champagne. [**This/It**] happened again when I drank sparkling cider.

In Ex. 1.14, the generic pronouns "You" and "One", may be used interchangeably without altering the meaning of the text. So too in Ex. 1.15, the event reference pronouns "This" and "It" could both be used to provide the same meaning.

1.3 Insertion and Deletion of Pronouns

Recent work has focussed on the problem of translating a pronoun in the source-language text as a pronoun in the target language. This approach ignores the insertion and deletion of pronouns in translation, and where and when these actions are appropriate. Pronouns that were not present in the source-language text may be added to the translation (*insertion*), and pronouns that were present in the source-language text may not appear in the translation (*deletion*). Consider the following example in which the possessive pronoun "its" is added to the English translation of the German sentence in Ex. 1.16 (Becher, 2011):

(1.16) Deshalb bleibt XyzTech mit positivem Cash Flow und gutem Ergebnis im Konzern.

(1.17) As a result, we shall retain XyzTech, with **its** positive cash flow and good earnings.

Here a possessive pronoun is required in the English translation, but it would not be natural to include one in the original German sentence.

In general, a pronoun may be inserted/deleted for syntactic reasons (i.e. due to requirements of the target language), or stylistic reasons. Different styles may be used for different reasons. Translators may have a personal style which is appropriate in some scenarios, whereas other scenarios might require adherence to a particular *house style*, or a certain degree of formality.

In the case of pro-drop languages, such as Spanish or Czech, subject position pronouns may be omitted if they can be inferred from text via other means. English, German, and French — the languages central to the work in this thesis — are not pro-drop languages. Research on pro-drop in SMT is left for future work.

In contrast to manual translation, an SMT system will typically translate pronouns in the source-language text as pronouns in the target language, but there is still scope for insertion or deletion. In the case of SMT, when a pronoun is inserted/deleted, we could ask: Is it acceptable that the pronoun has been inserted/deleted? This thesis explores some of the reasons why pronouns may be inserted in the target language, but leaves open the wider questions of what actions are desirable and how best to model this in SMT. These issues are left for future work.

1.4 Pronoun Forms and Functional Types

In linguistics, pronouns are typically categorised according to their *form*. The following list of categories is commonly found in both English and German grammar books[1], and indeed it can be applied to both French and German (the other languages studied in detail in this thesis).

- Personal: Classified by person, number and case. English has first, second and third-person pronouns and divides them into singular and plural. The personal pronouns in English are: I/me, we/us, you, he/him, she/her, it and they/them
- Reflexive: Used when a person or thing acts on itself. For example, "John talks to **himself**". Reflexive pronouns co-refer with an NP within the same clause
- Reciprocal: Used when there is a reciprocal relationship between two people or things. For example "The boys don't like **each other**". Reciprocal pronouns refer to an NP within the same clause

[1]Dippmann (1987) and Engel (1988) provide clear explanations of the categorisation of German pronouns by form (personal, reflexive etc.). Huddleston (1988) provides a similar categorisation of English pronouns.

- Possessive: Used to indicate possession. The "possessives" group comprises both possessive pronouns and possessive adjectives. The possessive pronouns in English are: mine, yours, his, hers, its, ours and theirs. The possessive adjectives are: my, your, his, her, its, our, their

- Demonstrative: The "demonstratives" group comprises both demonstrative adjectives, which qualify nouns (e.g. "**This** apple tastes good") and possessive pronouns which may be substituted for nouns (e.g. "**This** tastes good"). Both may use proximity to / distance from the speaker to distinguish one entity from others. For example "Is **this** book yours? / Is **that** book yours?" (possessive determiners) or "Is **this** yours? / Is **that** yours?" (demonstrative pronouns). Demonstrative pronouns may also be used in discourse, without spatial dimension, to refer to something that is currently being said or was previously said. For example "I told my friend he was good at sports. He liked **that**". The English demonstrative pronouns are: this, that, these and those

- Indefinite: Used to refer to one or more unspecified persons or things. For example "**Everyone** likes cats". The set of indefinite pronouns in English includes combinations of: some/any/every/no + thing/one/body

- Relative: Used in relative clauses to refer to people or things previously mentioned in an earlier clause in the sentence. The English relative pronouns include: who, whom, whose, what, which and that

- Interrogative: Used to ask questions about a person or thing. For example "**Who** said that?". The English interrogative pronouns include: who, whom, whose, which, what, etc.

The above categorisation of pronouns mirrors that laid out by the *Interlingua Grammar* (Gode and Blair, 1951) which describes features that are common to a collection of source languages: English, Spanish, Portuguese, Italian, French, German and Russian. The grammar describes a common set of pronoun categories: Personal, reflexive, possessive, demonstrative, indefinite and relative. Reciprocal and Interrogative pronouns are not explicitly mentioned in the grammar definitions. However, pronouns from these categories do exist in the list of "grammatical words" that make up the *Interlingua-English Dictionary*. This list of words is described as "indispensable for the operation of the language".

However, some categories may not apply to all languages. For example, many languages such as Manam and Frisian lack reflexive pronouns and instead use personal pronouns in their place (Kiparsky, 2002).

Other categorisations and sub-categorisations of pronouns exist. For example, pronouns may be sub-categorised as subject or non-subject position, or anaphoric pronouns may be sub-categorised as inter- or intra-sentential.

For the purpose of translation, both manual and automated, pronouns in this thesis are categorised according to their *function*. Pronouns are categorised as belonging to one of eight *functional types*:

- Anaphoric/Cataphoric: Co-refers with an NP (its *antecedent*) – for anaphora the pronoun follows the antecedent and for *cataphora* the pronoun precedes the antecedent (e.g. "If **she** is in town, **Mary** will join us for dinner")
- Event reference: Refers to an event – this could be a verb, verb phrase, clause or an entire sentence
- Extra-textual reference: Refers to something that is not explicit in the text
- Pleonastic: Does not refer to anything, but is syntactically required (e.g "**It** is raining")
- Speaker reference: Refers to the speaker / author
- Addressee reference: Refers to a specific person or group of people in the audience
- Generic reference: Refers to people, or other living entities, in general
- Other: Any pronoun that does not fall into one of the categories above. This category includes indefinite pronouns (e.g. "anyone") and some numbers/quantifiers that are used as pronouns but are not themselves bare pronouns (e.g. "others", "each", "both")

These functional types closely match those presented in work on *functional grammar* in the field of linguistics. This is described in more detail in Section 2.1.1.

The categorisation of pronouns by their functional type helps to address the problem of functional ambiguity introduced in Section 1.1.2. For example the personal pronoun "it", which as described in Section 1.1 may be used as an

anaphoric, pleonastic or event reference pronoun. This combination of a pronoun form and function is referred to as a *form-function* pair.

1.5 Thesis

When faced with the translation of a pronoun in the source-language text it is useful to first identify its function. This is because different pronoun form-function pairs have different translation requirements in the target language. Not only may manual translation handle these form-function pairs differently but they raise different challenges to phrase-based, syntax-based and 'discourse-based' SMT (in part, because of the locality of information needed to make an acceptable choice). Additional pronoun-specific features may be used to further sub-categorise pronouns once their function has been determined. This framework of categorising pronouns according to their function (and additional features) may be used to support analyses of manual and automated translation, and in the design and evaluation of SMT systems.

1.6 Thesis Overview

Central to the thesis is the categorisation of pronouns according to the function that they perform in text. The eight functional types described in Section 1.4 provide a framework for categorising pronouns that is used throughout the thesis and forms the basis of the four strands of work outlined below: Corpus annotation, corpus analysis, the design of a post-editing SMT system and the development of a pronoun test suite to support SMT system design and evaluation.

1.6.1 Annotation

Work began with the construction of the ParCor corpus (Guillou et al., 2014) of parallel English-German texts in which pronouns and their features were labelled by human annotators. Pronouns in both source and target-language texts were labelled with respect to their *functional type* and location and, where relevant, their antecedent(s). Pronouns are categorised as belonging to one of the eight functional types described in Section 1.4. These eight functional types were selected on the basis that for SMT, pronouns of different types should be handled

differently. For example, anaphoric pronouns in German (and many other languages) should agree with the head of their antecedent in terms of number and gender[2]. Pleonastic pronouns, such as the "es" in "**Es** regnet" ("**It** is raining"), have no antecedent and therefore agreement is not relevant.

The annotation guidelines have been used in the annotation of English and German texts. They were designed to be as language-independent as possible, but additions or modifications may be required for other languages. I would anticipate that the same guidelines could be applied to French with few, if any, modifications. In the case of pro-drop languages such as Spanish or Czech, the guidelines and annotation scheme would need to be extended to handle the annotation of omitted pronouns. The current annotation scheme and guidelines cater for the annotation of explicit pronoun tokens only. Additional language-specific guidelines could also be added for the handling of other methods for expressing pronouns. For example, in Arabic, object position pronouns are expressed as suffixes attached to the verb. The current annotation scheme and guidelines cater for the annotation of complete pronoun tokens only. A different strategy would be required to accommodate the annotation of these suffixes in Arabic.

The ParCor corpus may serve as both a resource for understanding the differences in pronoun use between a pair of languages and as a gold-standard test set for SMT experiments. Construction of the corpus is described in Chapter 4.

1.6.2 Analysis

In the most comprehensive study on the translation of pronominal coreference to date, Hardmeier (2014) concludes that current models for pronoun translation are insufficient. To address this, he suggests that *"...future approaches to pronoun translation in SMT will require extensive corpus analysis to study how pronouns of a given source language are rendered in a given target language"*. The analysis strand of this thesis reports on such a corpus analysis. The first stage of the analysis looks at differences in pronoun use between original documents and their human-authored translations (see Chapter 5). Identifying and understanding systematic differences in pronoun use between a pair of languages may help inform the design of SMT systems and help us to infer practices that SMT systems should adhere to. The analysis makes use of the manual anno-

[2]Different pronoun-antecedent agreement features exist for different languages.

tations in ParCor and automatic word alignments between the English texts and their (human-authored) German translations. Similarities and differences in terms of pronoun use in English and German are observed at the corpus, document and sentence levels. The analysis reveals that the German translations in ParCor contain more anaphoric and pleonastic pronouns than their English originals, reflecting differences in pronoun use. This raises a particular problem for reference-based automatic evaluation of pronoun translation where pronoun use in the source-language text and reference translation diverges. In-depth investigations reveal some possible explanations for the differences in pronoun use. For example, relative pronouns corresponding to *relativizers* in the English text were found to account for some of the insertions of anaphoric pronouns in German.

The second stage of the analysis looks at how well state-of-the-art SMT systems perform at pronoun translation (see Chapter 6) and uses this information to identify the pronoun types where future efforts would be best directed. The analysis of SMT expands upon the findings from the analysis of manual translation. From the set of anaphoric pronouns, which are more frequent in German translations than original English documents, are selected the English pronouns "it" and "its" (possessive) and the English that- and null-relativizers. The pronouns "it" and "its" both have many possible translations in German due to the use of three grammatical genders: Masculine, feminine and neuter. Their translations, however, are subject to different agreement constraints in German. The English relativizers are of interest as they may trigger the insertion of a relative pronoun in German, as revealed by the analysis of manual translation. The analysis also reveals biases in the training data, which represents a particular problem that SMT systems must overcome.

1.6.3 SMT Design: Post-editing

Chapter 7 describes an automatic post-editing submission to the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015). The task focusses on the translation of subject position instances of the English pronouns "it" and "they" into French. Motivated by the need for functional disambiguation the post-editing method uses two rules to amend the output of a baseline phrase-based SMT system; one for anaphoric pronouns and the other for non-anaphoric pronouns. The underlying assumption of the post-editing approach is that through

using information from the source-language text, incorrect translations in the baseline SMT output can be identified and subsequently corrected. Using surface form, position (subject vs. non-subject), pronoun *type* and antecedent information about each source-language pronoun (obtained via external tools), the relevant rule is used to predict a suitable candidate pronoun translation. If the predicted translation and that in the baseline SMT output differ, the pronoun in the SMT output is replaced with the predicted value.

The performance of the external tools used in post-editing is measured, and an oracle experiment is conducted using ParCor-style manual annotations in place of the output of the external tools. This provides an assessment of the suitability of the post-editing method. The analysis showed that the design of the rules and their strict application to the SMT output are the biggest factors in the failure of the system. The anaphoric rule, which only generates (gendered) personal pronouns, ignores the possibility of using demonstrative pronouns (which are gender-neutral in French) and so lacks flexibility. The non-anaphoric rule is unable to distinguish between pleonastic and event reference instances of "it", and is restricted to replacing "incorrect" translations with "ce". Furthermore the strict application of the rules to the SMT output ignores the problem of introducing additional disfluencies into the translation.

1.6.4 Evaluation

No suitable automatic evaluation metric exists for pronoun translation and manual evaluation is both time consuming and costly. Framing the evaluation of pronoun translation as a pronoun selection task (Hardmeier, 2014) helps to remove the problem of biasing the human evaluator to what is output by the SMT system, and provides a relatively quick way to conduct manual evaluation. This method was used in the analysis of two state-of-the-art SMT systems in Chapter 6. However, to be able to make real progress in the design and implementation of discourse-aware SMT systems, automatic evaluation methods which allow for rapid development are required.

Chapter 8 presents PROTEST, a pronoun test suite to support the evaluation of SMT systems. The test suite consists of a set of 250 hand-selected pronoun tokens and an automatic evaluation script which identifies those SMT translations that match the reference, and refers the remainder for manual evaluation.

The test suite pronoun tokens are extracted from the test set of the DiscoMT 2015 shared task on pronoun translation for the English-French language pair. Pronoun tokens are categorised according to their function and sub-categorised according to a range of other features. The methodology behind the categorisation of pronoun tokens and the evaluation script are language independent, and extensions to other languages are considered.

1.7 Contributions

The main contributions of this thesis are:

- The ParCor annotation guidelines for the manual annotation of pronouns according to their function. These guidelines have been used to annotate the ParCor corpus and the English source-side texts of the *DiscoMT2015.test* dataset for the shared task on pronoun translation at the 2nd Workshop on Discourse in Machine Translation.
- The ParCor corpus of parallel English-German documents for which pronouns have been manually annotated. The corpus is available to download from OPUS, the online parallel corpus website[3].
- A corpus analysis of manual translation, the aim of which is to better understand the options for translating different types of pronouns.
- An analysis of pronoun translation for two state-of-the-art SMT systems, the aim of which is to understand how well systems translate different types of pronoun. The analyses of manual and automated translation provide insight into the need to handle pronouns according to their function in the source language and their translation requirements in the target language.
- The design and development of an automated post-editing system submitted to the DiscoMT 2015 shared task on pronoun translation.
- An analysis of the performance of the automated post-editing system, including an oracle experiment and an investigation into those factors affecting performance. This represents the first step of a wider analysis of all shared task systems.

[3]ParCor: http://opus.lingfil.uu.se/ParCor/

- A methodology for semi-automatically evaluating the translation of selected pronoun tokens by a machine translation system. Those pronoun translations that are not automatically matched to the reference translation are referred for manual evaluation.

- The PROTEST pronoun test suite for evaluating the translation of 250 hand-selected pronoun tokens in the *DiscoMT2015.test* dataset. The test suite will be released in an online repository.

1.8 Relation to Published Work

Some of the work in this thesis has already appeared in conference and/or workshop papers. In particular, the following chapters contain extended and/or updated versions of previously published papers:

- Construction of the ParCor corpus (Chapter 4) was a joint project between the University of Edinburgh and the University of Uppsala. The corpus and its annotation guidelines are presented in Guillou et al. (2014).

- The corpus analyses of manual translation and state-of-the-art SMT output described in chapters 5 and 6, extend Guillou and Webber (2015).

- Chapter 7 extends the description of an automatic post-editing system submitted to the DiscoMT 2015 shared task on pronoun translation (Guillou, 2015).

- The design and development of the PROTEST test suite for pronoun translation was a joint project with Christian Hardmeier (University of Uppsala). The content of Chapter 8 is based on Guillou and Hardmeier (2016).

Chapter 2

Literature Review

Pronouns have been studied widely in both linguistics and computational linguistics. The taxonomy of pronouns in functional grammar matches closely the set of functional type labels introduced in Section 1.4 which is used throughout this thesis.

Work on pronoun translation in SMT began with the efforts of Hardmeier and Federico (2010) and Le Nagard and Koehn (2010). This and the work that followed it has focussed on the problem of translating anaphoric pronouns, specifically the translation of the third-person pronouns "it" and "they" from English into languages with grammatical gender. This interest in translating anaphoric pronouns has culminated in the introduction of two shared tasks at the 2nd Workshop on Discourse in Machine Translation (DiscoMT) at EMNLP 2015. The first on *pronoun prediction* and the second on *pronoun translation*. Both tasks were concerned with English-to-French translation. Interest in other discourse-level phenomena, including lexical consistency and discourse connectives, also continues to grow and has been encouraged by the DiscoMT workshops.

There are three main directions of previous research that are relevant to this thesis. The first is the work in the broader fields of linguistics and computational linguistics on understanding the use of pronouns and their categorisation according to function. The second is the work on pronoun translation by both statistical and non-statistical MT systems. The third is the work on analysing manual translation, which comes rather from the fields of linguistics and translation studies than from computational linguistics. This chapter provides a review of the work carried out in both strands of research.

2.1 Pronoun Categorisation and Modelling

2.1.1 Pronoun Categorisation and Distribution in Linguistics

Pronouns are a universal feature of language (Greenberg, 1963). They belong, together with noun phrases (NPs), to the wider class of *referring expressions* which are used to identify entities and events in discourse. Michael (1970) traces the first use of *pronouns* as a separate part of speech to Dionysius Thrax, in the second century BC. According to early definitions pronouns are used as a "noun substitute".

In linguistics, pronouns are typically categorised according to their *form*. As outlined in Section 1.4 the following list of pronoun forms is typically provided in English grammar books: Personal, reflexive, reciprocal, possessive, demonstrative, indefinite, relative and interrogative. These pronoun categories, however, are not specific to English. A similar set of categories also appears in the *Interlingua Grammar* (Gode and Blair, 1951) which describes features that are common to a collection of source languages: English, Spanish, Portuguese, Italian, French, German and Russian. The Interlingua Grammar excludes categories for reciprocal and interrogative pronouns, but does include lexical entries in the list of "grammatical words" for some pronouns from these categories.

Another way to categorise pronouns is according to the *function* that they perform and it is this approach that led to the development of the taxonomy of pronouns used throughout this thesis. These pronoun functions are described in work on *functional grammar*, a general theory first introduced by Dik (1978), whereby the properties of natural language are defined according to their use. An excellent introduction to functional grammar is provided in Halliday (2004) and forms the basis of this section.

In functional grammar the class of a word indicates its potential range of grammatical functions. The functional potential of pronouns is defined by their location on each of the following vectors: Countability (count/mass), animacy (conscious/non-conscious) and generality (general/particular).

The category of English personal pronouns is subdivided into speech roles, other roles and generalised "one". *Speech roles* encompass the listener (i.e. *addressee reference*) and speaker (i.e. *speaker reference*). The speaker category is further subdivided into *speaker only* (e.g. "I"), *speaker plus listener* and *speaker plus other(s)* (e.g. "we"). In Wales (1996), the ambiguous pronoun "we" is di-

vided into *inclusive we* (referring to the speaker and addressee) and *exclusive we* (referring to the speaker and another). The concept of *"generalised one"* is equivalent to the *generic reference* function used in this thesis and has been expanded to include generic "you" (see also Wales (1996)) and "they". The *other roles* category includes the plural pronoun "they" and the singular pronouns "he" and "she" (*conscious* entities) and "it" (*non-conscious* entities).

In Halliday (2004) two forms of *phora* or *pointing* are described: Endophoric and exophoric. In the case of *endophoric* reference the entity to which the reference item (e.g. pronoun) refers is recoverable from the text. Endophoric reference covers both *anaphoric* reference in which the pronoun follows its antecedent, and *cataphoric* reference in which the pronoun precedes its antecedent. This contrasts with *exophoric* reference (referred to as *extra-textual reference* in this thesis) in which the entity to which the reference item refers is recoverable from the environment of the text, but not the text itself. Endophora and exophora achieve the effect of "pointing" either via *co-reference* (pointing to the same referent) or *comparative reference* (pointing to another referent of the same class). Personal and demonstrative pronouns, when used anaphorically, establish a relation of co-reference. In comparative reference, comparison is achieved via reference to general features of identity/similarity/difference or to particular features of quality or quantity, e.g. "same", "similar", "other" (general features) and "better", "fewer" (particular features).

Pleonastic "it" is referred to as *anticipatory it* in Halliday (2004). *Event reference* is not explicitly mentioned, but the use of demonstrative pronouns to refer to extended passages of text is highlighted. Event reference has also been referred to as *abstract anaphora* (Asher, 1993), i.e. anaphora involving reference to abstract entities such as events and states.

The syntactic behaviour of pronouns, anaphors and other referential expressions is described in *government and binding theory* (Chomsky, 1981). A set of conditions defines the binding domains for reflexive and reciprocal pronouns (belonging to the set of "anaphors"), pronouns (excluding reciprocals and reflexives) and other referential expressions (e.g. noun phrases). In the case of anaphors, the condition asserts that the anaphor must be locally bound, i.e. anaphor and antecedent must both be contained within the same clause. The other conditions state that pronouns must not be locally bound, and that other referential expressions can not be bound.

2.1.2 Modelling Pronouns in Computational Linguistics

Work on modelling pronouns and referring expressions in general, includes the construction of discourse models (Webber, 1978) and *centering theory* (Grosz et al., 1983, 1995).

Webber (1978) asserts that none of the following forms of anaphora can be understood in purely linguistic terms: *Definite anaphora* (using definite pronouns and noun phrases), *one-anaphora* (e.g. "My car is the red *one*"), *verb-phrase deletion* (e.g. "Whenever Wendy buys herself a new hat, Phyllis does ∅ too" [∅ = buy herself a new hat]). Instead they should be explained using a *discourse model*. The basic idea behind the discourse model is that the speaker has a model of something that they wish to communicate to the listener (i.e. the *addressee*), and their aim is to direct the listener to synthesise a similar model.

The work by Webber (1978) aims to address two complementary tasks that deal with anaphoric language. The first is identifying what a text potentially makes available for anaphoric reference. The second is in constraining the candidate set of a given anaphoric expression to a single possible choice. Whilst the second task had already received considerable attention (i.e. from work on anaphora and coreference resolution), the first had not. Webber (1978) proposes the construction of sentence-level semantic and syntactic representations from which the discourse entities (entities naturally evoked by the discourse) can be identified, and the methods of identification for each type of anaphora.

Centering theory (Grosz et al., 1983, 1995) provides a system of rules and constraints that govern choices made by discourse participants in terms of the type of referring expressions used, e.g. proper nouns, personal or reflective pronouns, etc. Discourse is viewed as dynamic with the current state (i.e. current point within an utterance) determining the centres of attention (i.e. entities which are being discussed). Centres are defined as entities that link the utterance to which they belong to other utterances in the same discourse segment. Grosz et al. (1995) define rules in terms of these centres, which are either forward-looking (to the next utterance) or backward-looking (to the previous utterance). The rules describe constraints over the possible options for realisation (e.g. when it is better to use a pronoun or a noun phrase) and preferences for types of transitions (i.e. continuity within the discourse is preferred to change). The main claim of centering theory is that the more a discourse adheres to centering constraints, the more coherent

it will be and the easier it will be for the listener to infer meaning from it. Whilst the work presented in Grosz et al. (1995) is theoretical, centering theory has been applied to the design of a number of algorithms for anaphora resolution (Brennan et al., 1987; Strube and Hahn, 1999; Tetreault, 2001). Anaphora and coreference resolution are discussed in Section 2.3.2.

2.2 Statistical Machine Translation

There are a number of Statistical Machine Translation (SMT) paradigms currently in use. Each provides a different approach to translation, but they all have one thing in common: They translate sentences in isolation. As noted in Section 1.1.1, this poses a problem for translating coreferential pronouns. In many languages, coreferential pronouns and their antecedents must agree in terms of features, and these features may vary between languages. For example, in German, French and Czech, an anaphoric pronoun and its antecedent(s) must agree in terms of number and gender. In many scenarios, a pronoun and its antecedent appear in different sentences, i.e. *inter-sentential* coreference. Under the current framework of sentence-by-sentence translation the pronoun will be translated without knowledge of its antecedent. Therefore the correct form of the pronoun in the target language cannot be guaranteed. However, this is not the only scenario in which pronouns may be translated incorrectly. It is possible that pronouns which occur in the same sentence as their antecedent (i.e. *intra-sentential* coreference) may be translated inappropriately if the pronoun and the antecedent fall into different translation units (n-gram, syntactic tree, etc.).

Phrase- and syntax-based systems are commonly used in SMT research. Both have been used in previous work on translating pronominal coreference and are used to provide the state-of-the-art SMT system output used in the analysis of pronoun translation by current systems (Chapter 6). However, many other paradigms also exist. Some of these are described below, with reference to their potential suitability for translating pronouns.

2.2.1 Phrase-based MT

Phrase-based MT (Koehn et al., 2003) extends the early word-based models – instead of translating individual words, phrase-based models aim to translate se-

quences of words, or *phrases*. These phrases may be linguistically motivated (i.e. cover a complete noun or verb phrase), but this is often not the case, and nor is it a requirement. Under the phrase-based paradigm, the target-language translation is constructed from left-to-right by stitching together phrases that cover the tokens in the source-language sentence. Translation is complete when all of the tokens in the original source-language sentence are covered by a phrase in the translation output. Multiple candidate translations, *hypotheses*, are constructed using a lattice structure. Each hypothesis is scored using a log-linear combination of features and the 1-best translation or an n-best list is returned. This method is simple and, despite lacking any linguistic knowledge, provides state-of-the-art translation for many language pairs. It has also been the SMT paradigm used for much of the work on pronominal coreference in SMT to date.

Hierarchical phrase-based MT (Chiang, 2005), also known as *Hiero*, is an extension of phrase-based MT. It aims to combine the strengths of phrase-based and syntax-based translation, using synchronous context-free grammar rules constructed using typical phrase-extraction methods of phrase-based MT – that is non-linguistically motivated sequences of tokens. Hiero phrases contain both terminals (words and punctuation symbols) and non-terminals (variables) to achieve a hierarchical structure in which phrases may be nested within each other. This nesting would perhaps allow for longer range pronoun-antecedent dependencies to be captured than in traditional phrase-based MT. However, as the non-terminals are unlabelled (just indexed instances of the symbol "X"), there is no constraint from the root label as to what phrase can substitute the non-terminal, and therefore no structure that could be leveraged to control for pronoun use.

2.2.2 Syntax-based MT

One of the reported issues with phrase-based MT models is that they do not incorporate sufficient linguistic knowledge to produce grammatical output (Och, 2003). Ahmed and Hanneman (2005) explain that such methods rely only on a language model to capture syntax. However, due to the small n-gram window, phrase-based MT cannot model long-range dependencies (e.g. pronoun-antecedent links) or even some local dependencies (e.g. those associated with topicalisation or question formation).

Syntax-based MT overcomes this problem by modelling the entire sentence

using a syntax tree. These trees may be generated using a monolingual parser run over the source-language string (in tree-to-string translation) or on the target side as the decoder constructs the translation (in string-to-tree translation). The overall aim of syntax-based MT is therefore to combine an explicit syntax representation with the benefits of statistical methods, namely the reduction in human effort in constructing the model. Human effort is then shifted from the construction of the translation model to development of the monolingual syntax parsers that are central to the translation process (Ahmed and Hanneman, 2005).

Syntax-based systems are still limited to the sentence level and are more complex to build and work with than phrase-based systems. They also exhibit comparatively lower BLEU scores than their phrase-based counterparts for some language-pairs. However, syntax-based systems may be suitable for addressing the problem of pronoun translation, due to their ability to handle longer range dependencies. They may be effective in targeting intra-sentential anaphoric pronouns as these are syntactically governed. In the case of inter-sentential anaphoric pronouns, syntax-based MT is unlikely to have any benefit over phrase-based MT.

Syntax-based SMT, in the form of the TectoMT system (Žabokrtský et al., 2008), has been used as the underlying SMT paradigm in work on pronominal coreference translation by Novák (2011) (cf. Section 2.3.1).

2.2.3 Factored MT

Another limitation of phrase-based MT systems is that the sequence of surface word forms over which they operate is limited, restricting the contextual information that can be taken into account. One attempt to allow for the inclusion of additional linguistic information in the translation process is factored MT (Koehn and Hoang, 2007). Factored MT allows for additional annotation at the word level using many possible factors including part-of-speech, lemma, gender, and case. A more complex representation is therefore maintained throughout the translation process. The motivations for the use of factored MT include the use of lemmas to overcome data sparsity issues and that the availability of morphological, semantic and syntactic information to the translation model allows for indirect modelling.

Factored MT introduces additional *mapping* steps that either *translate* input factors to output factors (phrase level), or *generate* new output factors from existing (output) factors (word level). Translation follows the traditional phrase-

based approach, but with the additional decomposition of the translation into a sequence of mapping steps. To implement a factored MT system, several other changes are necessary. Firstly, the training data used to build factored MT systems must also be annotated with the same extra features used in the translation process. Secondly, sequence models may be defined for any factor or set of factors. These sequence models act in the same way as a language model. For example a part-of-speech tag sequence model may be defined over high-order n-grams. Thirdly, as the additional modelling of the factors increases the computational complexity of decoding, additional pruning may need to be considered.

In terms of incorporating coreference information, a coreference resolution system could be used to annotate the training data. A sequence model, or similar, over pronoun-antecedent features would also be required. Whilst factored MT may seem an ideal candidate for incorporating additional coreference information, selecting the correct combination of features to obtain an improvement may be complex. Indeed this is a general problem of factored models. Results may also be difficult to replicate across language pairs with different feature-agreement constraints.

2.2.4 Document-level Translation

As pronominal coreference is a discourse-level phenomenon, document-level decoding may be suitable for handling it. Recent work in document-level decoding includes the development of cache-based models (Tiedemann, 2010a,b; Gong et al., 2011), post-editing methods (Xiao et al., 2011) and the optimisation-based Docent decoder (Hardmeier et al., 2012).

Cache-based translation (Tiedemann, 2010a,b) works by inserting into the cache those translation options used in the best translation hypothesis for each of the previously translated sentences in the document. A decay factor is added to the translation model in order to incorporate the notion of recency. When decoding source-language phrases that have an entry in the cache, a cache translation score is computed. This score forms an additional feature in the log-linear SMT model. One of the major problems with the technique, when applied to domain adaptation, is that it assumes that the cached hypothesis translations are accurate. Additionally, the translation cache is initially empty and is therefore of little use in the translation of sentences that appear early in the document. In terms of

pronoun translation, these problems may be less of a concern. Cache-based models could be used to record the translation of each antecedent so that when the pronoun comes to be translated, agreement can be encouraged/enforced. Even if the translation of the antecedent is incorrect, pronoun-antecedent agreement may still be desirable, and having an initially empty cache is of little consequence in this setting.

The extensions that Gong et al. (2011) propose, which are useful for addressing the problems of starting with an empty cache in the general SMT setting, may be of some use in pronoun translation. The *static cache* stores relevant phrase pairs from similar documents. This could be used to store a set of approved pronoun-antecedent translation pairs. The *topic cache* stores a set of target-language topic words taken from the target-side text of similar bilingual document pairs. It could be useful in improving the quality of antecedent translation where the system is trained on data from a different domain to that of the documents, or where the antecedent head word is ambiguous (e.g. "bank" may be used in the sense of finance or a river bank).

The structured topic cache extension of Louis and Webber (2014) for French-to-English domain-adapted translation of biographies, proves beneficial for the pronoun "he". In evaluating the structured topic cache translation versus a phrase-based baseline, Louis and Webber (2014) consider *impact words* — words present in the reference translation and structured topic cache translation, but not the baseline translation. In 30 documents, 36 instances of "he" and 36 of "his" are better translated. This improvement is not cited for any other pronouns. The technique of measuring improvements against a reference translation at the sentence level should be treated with caution (cf. Section 2.5.1) as the translation of anaphoric pronouns should also consider the head of the antecedent.

The two-step translation process proposed by Xiao et al. (2011) starts with the output translation from a baseline SMT system, which is then corrected at the lexical level. They apply the method to the problem of lexical consistency, identifying ambiguous words in the source language and replacing the translation of each with the most frequently occurring translation from the baseline SMT output. A similar approach could be taken to correcting the translation of pronouns, assuming that the pronoun's function and antecedent (in the case of anaphoric pronouns) is known. This idea is explored in detail in Chapter 7.

The Docent decoder (Hardmeier et al., 2012) provides a general framework

for document-level decoding. Rather than enforcing changes in a post-editing scenario (Xiao et al., 2011), translation is treated as an optimisation problem. Changes are made at the phrase level and are iteratively applied to the output of a baseline SMT system. Change operations include swapping phrases, changing phrases (i.e. selecting a new phrase from the phrase table), and re-segmenting phrases. Scoring is performed after each operation, with the change accepted if the amended version of the document receives a higher score than the current version. The process terminates if there is no change in score between a fixed number of successive steps, or if a predefined number of iterations is reached. This approach is proposed as an alternative to the dynamic programming beam search method which is suitable for decoding sentences in isolation, but would be computationally infeasible for a complete document due to the explosion of the search space. The Docent framework may be used to target a range of discourse-level phenomena. Linguistic knowledge is incorporated using decoder features. The challenge for MT researchers is in designing feature functions that target specific phenomena such that they make beneficial changes to the translation output and also in tuning to meaningful metrics. As described in Section 2.5.1 BLEU, the dominant metric in MT evaluation, is a general-purpose metric and is therefore not suitable for evaluating specific discourse-level phenomena, or for tuning discourse features. The use of BLEU for tuning is perhaps the major limiting factor of this and other document-level approaches. It may also explain why little improvement was observed for pronoun translation in the experiments by Hardmeier (2014).

2.2.5 Semantic-based MT

Early attempts to incorporate *semantic role* information in SMT include post-processing (Wu and Fung, 2009) and feature-based (Liu and Gildea, 2010) approaches. The two-pass model used by Wu and Fung (2009) uses semantic parses of the source text and the output of a phrase-based SMT system. If the semantic frames from the source text and MT output parses differ, segments (linguistically motivated phrases) of the MT output are reordered to better match the semantic frames of the source text. Liu and Gildea (2010) integrate features into a string-to-tree syntax-based SMT system. They parse the source-language text and project the semantic roles to the target language. Two features make use of

this information to reorder the MT output, and to penalise the deletion of semantic roles. The aim of the two approaches is similar — to preserve the semantic roles from the source text, in the MT output.

The use of a modality/negation annotation scheme in a syntax-based SMT system marks another example of the incorporation of higher-level *semantic* information to produce a *semantically informed* SMT system (Baker et al., 2012). In their approach, Baker et al. (2012) use tree-grafting procedures to add semantic information, originally projected from the target language, to syntactic parse tree fragments in the source language during training time. Grammar rules are derived from the partial trees and used by the decoder to generate translations. Using this approach, the authors achieve a BLEU score improvement over a Hiero-based SMT system.

The incorporation of semantic information in SMT is not limited to the work by Baker et al. (2012). Abstract Meaning Representation (AMR) (Banarescu et al., 2013), is a semantic representation scheme developed for use in various natural language processing tasks. It may be used in machine translation experiments as a form of inter-lingua. Under this paradigm an English sentence may be parsed into an AMR and from the AMR a sentence may be generated in another language. This generated sentence would then be a translation of the original English sentence in which the original meaning is preserved.

AMR expresses intra-sentential coreference via the use of variables with two instances of the same variable being linked. Whilst the AMR formats provide a means to move beyond the sentence level, with the possible linking of AMRs through coreferential links, the current focus is on the generation of AMRs at the sentence level. AMRs could, therefore, be used to tackle the translation of intra-sentential anaphoric pronouns at present, but not inter-sentential ones. The linking of inter-sentential anaphoric pronouns to their antecedents under the AMR paradigm, would require changes to AMR parsing. However, as AMRs abstract away from the sentence and syntactic layers, information such as pronoun surface form and syntactic roles, which would be useful in pronoun translation, are not available. AMRs would therefore need to be augmented with additional information before they could be useful for the pronoun translation task.

2.3 Pronoun Translation in SMT

2.3.1 Analyses of Pronominal Coreference in SMT Output

Analyses of pronoun translation by state-of-the-art SMT systems illustrate problems that arise when an anaphoric pronoun is translated without knowledge of its antecedent. Novák (2011) highlights how an English-to-Czech TectoMT system (Žabokrtský et al., 2008) always translates the English pronoun "it" into a third-person neuter pronoun in Czech, resulting in 44 out of the 64 anaphoric instances (68.75%) being translated incorrectly. Le Nagard and Koehn (2010) made similar observations about their English-to-French phrase-based SMT system with the English pronouns "it" and "they" too often translated into masculine forms. They report pronoun translation accuracy of their baseline system at 69%. Because these systems tended to default to the majority class in the training data, correct translation arises by accident, rather than by design. Hardmeier and Federico (2010) made similar observations in German-to-English translation. They claim that the extent of the errors varies depending on the pronoun being translated, with around 90% of demonstrative pronouns translated correctly compared with only a third of feminine pronouns translated correctly. Pronouns of polite address and reflexive pronouns are almost always translated incorrectly.

Weiner (2014) assesses English-to-German translation and opts for a different sub-categorisation of pronouns to Hardmeier and Federico (2010). Where Hardmeier and Federico (2010) sub-categorise pronouns according to their *form* (personal, demonstrative, reflexive, etc.), Weiner (2014) concentrates on anaphoric pronouns and categorises them according to their *surface form*: The pronouns "he", "she", "it" and "they" (nominative) and their objective and possessive forms. Instances of "he", "she" and "they" are almost always translated correctly by the baseline system. This is perhaps not surprising given that "he" and "she" are unambiguous in English, and because plural pronouns in German are un-gendered. The pronoun "it" is the one most often incorrectly translated and is therefore considered the hardest to translate. The accuracy of the translation of "it" by the baseline system was 47.6% for the news text genre and 47.2% for TED, based on sample sizes of 42 and 36 instances respectively. Weiner suggests that part of the problem could be the bias for translating "it" as "es"[1]. This

[1] "es" has other uses in German: being used as both an event reference pronoun and a pleonastic pronoun in addition to an anaphoric pronoun.

"it/es" bias is also observed in Section 6.4.3. Despite sub-categorising anaphoric pronouns as inter- and intra-sentential elsewhere in the project, Weiner does not use this sub-categorisation for the analysis of "it" translation.

These observations not only highlight that pronominal coreference poses a problem for SMT systems, they also indicate the problem of biases toward translating some source-language pronouns with certain target-language forms. This problem of translation bias is investigated in greater detail in Chapters 5 and 6. The observations made by Hardmeier and Federico (2010) and Weiner (2014), whereby some pronouns are translated with greater accuracy than others, also supports the main claim of this thesis: That functional differences necessitate the different handling of different groups of pronouns.

The analyses presented in previous research are limited in a number of ways. Some are conducted for a small number of pronouns (with the analysis by Novák (2011) conducted only for "it") and all are limited to instances where a source-language pronoun is translated as a target-language pronoun, with no investigation into other scenarios (i.e. insertions, deletions, explicitations in which the pronoun is replaced with an NP, etc.). The analyses are also typically performed for a small number of pronouns, with some studies considering fewer than 50 instances of a given pronoun. Given the manual nature of the work this is perhaps (largely) unavoidable.

2.3.2 Coreference Resolution

Anaphora resolution and the related task of coreference resolution have been the subject of considerable research within Natural Language Processing (NLP). Excellent surveys are provided by Strube (2007) and Ng (2010). Indeed, it is the availability of anaphora and coreference resolution systems that has made possible the recent work on pronominal coreference in SMT. Many of the methods proposed for translating pronouns have made use of knowledge provided by external anaphora/coreference resolution systems.

The Stanford Coreference Resolution system (Raghunathan et al., 2010; Lee et al., 2011), used in this thesis, is based on a multi-sieve approach. The sieve consists of several deterministic coreference resolution models, applied one at a time to the output of the previous model, starting with those that offer the highest precision. This is in contrast to other methods which apply only a single model,

running the risk of the large number of low precision features overwhelming the (often much) smaller number of high precision features. The sieve approach was shown to outperform a number of state-of-the-art systems in the CoNLL 2011 shared task on coreference resolution (Lee et al., 2011). The Stanford system also contains a sieve for pleonastic "it" detection. Unlike dedicated non-anaphoric "it" detection methods such as NADA (Bergsma and Yarowsky, 2011), the Stanford sieve does not consider the whole sentence, rather it uses a simple string-match method to identify instances from a set of fixed phrases.

The Stanford Coreference Resolution system and other more recent state-of-the-art anaphora and coreference resolution systems (e.g. BART (Versley et al., 2008), IMSCoref (Björkelund and Farkas, 2012), and the Berkeley Coreference Resolution System (Durrett and Klein, 2013)), are suitable candidates for incorporation within an SMT system. They have all been shown to perform well when compared with other systems in their "class" and are freely available. However, even current state-of-the-art anaphora and coreference resolution systems suffer from inaccuracy – a pronoun's referent may not be identified, or the wrong referent may be identified, or a pronoun may be identified as anaphoric when in fact it has a different function.

2.3.3 Integration of Anaphora and Coreference in SMT

Efforts in SMT have focussed primarily on the translation of English into languages with grammatical gender, including French, German and Czech. These target languages require anaphoric pronouns to take the grammatical gender that agrees with the pronoun's antecedent. For example, when translating into German, an instance of anaphoric "it" may be translated using a pronoun of neuter, feminine or masculine gender, i.e. "es", "sie" or "er" (cf. Section 1.1.1). These pronouns also need to be inflected according to the role that they perform with in the sentence – with four grammatical cases used in German. However, that is not to say that the translation of pronouns in the opposite direction is necessarily easier. As demonstrated by Hardmeier and Federico (2010), problems also exist for German-to-English translation. It is therefore worth considering not only the problems that exist for different language pairs, but for each pair, considering both translation directions. Given differences in grammar between different languages, it is likely that doing so will reveal different pronoun translation sub-problems.

Methods addressing the problem of pronoun translation may be categorised according to the point at which they are applied within the translation pipeline. Pre-annotation approaches such as the annotation projection method of Le Nagard and Koehn (2010) (also used by Guillou (2012)) are applied prior to translation. Methods applied during decoding make use of feature functions which are integrated in the log-linear SMT model. Examples include the word dependency model of Hardmeier and Federico (2010) and classifiers used to predict the correct translation of a pronoun using information from the source-language text and its translation (Novák et al., 2013; Weiner, 2014; Hardmeier, 2014). The final approach used to date is to automatically post-edit the pronouns in the SMT output (Weiner, 2014; Luong et al., 2015) (and cf. Chapter 7). Each of the approaches are described in more detail in the following sections.

Despite attempts to integrate coreference resolution within SMT systems, there has been little success. The results of the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015) serve to illustrate this. Of the six different systems submitted to the shared task, four of which make use of coreference resolution, none was able to beat the baseline system — a pure phrase-based SMT system with no discourse-specific features.

2.3.3.1 Pre-annotation Approaches

Pre-annotation involves annotating the source-language training and test data texts with features. The aim is to provide additional information that can be used to influence the translation of pronouns. The advantage of pre-annotation methods is their simplicity. However, errors introduced in the annotations as a result of incorrect anaphora/coreference resolution may lead to incorrect pronoun translations from which the decoder has no chance of recovery. The use of annotated training data can also result in increased data sparsity.

Le Nagard and Koehn (2010) detect and exclude instances of pleonastic "it" and focus on gender-correct translation of the third-person pronouns "it" and "they" in English-to-French translation. They implemented the coreference resolution algorithms described by Hobbs (1978) and Lappin and Leass (1994) and used this to obtain the antecedent of each instance of "it" and "they". For both training and testing data their method identifies the antecedent of each occurrence of "it" and "they" in the source-language text and the antecedent head translation in the target language. Grammatical gender of the antecedent trans-

lation is extracted and based on that gender, the occurrence of the pronoun in the source-language text is replaced with *it-masculine*, *it-feminine* or *it-neutral* for "it" or *they-masculine*, *they-feminine* or *they-neutral* for "they". A phrase-based MT system is trained using the annotated English source-language text and raw French target-language text and used to translate (source-language) test texts. In order to simplify the problem of obtaining the target-language translation of the antecedents, the output of a baseline SMT system is used. The baseline system is trained on un-annotated copies of the same parallel training corpus texts. This model is in effect a two-step translation process. The authors report no improvement over their baseline system and attribute this to poor performance of the coreference resolution algorithm(s), which they claim labels only 56% of source-language pronouns correctly. Another possible source of error in this method of two-step translation is that it assumes that the antecedent head will be translated in the same way in both translation passes. If the antecedent head is translated differently (i.e. is a different word with a different gender) in the second pass, this may lead to further errors as the pronoun is encouraged to be translated with the wrong gender (or at least one that does not match the gender of the antecedent). However, the effects of this error may be small in practice. Guillou (2012) identified only a very small number of antecedents that were translated differently between the two passes in a re-implementation of the method for English-to-Czech translation.

The work by Guillou (2012) applies the *source-side annotation projection* method of Le Nagard and Koehn (2010) with a number of differences. Firstly, the technique is applied to English-to-Czech translation. Secondly, a wider range of third-person personal pronouns is considered. Thirdly, the suitability of source-side annotation-projection is assessed. Experiments are conducted under *perfect* conditions by using a *gold-standard* manually annotated corpus in place of coreference resolution system output. This allows for the assumptions of perfect identification of coreferring pronouns and their antecedents. These assumptions could not be made even if a state-of-the-art coreference resolution system had been used, as such systems cannot yet achieve sufficiently high levels of accuracy. The output of the SMT system is evaluated against a baseline system (without the use of annotation-projection) using manual and automated methods (cf. Section 2.5.1). As reported by Le Nagard and Koehn (2010) the results showed little improvement over the baseline system. There are several explanations for

the poor performance, with training data size and data sparsity being possible factors. It is therefore not possible to provide a verdict as to the suitability of the method without first resolving these issues. Another factor is that the annotation-projection method focusses on the source side. As can be drawn from the rather limited successes of these methods in improving the translation of coreferential pronouns, this may not be sufficient. Instead operations on the target side should also be considered (Guillou, 2012).

Annotation-projection methods have been used elsewhere in SMT. Gimpel and Smith (2008) use it to capture long-distance phenomena within a single sentence in the source-language text via the extraction of sentence-level contextual features. These features are used to augment SMT translation models and better predict phrase translation. Similar techniques have also been applied to multilingual Word Sense Disambiguation whereby the sense of a word may be determined in another language (Diab, 2004; Khapra et al., 2009).

2.3.3.2 Decoder Features

Hardmeier and Federico (2010) introduce a word dependency module integrated in an English-to-German SMT system as a decoder feature. BART (Versley et al., 2008) is used to automatically identify coreferential pronouns and their antecedents in the source-language text. The coreferential links, which may be either intra- or inter-sentential, are encoded in the input to the decoder but translation remains at the sentence level. A decoder-driver manages the order in which the sentences are translated, such that the sentence that contains the pronoun's antecedent is translated before the sentence containing the pronoun. The translated sentences are then re-ordered according to the original order of the source document. Within the word dependency model, antecedent word forms are replaced using a tag representing the number and grammatical gender of the word. Pronoun-antecedent source word pairs representing the coreference links are tracked during decoding. When the pronoun in a pair is translated, the model adds an additional score representing the probability of the pronoun translation given the antecedent translation. This score is used in the decoder's search process. Using their own BLEU-inspired automated evaluation metric (cf. Section 2.5.1.1), they obtain small improvements in precision and recall, with the improvement in recall being significant. The largest gains in performance stem from translation of the pronoun "it", which is translated better by the system

that contains the word dependency model than by the baseline system. This improvement is encouraging as correctly translating "it" requires the accurate identification of the antecedent and extraction of its grammatical gender.

More recent work has framed the translation problem as one of cross-lingual pronoun prediction. Hardmeier (2014) worked on predicting the translation of instances of "it" and "they" for English-to-French translation. The prediction model takes the form of a neural network classifier that predicts for each instance, one of six classes: The masculine and feminine singular and plural third-person subject pronouns ("elle", "elles", "il" and "ils"), the impersonal pronoun "ce" or other (i.e. none of the previously mentioned pronouns is to be used). The classifier uses a combination of features describing the source-language pronoun and its immediate context, and the target-language antecedent candidate. Anaphoric links are treated as latent variables within the neural network – the classifier implicitly incorporates anaphora resolution, considering the set of nearby noun phrases as potential antecedents. The pronoun prediction model is integrated as a feature in the Docent decoder framework (Hardmeier et al., 2012). Alternatively an external anaphora resolution module, or gold-standard coreference annotated texts may be used used to identify the antecedents. Similar results are achieved using gold-standard annotations over the English ParCor (Guillou et al., 2014) texts as compared to implicit anaphora resolution. In general, the method yields very little improvement in terms of pronoun translation accuracy, which suggests that coreference resolution is not the only problem that SMT systems must contend with. The method's ability to disambiguate the different functions of "it" is not clear, but an inability to do so accurately could explain some of the translation errors. A version of the system was submitted to the DiscoMT 2015 shared task on pronoun translation and was ranked third out of the six participating systems (Hardmeier, 2015b). For this version the neural network classifier was trained with latent anaphora resolution (i.e. anaphoric links are modelled as latent variables) but the output of the Stanford coreference resolution system was used at test time.

Weiner (2014) applied two different approaches to improving pronoun translation using classifiers to predict the correct translation of instances of "it" in English-to-German translation. The first is a Discriminative Word Lexicon (DWL), a maximum entropy model that aims to predict the probability that a given target-language pronoun should be used in the translation. As with the approach

taken by Hardmeier (2014), the model makes use of a combination of source and target-side features including bag-of-words and bag-of-ngrams, alignments between the source and target-language words, previous words, nouns and the antecedent. A collection of DWLs are used, one for each target-language word. The second approach is a Source Discriminative Word Lexicon (SDWL). It is similar to the DWL but uses only features from the source side. A collection of SDWLs are used, one for each source-language word: "he", "she", "it" and "they". Both approaches were evaluated in terms of their ability to predict correct pronouns. Although both performed well for the pronouns "he", "she" and "they", neither performed well for "it". As the authors had identified that the translation of "it" was the most difficult for the baseline SMT system and the prediction for "it" was deemed poor, no attempt was made to integrate the classifiers in an SMT pipeline.

In contrast with the other approaches, Novák et al. (2013) use a syntax-based (TectoMT) framework. They present a discriminative model for the translation of "it" in English-to-Czech SMT. The feature is applied at the stage at which an English tectogrammatical tree is mapped to a corresponding Czech tree. The model considers several functions of the pronoun "it": *Referential* (referring to a noun or noun phrase), *anaphoric* (referring to a verb, verb phrase or larger segment of text) and *pleonastic*. It also considers the possibility of translating "it" as a personal pronoun, as a demonstrative pronoun, or dropping it from the translation altogether. BLEU scores of translations obtained by the baseline system and the discriminative model system show almost no difference. Manual evaluation was conducted for a sample of 50 sentences containing at least one instance of "it" and for which the translation produced by the two systems differed. This revealed that the discriminative model system produced a better translation than the baseline system for almost half (24 out of 50) of the sentences.

In contrast with methods used to pre-annotate the training data, the advantage of using feature functions is that the decoder has direct access to the pronoun-antecedent link information. This may help to mitigate the effect of errors introduced by external tools such as anaphora resolution systems, as the decoder has a chance to recover from bad decisions. However, tuning is a problem given the lack of an appropriate pronoun-sensitive, automated evaluation metric. If tuned for BLEU, which has been shown to be insensitive to small changes in pronoun translation (Hardmeier and Federico, 2010; Guillou, 2012), the feature

function may be given too little weight, with the risk of rendering it useless.

2.3.3.3 Post-editing

Depfix (Rosa, 2014) is a general tool for English-Czech post-editing. It incorporates a number of rules used to drop superfluous nominative pronouns and amend the morphological inflection of *some* pronouns when the incorrect case is used. The number of pronouns currently handled by Depfix is small and the focus is on addressing *simple cases*. Additional rules would need to be added for other pronouns.

Meyer et al. (2011) present a method which replaces pronoun translations in the SMT output with those suggested by a classifier, as a form of post-editing. Classifiers are trained to use an optimal combination of features (derived from the target side) to infer the correct gender of the pronoun for English-to-French translation of the pronoun "it". Unlike previous methods, coreference resolution does not play a part. Whilst the system corrects erroneous pronoun translations in the SMT output, it also replaces correct translations with incorrect ones. The net result, however, shows an improvement of about 10% in pronoun translation accuracy. In an extension to this work, Popescu-Belis et al. (2012) claim that the main problem of any approach that relies on identifying the pronoun's antecedent is that even state-of-the-art anaphora resolution systems perform poorly. They recommend the use of alternative methods, which do not require identification of antecedents. Their system uses the classification approach in Meyer et al. (2011), but in addition to the use of classifiers to correct pronoun translations, the method also judges whether a candidate translation (in the SMT output) should be changed in the first place. They report improvements in pronoun translation accuracy and also in BLEU scores. This is contrary to the results of efforts that concentrated on annotation-projection methods.

In addition to decoder feature methods, Weiner (2014) also applied two post-editing methods to the translation of "it". These approaches showed some small improvements over the baseline system performance. The first approach aims to directly detect and replace incorrect translations and makes use of anaphora resolution information from the source-side text and part-of-speech tagging over the SMT output. This information is used to identify those pronouns that do not agree with their antecedent and in the grammatical rules used to select the correct replacement. The method performed well but does not take grammatical

case into account, which is important in selecting the correct surface form of German pronouns. The method could be extended to include information about the verb and the semantic role of the pronoun to generate pronouns of the correct case. The second approach makes use of the same information, but performs N-best list re-ranking. The aim is to rank sentences that contain correct translations of the pronoun higher than those that contain incorrect translations. The N-best list re-ranking approach performed a little better than the direct post-editing method for the news text genre but little difference was observed between the two methods for the TED genre. This result is interesting given that the N-best list may in fact not contain the correct pronoun translation, a problem that direct post-editing methods do not have to contend with.

Two post-editing systems were submitted to the DiscoMT 2015 shared task on pronoun translation. The first is described in Chapter 7. The second incorporates pronoun prediction from a classifier (Luong et al., 2015). Common to both systems is the use of the Stanford coreference resolution system (Lee et al., 2011) to provide coreferential links for the English source-language text. Luong et al. (2015) used the coreference resolution information and the output of a baseline SMT system as features for their classifier. Their system was ranked first out of the six participating systems, but as with all systems, it failed to beat the official shared task baseline system.

The advantage of post-editing methods for pronoun translation tasks is that in the case of anaphoric pronouns the translations of both the pronoun and its antecedent are already known. There is therefore no need to keep track of this information within the decoder unlike those methods which use decoder features.

2.3.3.4 Other methods

Novák (2011) makes a number of suggestions as to how to incorporate coreference resolution within an SMT system, applied to both the source and target-sides of translation. The focus is on the problem of English-to-Czech translation and suggestions are made with reference to the TectoMT framework (Žabokrtský et al., 2008). With respect to coreference resolution applied to the source side, Novák (2011) expands on the methods of Le Nagard and Koehn (2010) and Hardmeier and Federico (2010). One suggestion is the use of longer coreference chains, rather than simply the closest mention of the antecedent, in order to pick antecedent translations more confidently. On the target side, a coreference resolution system

may be used to enrich a target-language tree model and give rise to new dependency relations that do not appear in the original corpus. The ideas are presented in the form of a position paper and no assumptions can be made as to whether any of the proposed work has yet been implemented. However, the suggestion to build coreference chains using a coreference resolution algorithm on the source side to provide more robust predictions as to the grammatical properties of the antecedents, seems logical.

2.4 Pronoun Translation in Non-statistical Systems

Although the focus of this thesis is on statistical machine translation, work on pronoun translation in non-statistical systems should not be ignored. Early and contemporary work on pronoun translation using both rule-based and example-based systems provides a useful comparison with work in SMT.

2.4.1 Rule-based MT

Unlike statistical MT which requires a considerable amount of parallel data with which to train translation systems, rule-based MT makes use of linguistic rules for the analysis of source-language text and for the generation of target-language text. The rule-based MT paradigm includes *direct*, *transfer-based* and *interlingua* methods. In the direct method, the transfer from source to target language is made at the word level with some simple grammatical adjustments. The transfer-based method achieves translation in three stages: Analysis of the source sentence and conversion to an abstract representation, mapping between source- and target-language representations, and generation of the target-language text. In the interlingua method the source-language sentence is transformed into a language-independent *interlingual* representation from which the target-language text is generated.

In the design of rule-based MT systems, specific rules may be designed for the translation of pronouns. These rules require input from linguists with an understanding of both the source and target language, and different rules will be required for different language pairs. Despite their linguistic motivations, these rules may also be insufficient in handling some cases as highlighted in the analysis of deficiencies of rule-based MT systems by Ferrández and Peral (2003). The

authors compared 16 commercial rule-based MT systems, including SYSTRAN[2]. They concluded that none of the systems can meet all of their criteria for a *complete solution*, i.e. can translate texts of any domain, resolve inter-sentential anaphora, identify co-reference chains within a text, and translate dropped pronouns into the target language.

The general strategy for targeting pronoun translating in rule-based systems is to apply anaphora resolution over the source-language sentences (possibly conjoining two sentences to simulate inter-sentential scenarios (Mitkov et al., 1995)) and integrate the output into the MT system via rules. Work in the 1990s was conducted for a range of language pairs, translation directions and systems. An excellent survey of this work is provided in (Mitkov, 1999b). Interest in this research area appears to have reached a peak at the end of the 1990s, with a special issue on anaphora resolution in machine translation (Mitkov, 1999a).

Pro-drop presents a problem for both statistical and non-statistical MT. When translating from a pro-drop language such as Japanese into a non pro-drop language such as English, it is necessary to identify pronouns omitted from the source-language text and to ensure that they are inserted into the target-language translation. Nakaiwa and Ikehara (1995) developed resolution methods for both intra- and inter-sentential dropped pronouns. These methods were integrated in a transfer-based Japanese-to-English MT system (ALT-J/E) in the form of an algorithm which defines how omitted pronouns in Japanese are to be translated into English. Ferrández and Peral (2003) follow a different approach for Spanish-English translation. Their method first identifies the position of dropped subject pronouns in Spanish and then inserts the pronoun into the sentence prior to translation into English. Translation is carried out using the AGIR interlingua MT system. The authors highlight that their method works particularly well for omitted plural pronouns, and that it outperforms SYSTRAN in the handling of pro-drop.

Increased interest in pronoun translation in SMT has been accompanied by renewed efforts in rule-based MT. Recent efforts have included a study to compare the performance of the transfer-based Its2 system and a statistical MT system, on the translation of omitted pronouns for Spanish-to-French and Italian-to-French (Russo et al., 2012a,b). The conclusions of this study were that the SMT systems

[2]SYSTRAN (www.systransoft.com) was originally rule-based but is now a hybrid i.e. rule-based / statistical system.

generally outperform the Its2 systems and that instances of personal pro-drop are typically easier to translate than impersonal pro-drop for both language pairs. Poor performance of the Its2 system is attributed to a lack of generation rules for impersonal pro-drop and to the lack of an anaphora resolution module in the case of personal pro-drop. In their submission to the DiscoMT 2015 shared task on pronoun translation (for English-to-French), Loáiciga and Wehrli (2015) incorporate an anaphora resolution into the Its2 system. Rules are provided for the translation of 3rd personal pronouns but not for impersonal pronouns, which may explain poor performance as compared to the baseline and other participating systems.

2.4.2 Example-based MT

Example-based machine translation (Nagao, 1984) is based on the concept of analogy and makes use of corpora containing texts which have already been translated. Given a new source-language sentence that is to be translated, *example* sentences that contain similar sub-sentential components are extracted from the corpus. The translations of these similar sub-sentential components are extracted from the examples and combined to construct a target-language translation of the source-language sentence.

There is little available literature on the specific problem of pronoun translation for example-based MT outside of work on pro-drop for Japanese-English translation. Kurohashi et al. (2005) include a module in their IWSLT 2005 system to cater for the omission of pronouns in Japanese-English translation. Their system handles the problem by using a language model of English which scores translations both including and excluding a generated English pronoun and selects the *best* translation. In the IWSLT 2006 system by the same group (Nakazawa et al., 2006), pronouns are estimated using information about modality and subject case. An extra node, representing an omitted pronoun in Japanese, is inserted prior to translation. Again, a language model is used to score translations with and without the generated pronoun. Pronoun translation, however, is not the main focus of the system and no investigation is made into whether pronoun translation is improved by the system.

2.5 Evaluation

Evaluation in Machine Translation falls into two main categories: Manual and automated. Whilst manual evaluation provides a definitive assessment of system performance, it is both costly and time consuming owing to human involvement. Automated evaluation metrics that correlate well with human judgements are therefore highly sought after.

Automated evaluation metrics typically score MT output against a single reference translation, and if multiple reference translations are available the set is typically rather small. In natural languages, there are often a great number of valid ways of expressing the same thing. Limiting automated evaluation to the use of a single reference translation, or at best a very small set, means that what may be deemed a *good* translation by a human, will be assigned a low score by the automated metric if it differs greatly from the reference(s). Paraphrasing techniques to generate synthetic reference translations may provide a possible solution to this problem but are outside of the scope of this project.

Manual methods for evaluating MT typically centre around assessing adequacy and fluency. *Adequacy* measures the extent to which the translation captures the meaning of the original source-language sentence. *Fluency* is a measure of the grammaticality and naturalness of the translation. Other methods that target specific phenomena, such as pronouns or discourse connectives, may focus on counting the number of correct translations.

2.5.1 Pronoun Translation

2.5.1.1 Automated Evaluation

BLEU (Papineni et al., 2002) is currently the dominant automated evaluation metric in SMT. Despite its wide adoption in assessing the general quality of translation, BLEU has been rejected as an unsuitable metric for discourse-level phenomena. Le Nagard and Koehn (2010) reject BLEU on the grounds that they expect to observe only a small number of changes between the output of their system and the baseline (i.e. only the pronouns). Little variation in scores is therefore be expected. Instead, they resort to manually counting the number of correctly translated pronouns in the MT output.

Hardmeier and Federico (2010) claim that as a general-purpose method BLEU is not suitable for specifically targeting the impact of their model on pronoun translation. Instead they propose a new BLEU-inspired method that defines both precision and recall with respect to a single reference translation and incorporates BLEU's notion of a *clipped count*. For each pronoun in the source-language text, the set of aligned target-language words in the reference set and the candidate translation are obtained. The clipped count of a given candidate word is taken to be the number of times that it occurs in the candidate translation, limited by the number of times it occurs in the reference set.

Guillou (2012) also rejects BLEU on the grounds that it is too general-purpose to reflect the changes of a pronoun-aware system over a baseline system. It is also not well suited to the evaluation of Czech translation in general due to the large vocabulary size of the Czech language (Bojar and Kos, 2010). The precision/recall metric of Hardmeier and Federico (2010) is also deemed unsuitable due to the way in which it compares pronoun translations in the MT output to those in the reference translation. The metric compares the translation of an anaphoric pronoun with that in the reference translation without also considering the translation of the head of the antecedent. This can lead to pronoun translations being scored as correct when pronoun-antecedent agreement does not hold, and with pronouns being scored as incorrect when they are in fact valid alternative translations (for which agreement holds). Instead, Guillou (2012) uses methods of automated and manual counting of correct translations. The automated evaluation method hinges on the requirement that a Czech pronoun must agree in number and gender with its antecedent. This is implemented as a count of the number of pronouns in the (Czech) SMT output that agree in number and gender with the Czech translation of their antecedent (as identified in the English source-language text and projected to the SMT output).

Weiner (2014) also makes use of an automated method for counting the number of correctly translated pronouns. Again a correct translation is defined as one where the antecedent and pronoun agree in terms of number and gender. Agreement is checked using the part-of-speech (POS) tags, extracted using the RF Tagger[3] which tags both number and gender, for the pronouns and antecedent heads. This relies on accurate POS tagging, which may not be possible over dis-fluent SMT output. Weiner therefore also suggests a method of manually correcting

[3]RFTagger: http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/

the pronouns in the SMT output and calculating how many of the gold-standard translations are achieved by the system. These methods are applied to the output of English-to-German translation systems and appear to ignore whether the pronoun translation takes the correct case.

Despite the dominance of BLEU in SMT evaluation, other automated evaluation metrics exist, including METEOR (Banerjee and Lavie, 2005), SemPos (Kos and Bojar, 2009), Translation Error Rate (TER) (Snover et al., 2006), Hybrid Translation Edit Rate (HyTER) (Dreyer and Marcu, 2012) and NIST (Doddington, 2002). Being general-purpose MT evaluation metrics like BLEU, they are also unlikely to be well suited to the particular problem of pronoun translation. Some of these general-purpose metrics were, however, used to assess the performance of systems submitted to the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015).

2.5.1.2 Manual Evaluation

Automated methods are prone to problems such as reliance on external tools and resources to identify the antecedent head and its number and gender, or reliance on the reference translation. In contrast, manual methods are slow and expensive. However, with no standard automatic metric available for pronoun translation, manual evaluation represents the most accurate method for assessing system performance. Simple methods rely on counting the number of pronouns translated correctly and incorrectly, and have been used by Le Nagard and Koehn (2010) and Guillou (2012). The manual evaluation in Guillou (2012) relied on a native Czech speaker to provide a number of judgements. These included whether the pronoun translation was correct, if incorrect whether it could still be understood by a native speaker, and which system (i.e. pronoun-aware vs. baseline) provided a better translation.

In Hardmeier's (2014) pronoun-selection task, human expertise is leveraged in such a way that the cost of gathering translation accuracy judgements is minimised. In this task, a human annotator is presented with the original source-language text and its translation. Up to five sentences of previous history are provided, for both the original source-language text and the translation. This ensures that the annotator has sufficient context to identify the antecedent of each anaphoric pronoun, as well as its translation. In the final sentence of each example block the pronoun is highlighted in the original source-language text,

and its translation is replaced with a placeholder value. The annotator's task is to select those pronoun(s) which may be used as valid replacements for the placeholder. When carrying out this task, minor dis-fluencies such as incorrect verb agreement or obviously missing words should be ignored, and the pronoun that best agrees with the antecedent in the SMT output should be selected, even if the antecedent is translated incorrectly. Pronouns produced by the system may then be compared with the gold-standard selections made by the annotator. This is similar to the method described in Weiner (2014), although there are some differences. Firstly, the pronoun-selection task allows for the selection of multiple possible valid pronoun translations. Secondly, because the pronoun translation is obscured by a placeholder value, the effect of the annotator being biased by the pronoun in the MT output is removed. The disadvantage of the pronoun-selection method, as with all manual evaluation methods, is that the task must be repeated each time a new translation is generated.

2.5.2 Other Linguistic Phenomena

BLEU scores have also been rejected as being unsuitable for evaluating the translation of discourse connectives (Meyer and Popescu-Belis, 2012). Meyer and Popescu-Belis (2012) instead manually count the number of discourse connectives correctly translated by the discourse-aware and baseline systems.

The ACT metric (Hajlaoui and Popescu-Belis, 2013) presents a semi-automated alternative to complete manual evaluation of discourse connectives. It scores the translation of source-language discourse connectives in the MT output against those in the reference translation. To identify the translation of a discourse connective, ACT first consults a dictionary of possible translations. If either the reference translation or the MT output contain more than one possible translation of a source-language connective, word alignments are used to identify the correct translation.

Cases where the connective in the MT output is identical to the one in the reference translation, or a synonym, are handled automatically and marked as "correct". Cases where the connective in the MT output is not equivalent to the one in the reference, or the connective is included in the reference but missing from the MT output are automatically marked as "incorrect". However, translations in which the connective is missing may still be correct as the connective may be

implicit in the translation (i.e. it can be implied from the text). It is possible that although the reference translation may contain an explicit connective, the use of an implicit connective could still be a valid translation option. Therefore such cases should not be handled automatically. In all other cases, manual assessment is required, to a higher or lower degree: Those cases where a connective appears in the MT output but not the reference, or where neither the MT output nor the reference contain a connective. The ACT score is computed as the ratio of the total number of correct translations to the total number of connectives in the source-language text.

The obvious drawback of ACT is the need for manual evaluation. Although the burden on the human evaluator is greatly reduced through automatic scoring of the *simple cases*, human involvement is still necessary. Manual evaluation of the remaining cases is required in order to produce a complete and definitive evaluation of system performance. The metric also relies on the reference translation and this raises the issue of what happens when there are big differences between the automated translation and the reference. Such differences may arise when the reference is a paraphrase of the original source-language text and the automated translation is a close translation of the source. Despite these concerns, a method with parallels to the ACT metric, is proposed for semi-automatic evaluation of pronoun translation in Chapter 8.

2.6 Analyses of Manual Translation

Analyses of manual translation may be useful in guiding the development of discourse-aware SMT systems. Indeed, in the most comprehensive study on the automated translation of pronominal coreference to date, Hardmeier (2014) suggests that extensive corpus analyses may be required to better understand the problems that SMT systems face.

2.6.1 Pronouns in Manual Translation

Whereas MT systems typically aim to provide a close translation of the original source-language text, human translators often work in a less constrained manner. The medical and legal domains may require translations that are close to the original source-language text in order to prevent changes in meaning. However, in

other genres such as news reporting or the translation of literary texts, translators may be granted a certain degree of artistic freedom, allowing them to produce fluent translations that appear natural in the target language yet still capture the meaning of the original text.

Human translated texts are often written in what has been called *translationese* (Gellerstam, 1986) — a dialect of the target language that exhibits elements of the text's source language, modulated by the translation process. Baker (2003) categorised the changes that translators typically make. These include *simplification*, *explicitation* (spelling things out) and *normalisation* (conforming to patterns or conventions of the target language). Translated texts are often shorter and tend to make use of certain discourse markers with greater frequency than the original texts (Koppel and Ordan, 2011). These differences are often marked and can be used to automatically detect whether a text is an original or a translation (Baroni and Bernardini, 2006; Koppel and Ordan, 2011).

Conducting an investigation into the most salient function words that can assist in detecting original vs. translated texts, Koppel and Ordan (2011) describe a number of differences between original English texts and translations from other languages. They consider English translations of French, Italian, Spanish, German and Finnish texts in the Europarl corpus (Koehn, 2005). In this corpus first, second and third-person personal pronouns are under-represented in the translations, compared to the original texts. They suggest that this could be due to simplification or to explicitation, where anaphoric pronouns are replaced with noun phrases.

2.6.2 Comparison of Pronoun use in Written and Spoken Genres

Pronoun use not only varies between original texts and translated ones, it also varies by genre. For the English-German pair, Ruiz and Federico (2014) found that TED Talks[4] (spoken language genre) contain three times as many first and second-person personal pronouns and twice as many instances of the third-person pronoun "it", as compared to the News Commentary texts[5] (written genre). This finding is mirrored in the analysis of manual translation in Chapter 4, in which

[4]The TED Talks were taken from the IWSLT shared task datasets.
[5]The News Commentary texts were taken from the WMT shared task datasets.

greater pronoun density is observed for TED Talks than the EU Bookshop documents. However, the usefulness of comparing their study to the one in this thesis is rather limited. Ruiz and Federico (2014) do not describe their method for identifying pronouns and so it is not possible to ascertain whether event reference and pleonastic pronouns are included in the counts for "it". Furthermore, differences between English and German in terms of pronoun distributions, or the quality of their MT systems at translating pronouns in texts of different genres, are not assessed. Despite the limitations, the overall recommendation is sound; that is, the TED Talks dataset with its high pronoun density provides a good starting point for further investigation into the translation of pronouns.

Ruiz and Federico (2014) also comment on the issue of ambiguous pronouns, highlighting the issue of translating the English pronoun "you" which has multiple possible translations in German: "man", "Sie", and a number of indefinite pronouns indicating "someone". In some scenarios local context will be sufficient to resolve such ambiguities, but in others correct pronoun translation will simply be down to chance. This theme of ambiguity runs central to the work in this thesis.

2.6.3 Comparison of Pronoun use for English-German

Analyses of pronoun use in manual translation by Becher (2011) and Kunz and Lapshinova-Koltunski (2015) are relevant to the work in this thesis. Both analyses were conducted for English-German translation and both highlight differences in pronoun use between these two languages.

Becher (2011) identifies instances of explicitation (addition) and implicitation (omission) of pronouns in documents from the *business letters* genre. Differences between the source-language text and target-language translation are expressed in terms of *additions*, *omissions* and *substitutions*. Pronouns are sub-divided into two categories. The *interactive pronouns* category contains the first and second-person personal pronouns. The *cohesive pronouns* category contains all other pronouns, including third-person personal pronouns, demonstratives, pronominal adjectives and pronominal adverbs.

In terms of interactive pronouns (i.e. speaker/addressee reference in this thesis), Becher reports a greater number of additions in the German-to-English direction than the opposite direction, and a smaller increase in the number of

omissions in the English-to-German direction. The net effect is that more interactive pronouns are added in German-to-English translation. Possible reasons for differences in pronoun use include the conversion of a passive sentence into an active sentence, and vice versa, and the addition of personal pronouns to make texts clearer and avoid misunderstandings. As Becher does not list the pronouns that are marked for each language, it is not possible to draw conclusions based on general knowledge such as the tendency to use "man" (generic, third-person) in German where the corresponding English sentence is passive. The omission of pronouns in translation may be less likely (than additions), as it could lead to misunderstandings. Substitutions occur infrequently in the corpus and so it is not possible to draw concrete conclusions from the analysis of these instances.

In terms of cohesive pronouns the results of the analysis suggest that translators like to add pronouns when the opportunity arises. Looking at the overall patterns, there are a greater number of additions in the German-to-English direction and more omissions in the opposite direction. This is the same observation that was made for interactive pronouns and can be explained by the following actions. Firstly, demonstrative pronouns may be substituted for definite articles or vice versa. Becher asserts that translators omit pronouns when they believe that the reader will be able to infer the coreference relation from the text, and add pronouns when they believe this inference is not possible. Secondly, German offers a wide range of both pronominal adjectives and pronominal adverbs, many of which have no obvious equivalent in English. English translators may either paraphrase the German pronominal adjective/adverb or opt not to translate it at all, with the latter option being common. Thirdly, in the German-to-English direction, pronouns may be added as a result of the preference in English of marking possession through the use of possessive pronouns. In German, the relationship between the possessor and possessed is often implicit in a text, so possessive pronouns may be omitted in the English-to-German direction. Becher cautions that this difference may be a result of the common use of possessives in business-related texts, which may not be observed in other genres. Indeed for the ParCor corpus (see Chapter 4) more pronouns are observed in the German translations than the English original texts, for both the TED Talks and EU Bookshop corpora. Possessive pronouns do occur in both TED Talks and EU Bookshop documents, but are less common than other categories of pronoun form.

Kunz and Lapshinova-Koltunski (2015) provide an analysis of English and

German original texts, comparing frequency distributions[6] of cohesive types and sub-types in the GECCo corpus (Lapshinova-Koltunski and Kunz, 2014). The GECCo corpus contains a collection of English and German texts and their translations (into the opposite language). The corpus is manually annotated for (co)reference, substitution, ellipsis, conjunction and lexical cohesion. In terms of reference, the study reveals that English uses more personal pronouns and demonstrative determiners than German, and that German uses more demonstrative pronouns, pronominal adverbs and comparatives than English. Possible reasons for the differences between English original texts and German translations are that a) personal pronouns are used in English where the demonstrative pronouns "der" and "die" could be used in German and b) some uses of the pronoun "it" in English could be realised with pronominal adverbs in German (also noted by Becher).

Both analyses were conducted from a translation studies perspective, with no consideration as to the implications for SMT. With respect to SMT, an understanding of the differences in pronoun use as licensed by the source and target languages is important. Not only are differences present in the data on which models are trained (so the models may learn them), but also in the reference translations against which MT output is compared. If the goal is to one day provide accurate translation of pronouns, we must also understand where and when it is acceptable and necessary to include/omit pronouns.

2.6.4 Parallel Corpora with Coreference Annotation

Parallel corpora in which pronominal coreference is annotated may serve many purposes. They provide gold-standard coreference annotation over a set of texts which may be used as a test set in SMT experiments. They also provide a means of comparing pronoun use across a pair of languages, and ultimately, may serve as a useful resource for identifying systematic differences in pronoun use between those languages. Identifying and understanding these systematic differences may help to inform the design of SMT systems and help us to infer practices that SMT systems should adhere to. Analyses of manual translation may be conducted using raw texts. However, the provision of annotated corpora can simplify the summarisation of differences between source and target-language texts at higher

[6] As the comparison is between English and German original texts, rather than original texts and their translations, raw counts are not provided.

levels, as we will see in Chapter 5. At the time of writing, there are few parallel corpora in which pronoun coreference is annotated.

Popescu-Belis et al. (2012) annotated a portion of the English-French Europarl corpus, using the *translation spotting* method to annotate pronouns (Cartoni et al., 2011). Using this method ∼400 tokens of the English pronoun "it" in the source-language text were manually annotated with their translation in the target language. This parallel annotated data was used to train classifiers to predict the French translation of new instances of "it", resulting in a small increase in BLEU score.

The Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0) (Hajič et al., 2012) contains the original English plus a close translation into Czech of the Penn Treebank corpus (Marcus et al., 1993). The corpus offers rich linguistic annotation over a number of layers that are provided in addition to the raw texts – the word, morphological, analytical and tectogrammatical layers. The *word layer* contains the tokenised plain text over which the *morphological layer* adds part-of-speech tags and lemmas for each token. The *analytical layer* represents the sentences as (surface-level) syntactic parse trees. The *tectogrammatical layer* is a linguistic representation that combines syntax, semantic labelling, anaphora resolution and argument structure. The representation is based on the framework of the Functional Generative Description (Sgall et al., 1986). Unlike the other three levels, the tectogrammatical layer may contain nodes that do not correspond to surface-level components of the sentence, such as pro-drop. The corpus is relatively small (approximately 50,000 sentences) compared with other corpora used in training SMT systems. It was also originally designed as a multi-purpose linguistic resource: Its use in SMT experiments was not its primary goal. The corpus has, however, been used in SMT experiments. At the time of writing, it has been used in experiments on the translation of pronouns (Guillou, 2012) and discourse connectives (Meyer and Poláková, 2013).

The GECCo corpus (Lapshinova-Koltunski and Kunz, 2014) is an English-German parallel corpus containing a collection of documents from the written and spoken text genres. The texts were taken from the CroCo corpus (Hansen-Schirra et al., 2012) and the original language (English or German) of each text is known. Cohesive devices are labelled as belonging to one of five types, based on those suggested by Halliday and Hasan (1976) for English and adapted for the multilingual setting: (co)reference, substitution, ellipsis, conjunction and lex-

ical cohesion. The rationale behind these categories is that they apply to both English and German and could also be applied to other languages. This use of common categories is necessary in order to be able to draw comparisons between languages, both in the manual and automated translation settings. In terms of the *reference* category, personal and demonstrative pronouns and comparatives ("bigger", "better" etc.) are labelled in the corpus. Coreference chains are also annotated. At the time of writing the corpus is not available to the public.

Chapter 3

Resources

A number of tools and resources were used to complete the work described in this thesis. These are described below.

3.1 Tools

3.1.1 MMAX-2

MMAX-2 (Müller and Strube, 2006) is a graphical desktop-based annotation tool suitable for the annotation of many linguistic phenomena, including coreference, in a text. The tool allows for multiple annotation layers, each with its own annotation scheme (i.e. guidelines/rules) defining the attributes to be labelled. Spans of text within MMAX-2 are referred to as *markables*. These are defined in terms of their span and a set of attributes describing their properties. Information about markables is stored in XML-format files, with each file representing a single annotation layer.

The annotation of markables and their attributes, as well as the appearance of the text within MMAX-2, are customised via XML-format stylesheets. The *Scheme* stylesheet contains details of attribute types (both attribute name and value) and the dependencies between them, and reflects the annotation scheme (i.e. guidelines/rules) of the corpus. It also defines the options available in linking markables together (i.e. in the case of linking a pronoun and antecedent, an option might be "mark as coreferent") and the graphical representation of the links. The visual appearance of the text is defined via the *customisation* and *style* stylesheets. Customisations include font colours, styles (bold, italic, etc.) and

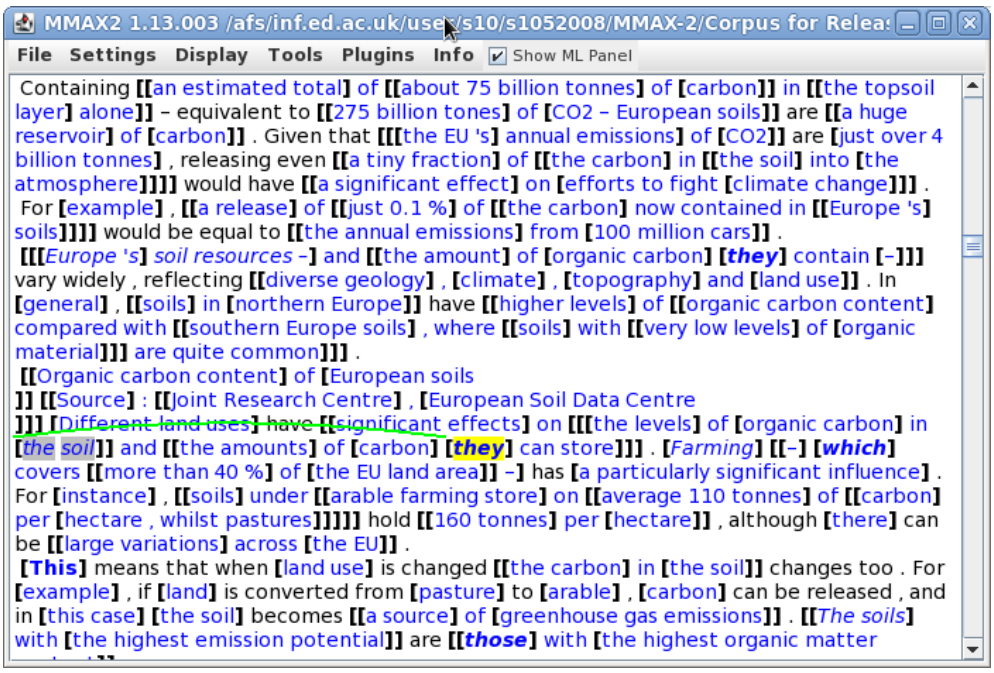


Figure 3.1: MMAX-2: Main window

text highlighting. Styles are global stylesheets that define the overall appearance of text and markables. Both scheme and customisation stylesheets exist for each annotation layer.

MMAX-2 is a tool consisting of three windows. The main window (see Figure 3.1) contains the text to be annotated and the attributes window (see Figure 3.2) contains the attribute values that the annotator has set for the anaphoric pronoun "they" highlighted in the main window. The final window is the markable control window (see Figure 3.3) through which the annotator can determine which of the annotation layers is *active* for annotation. MMAX-2 is used in the annotation of the ParCor corpus (Guillou et al., 2014) described in Chapter 4.

3.1.2 LFAligner

LFAligner[1] is tool for sentence-aligning bilingual parallel texts and constructing translation memories. It provides a wrapper for Hunalign (Varga et al., 2005), the underlying sentence-alignment tool.

The input to LFAligner is a bilingual corpus of tokenised and sentence-segmented

[1]LFAligner: http://sourceforge.net/projects/aligner/

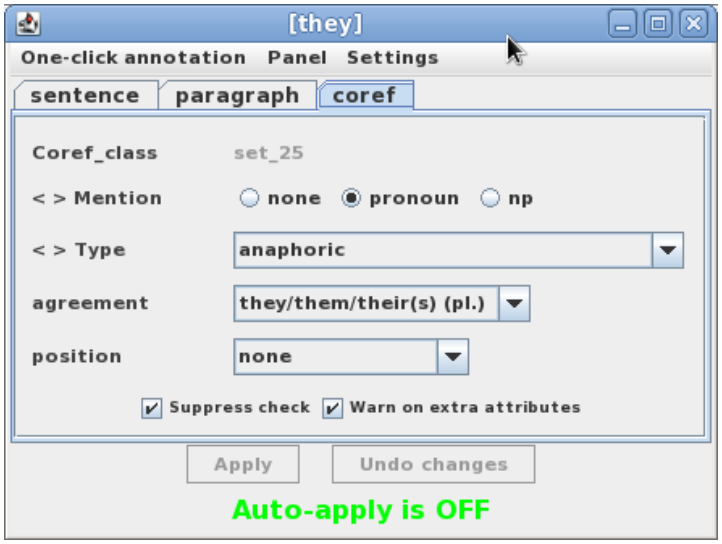


Figure 3.2: MMAX-2: Attributes window

Figure 3.3: MMAX-2: Markable control level window

texts, and optionally, a bilingual dictionary. The dictionary is used in combination with the Gale and Church (1993) sentence-length based alignment information to construct the sentence-alignments.

LFAligner is used to sentence align the parallel texts of the ParCor corpus (Guillou et al., 2014), the annotation of which is described in Chapter 4.

3.1.3 Berkeley Parser

The Berkeley parser (Petrov et al., 2006) is a constituent parser with pre-trained grammars for English, German and a number of other languages. It is incorporated in both the English and German pre-processing pipelines used in the annotation of the ParCor corpus. The pre-processing pipelines are described in Section 4.6.1.

3.1.4 German Markables Pipeline

The *German markables pipeline* used in the pre-processing of the German ParCor texts (see Section 4.6.1) is described in Broscheit et al. (2010) and Versley et al. (2010). The pipeline first parses the texts using the Berkeley Parser (Petrov et al., 2006). It then extracts nominal (both minimal and maximal noun projections) and pronominal mentions from the resulting parse trees. Morphological tagging described in Broscheit et al. (2010) provides number and gender information for the mentions as well as their type (definite/indefinite NP, name, personal/relative/reflexive pronoun).

The pipeline takes as input an MMAX-2 format annotation project and outputs a number of MMAX-2 format *markable* XML files, including files for Part-of-Speech tags, lemmas, and parses. The pronominal and nominal mentions, together with their morphological information are stored in a single top-level MMAX-2 format markable file.

3.1.5 Stanford CoreNLP

Stanford CoreNLP is a suite of Natural Language Processing tools with support for a number of languages, including English. The dependency parser is used in the English pre-processing pipeline for the annotation of the ParCor corpus (see Section 4.6.1). The dependency parser and coreference resolution (Lee et al.,

2011) components are used to extract information about pronouns in the English source texts as part of the automated pronoun post-editing process described in Chapter 7. The coreference resolution component uses a sieve-based approach, described in Section 2.3.2.

3.1.6 NADA

NADA (Bergsma and Yarowsky, 2011) is a tool for identifying non-anaphoric instances of the English pronoun "it". For each instance of "it" in the input text, the tool assigns a probability which may be thresholded to identify anaphoric vs. non-anaphoric instances. It is used in the English pre-processing pipeline for the annotation of the ParCor corpus (see Section 4.6.1), and in the automated post-editing experiment described in Chapter 7.

3.1.7 Moses

The Moses[2] toolkit (Koehn et al., 2007) provides a complete solution for the automatic training of SMT systems for any language pair given a corpus of parallel texts. The baseline SMT system used in the post-editing experiments described in Chapter 7 was created using the Moses toolkit.

3.1.8 Pronoun Selection Task Evaluation Tool

The pronoun selection task described in Hardmeier (2014) was conducted using a web-based evaluation tool (cf. Section 2.5.1.2). A customised version of this tool (i.e. with different pronoun options) was developed for the pronoun selection tasks described in Chapter 6.

3.2 Data

3.2.1 EU Bookshop Documents

The *EU Bookshop*[3] is an online repository providing a range of documents on topics connected with the EU's activities and policies. The majority of the documents are produced by EU institutions including the European Commission,

[2]Moses: http://www.statmt.org/moses/
[3]EU Bookshop: https://bookshop.europa.eu/

European Parliament, Council of the EU, EU agencies, and other bodies. At the time of writing, the EU Bookshop contains 100,000 titles in a variety of electronic formats including PDFs, E-books, CDs and DVDs in more than 50 different languages, including the 24 official languages of the EU.

The documents are intended for a wide audience with the majority aimed at non-experts. However, with the exception of materials aimed at children, most documents are written in a fairly formal style. Many of the documents are available in multiple languages, with translations provided by professional translation companies.

3.2.2 TED Talks

TED Talks[4] are lectures delivered at TED conferences to a live public audience. They are also recorded for online viewing by other members of the public around the world, and made available via the TED Talks website. The lectures address a wide range of topics and often have a strong persuasive style aimed at changing the beliefs or behaviour of the audience. Past presenters have included politicians, technology experts and Nobel prize winners.

The lectures are planned before being delivered to the public audience. As such, they are different from other transcriptions of spoken text, which would typically contain a greater number of dis-fluencies. The lectures are presented online in the form of videos and accompanying transcriptions. As part of the TED Open Translation Project, TED Talks are translated by volunteers across the world in order to provide subtitles for the hearing-impaired and those who do not speak English. There is only one translation of a given TED Talk per language.

3.2.3 TEDx Talks

TEDx Talks[5] are independently organised events, run under a free licence granted by TED. A TEDx event either involves a screening of an existing TED Talk, or a live presentation in the style of a TED Talk. In addition to *Standard* events (i.e. typical TED Talk format), other styles of TEDx Talk exist, including: *University* events (hosted at academic institutions), *Youth* events (organised by or catering

[4]TED: http://www.ted.com/
[5]TEDx: http://tedxtalks.ted.com/

to youths/children) and *Internal* events (hosted by companies/organisations as private events).

Translations, again by volunteers, are provided for some TEDx Talks but their availability is not as widespread as it is for TED Talks.

The TEDx Talks used in this thesis were originally delivered in German before being translated into English. They complement the TED Talks which were all originally delivered in English.

3.2.4 English-German Bilingual Dictionary

The large English-German dictionary used by the LFAligner to sentence-align the ParCor 1.0 corpus (Guillou et al., 2014) texts was created using bilingual n-gram pairs extracted from dict.cc[6], an online multilingual dictionary. The English-German and German-English dictionaries were constructed from the same dict.cc database extract file. This file was cleaned to remove information that was not part of the n-gram text (morphological information, abbreviations in parentheses etc.) and to split many-to-one entries into one-to-one entries. The resulting English-German dictionary contains 982,968 n-gram pairs and the German-English dictionary contains 993,535 n-gram pairs.

3.2.5 Dictionary of French Nouns

The dictionary of French nouns was constructed from an extract of the Lefff (Sagot, 2010) and augmented with entries from the bilingual English-French dictionary downloaded from dict.cc. The dictionary contains a list of 86,442 French nouns and their number and gender. It is used in the automated post-editing system described in Chapter 7.

[6]dict.cc online multilingual dictionary: www.dict.cc

Chapter 4

Construction of the ParCor Corpus

This chapter describes the construction of the ParCor corpus (Guillou et al., 2014), a collection of parallel texts in which pronouns are manually labelled according to their function. The aim of the corpus is two-fold. Firstly, it is intended to be used as a resource from which to learn systematic differences in pronoun use between two languages, thereby informing the design of SMT systems. Secondly, it is intended to be used as a gold-standard test set for use in developing and testing SMT systems.

The contributions of this chapter are the ParCor corpus itself, and the guidelines for the manual annotation of pronouns which were used to annotate the corpus. These same guidelines have also been used to annotate the English source texts of the *DiscoMT2015.test* dataset, described later in Section 7.9. The guidelines are intended to provide accurate annotation and be as language independent as possible. The high inter-annotator agreement scores in Section 4.7 suggest that the manual annotation task may be completed to a high degree of accuracy, once the annotator is familiar with the guidelines and the task. This chapter describes the application of the guidelines to both English and German texts, but they would also be suitable for the annotation of other languages, including French. In the case of some languages, especially those that exhibit subject pro-drop, extensions to the guidelines would be required, but the core guidelines outlined in this chapter should cover the cases that are common to different languages.

The content of this chapter extends Guillou et al. (2014). Extensions and additions include additional manual annotation and automatic annotation of unambiguous pronouns. Two German TEDx talks and their English translations were annotated, addressing the gap introduced because all texts in the ParCor

corpus are documents/transcripts in English. The unambiguous first-person singular pronouns have been automatically annotated for the English texts. So too has "man" (equivalent to generic "one" in English) for the German texts. The ParCor corpus was constructed as part of a collaborative project between the University of Edinburgh and the University of Uppsala. In addition to discussions regarding the annotation guidelines, which involved all co-authors, the following specific contributions are credited to my co-authors from Uppsala:

- Christian Hardmeier and Aaron Smith were responsible for the adaptation of the annotation guidelines for the TED Talks corpus, and the extraction of relevant illustrative examples. The annotation guidelines were originally defined for the EU Bookshop corpus, but differences between the two genres necessitated the addition of a number of extra guidelines and clarifications.
- Jörg Tiedemann arranged for the release of the corpus on the OPUS website.

4.1 Overview

Due to the lack of available parallel corpora in which pronominal coreference is annotated (cf. Section 2.6.4), it was necessary to construct one to perform the analyses presented later in this thesis. Work began with the development of annotation guidelines adapted from the MUC-7 Coreference Task Definition (Chinchor and Hirschman, 1998). These guidelines were used in the annotation of parallel English-German documents from the EU Bookshop online archive. The Europarl corpus, commonly used in SMT research was rejected as it does not contain document boundaries, and can contain multi-speaker sessions. Later, a number of parallel English-German TED Talks were added to the corpus as part of a collaborative effort with colleagues from the University of Uppsala. These two corpora were combined and released as the ParCor 1.0 corpus (Guillou et al., 2014).

Texts from these two corpora represent two very distinct genres. The formal written style of the EU Bookshop documents is complemented by the less formal (planned) spoken style of the TED Talks. These genres are described in more detail in Sections 3.2.1 and 3.2.2.

Following the initial release of ParCor 1.0, two German TEDx Talks and their English translations were annotated. The TEDx Talks complement the

collection of TED Talks and EU Bookshop documents, which were originally written in English and translated into German. In order to make claims about the two languages, independent of translation direction, it was necessary to know what effect (if any) translation direction had on pronoun use. TEDx Talks were chosen in preference to EU Bookshop documents as the original language is easily identifiable[1]. The TEDx Talks genre is described in more detail in Section 3.2.3.

To support the aims of the corpus as both a parallel resource and gold-standard test set for SMT, all pronouns in each text within the corpus (personal, possessive, demonstrative, relative, adverbial and generic) are marked according to their functional type (see Section 1.4). These functional types closely match those presented in work on functional grammar in the field of linguistics (see Section 2.1.1). The reason for categorising pronouns according to their function is, as noted earlier (in Section 1.1.2), that different types of pronouns function differently and may behave differently under translation. Previous work in SMT has focussed on anaphoric pronouns, with little attention paid to other types of pronouns or to the same pronouns used for other functions, and hence not serving as anaphors (i.e. *distractors*[2]). This corpus provides a starting point for work with a focus on handling pronouns differently, according to their *functional type*.

Pronouns occurring in the English and German texts were annotated using the MMAX-2 annotation tool (Müller and Strube, 2006). Following a pre-processing step in which pronouns and NPs (representing the potential set of pronoun antecedents) were identified by an automated pipeline, human annotators manually completed the annotation. The MMAX-2 tool is described in Section 3.1.1, the annotation scheme in Section 4.4 and the annotation process in Section 4.6. For each language, two human annotators worked in parallel until inter-annotator agreement was deemed to have reached an acceptable level. This is described in more detail in Section 4.7.

[1]The EU Bookshop does not make this information available online and tracing it requires some effort.

[2]Here, "distractors" refers to the set of pronouns which are not anaphoric but share the same surface form as anaphoric pronouns. For example "it" may be used as an event reference, pleonastic or anaphoric pronoun.

ID	Title	Tokens		Parallel
		English	German	Sentences
KEBC11002	Social Dialogue	32,000	31,572	1,391
KEBC12001	Demography, Active Ageing and Pensions	24,370	23,684	1,121
KH7911105	Soil	6,644	6,429	301
MI3112464	Road Transport	5,609	5,428	288
MJ3011331	Energy	10,854	10,853	471
NA3211776	Europe in 12 Lessons	23,311	21,761	1,191
QE3011322	Shaping Europe	11,005	10,819	485
QE3211790	Active citizenship	22,368	23,071	1,168
Total		136,161	133,617	6,416

Table 4.1: Documents taken from the EU Bookshop online archive

4.2 Data

4.2.1 EU Bookshop

The documents that make up the EU Bookshop corpus were taken from the EU Bookshop online repository described in Section 3.2.1. They represent the set of English-German parallel documents deemed to have *true* translations in German[3], that were available in E-book format at the time of data collection (4th February 2013). E-books were selected in preference to PDFs due to the ease with which the content of the documents may be extracted. The Calibre E-book management tool[4] was used to extract raw text from the E-books. The EU Bookshop documents are detailed in Table 4.1.

4.2.2 TED Talks

The English documents selected for the TED Talks corpus are taken from the English-French 2010 test set of the IWSLT 2013 dataset. The corresponding German translations were extracted from the English-German XML format files dated January 2012 as the English-German pairing for the 2010 test set contained

[3]As opposed to pairs of documents with the same title that contain different information specific to the English and German speaking audiences.

[4]Calibre E-book management tool: http://calibre-ebook.com/

ID	Title	Tokens		Parallel
		English	German	Sentences
767	Bill Gates on Energy: Innovating to Zero!	5,371	4,775	259
769	Aimee Mullins: The Opportunity of Adversity	3,414	3,430	143
779	Daniel Kahneman: The Riddle of Experience vs. Memory	3,564	3,566	181
783	Gary Flake: Is Pivot a Turning Point for Web Exploration?	1,280	1,163	65
785	James Cameron: Before Avatar ... a Curious Boy	3,265	3,054	172
790	Dan Barber: How I Fell in Love With a Fish	2,988	2,921	214
792	Eric Mead: The Magic of the Placebo	1,788	1,768	112
799	Jane McGonigal: Gaming Can Make a Better World	4,354	3,947	251
805	Robert Gupta: Music is Medicine, Music is Sanity	1,002	989	43
824	Michael Specter: The Danger of Science Denial	3,644	3,531	255
837	Tom Wujec: Build a Tower, Build a Team	1,301	1,161	81
Total		31,971	30,305	1,776

Table 4.2: Documents taken from the TED Talks in the IWSLT2013 2010 test set

different English transcriptions from the English-French pairing. The TED Talks are detailed in Table 4.2.

4.2.3 TEDx Talks

In order to provide evidence of pronoun use in German original texts and English translations, German TEDx Talks (See Section 3.2.3) and their English translations were extracted from the TEDx Talks repository on YouTube[5]. This complements the TED Talks in the corpus, for which there are English original texts and German translations.

Whilst there are many TEDx Talks originally recorded in German, only three talks had published translations into English at the time of data extraction. Of these, one of the talks has two speakers. As none of the TED Talks in the corpus follow a multi-speaker style and because different speakers will use pronouns differently, this talk was excluded from the corpus. This leaves two TEDx Talks, detailed in Table 4.3.

The original German and English translation subtitles were extracted in SRT (SubRip Text) format from the relevant YouTube videos (on 18th June 2014),

[5]YouTube: https://www.youtube.com

ID	Title	Tokens		Parallel
		English	German	Sentences
Wheelmap	Raul Krauthausen: Wheelmap.org (TEDxBerlin)	1,335	1,254	92
Komforzone	Prof. Dr. Gunter Dueck: Komfortzone Zukunft oder Wider die Gewöhnung (TEDxRheinMain)	4,043	3,908	277
Total		5,378	5,162	369

Table 4.3: Documents taken from the TEDx Talks available on YouTube

using the KeepSubs online service[6]. The texts were automatically cleaned and then manually checked with respect to sentence segmentation.

4.3 Annotators

Human annotators were employed to annotate the texts in the ParCor corpus (Guillou et al., 2014); two German annotators and three English annotators. All annotators are native speakers of the languages that they were asked to work with. The profiles of the annotators are displayed in Table 4.4. The values in the last column denote whether the annotator acted as the *primary* annotator (annotating all texts) or the *secondary* annotator, providing annotations used to calculate inter-annotator agreement. The annotations provided by the primary annotator are those used in the following chapters.

4.4 Annotation Scheme

The annotation scheme was adapted from the pronoun annotation guidelines in the MUC-7 Coreference Task Definition (Chinchor and Hirschman, 1998). The schemes for the EU Bookshop and TED Talks corpora are similar, as the aim is to provide comparable annotation for both corpora. However, there are a number of genre and language specific differences, which are highlighted in the sections below.

[6]KeepSubs: http://keepsubs.com/

ID	Native Language	Annotated	Primary/Secondary
EN1	English	TED	Primary
EN2	English	TED	Secondary
		EU Bookshop	Primary
		TEDx	Primary
EN3	English	EU Bookshop	Secondary
DE1	German	TED	Primary
		EU Bookshop	Primary
		TEDx	Primary
DE2	German	TED	Primary

Table 4.4: Annotator profiles

Central to the annotation scheme is that pronouns are marked as belonging to one of eight functional *types*: Anaphoric/cataphoric reference, event reference, extra-textual reference, pleonastic, addressee reference, speaker reference, generic reference, or other function (see Section 4.4.8). In addition, the annotation scheme defines the attributes which should be labelled for each *type* of pronoun and whether it should be linked to an antecedent. (Note that in the corpus annotations functional types are recorded under the *type* attribute.) The scheme forms the basis of the set of annotation guidelines given to the human annotators for the manual annotation phase. (A copy of the guidelines given to the annotators is included in Appendix A.) The annotation scheme is summarised below. As pronoun type defines the attributes that are recorded for each pronoun, the annotation scheme will be separately described for each pronoun type. Language and genre-specific differences are highlighted for each pronoun type, within the relevant section.

4.4.1 Anaphoric/Cataphoric

The anaphoric and cataphoric pronouns of interest are those which co-refer with an antecedent. An anaphoric pronoun occurs in a text *after* its antecedent, as in Ex. 4.1. A cataphoric pronoun occurs in text *before* its antecedent, as in Ex. 4.2.

(4.1) If **Mary** is in town, **she** can join us for dinner [anaphoric]

(4.2) If **she** is in town, **Mary** can join us for dinner [cataphoric]

The annotation scheme treats anaphoric and cataphoric pronouns as the same. For the remainder of this section, the term *anaphoric* will be used to cover both anaphoric and cataphoric pronouns.

For pronouns marked as anaphoric, the following attributes are also required:

- *Agreement*
- *Case/Position*
- *Antecedent(s)*

Agreement is used to provide additional information which can be used to disambiguate ambiguous pronouns. For example, the German anaphoric pronoun "sie" may co-refer with either a singular ("she/it") or plural ("they") antecedent. An SMT system, when faced with an instance of "sie" may not be able to decide on the correct translation based on the local context alone and may therefore produce an incorrect translation: For example, "they" when "she" or "it" would be correct. In the case of "sie" (i.e. anaphoric) two options, "singular" and "plural", are provided for the "agreement" attribute which is used to disambiguate between different uses of the same pronoun. The German pronoun "Sie" is used as the formal expression of the second-person personal pronoun (i.e. "you") and should be labelled as an addressee reference pronoun. Sentence-initial instances of "Sie" may be disambiguated using the pronoun type label.

Ambiguity is not just limited to "Sie/sie". It exists for other German pronouns including "die", "ihr", etc. It also exists for the English pronoun "they", which is typically used in the plural form, but is also increasingly used in the singular form to refer to a person of unspecified gender in place of gender-neutral "he" (Leech et al., 2009).

Position (subject or non-subject), in English, and *case*, in German, are used to identify the syntactic role of the pronoun in a sentence.

Antecedent(s): Each anaphoric pronoun is linked to its *nearest non-pronominal antecedent*. The NPs identified during automated pre-processing (cf. Section 4.6.1) form the set of candidate antecedents to which an anaphoric pronoun may be

linked. The annotators were instructed to select antecedents from this set wherever possible. If no suitable NP markable exists, the next closest automatically generated "NP" markable span should be amended so that it covers the necessary NP text. If no close match exists, a new NP markable may be created. When amending an NP markable or adding a new one, the following rules apply:

- The NP markable must contain the *head* noun
- If the head of the NP is a name, the entire name should be included in the NP markable. For example, given the name "Frederick F. Fernwhistle Jr.", it is not sufficient to simply mark "Frederick"
- The NP markable should include all text which may be considered a modifier of the NP. For example, "the big black dog" (where "dog" is the head) contains the modifiers "big" and "black"
- Determiners should be included for definite NPs. For example, "the big black dog" would be marked, not just "big black dog"

The guidelines for marking NP spans are taken from the MUC-7 guidelines (Chinchor and Hirschman, 1998) and mirror those used in the Tüba-D/Z corpus (Naumann and Möller, 2007).

An anaphoric pronoun will typically have a single NP as its antecedent. However, in those cases where a pronoun refers to multiple NPs (called *split reference*), it should be linked to each NP. Conjoined NPs such as "John and Mary" are treated as a single NP and may be marked as the single antecedent span of a plural pronoun such as "they". In the case where there is intervening text between NPs such that a single NP does not cover all of the elements of the antecedent, each sub-part should be marked individually.

(4.3) **John** likes documentaries. **Mary** likes films about animals. The last time **they** went to the cinema **they** compromised and saw a film about penguins.

In Ex. 4.3, "John" and "Mary" should be marked as NPs and both linked as antecedents for the instances of the pronoun "they". In this instance, both NPs can be recovered as antecedents of the pronoun from the MMAX-2 coreference markables XML file, via the *coreference set* of the pronoun.

Reflexive pronouns are rare in the TED Talks, and even rarer in the EU Bookshop documents. They are therefore treated as a *special case* (see Section 4.4.1.3). Third-person reflexive pronouns in TED Talks are annotated in the same way as all other third-person personal pronouns. In the case of the EU Bookshop documents, three reflexive pronouns were annotated in the German texts (none in the English texts), although the guidelines do not specify that the annotators should have done so.

4.4.1.1 EU Bookshop

Anaphoric pronouns may have no explicit antecedent in the text. For example the "they" in Ex. 4.4 has no explicit antecedent in the text, but one could infer from the context that "they" refers to the authors of the study and is therefore anaphoric. This is in contrast with the extra-textual reference type, for which the antecedent cannot be inferred from context.

(4.4) In this study **they** took 100 people and split them into two groups

In this case, the pronoun is marked not as *anaphoric* but as *anaphoric but no explicit antecedent*. *Split reference* pronouns (i.e. those that refer to multiple antecedents as in Ex. 4.3) are identified by counting the number of NPs to which the pronoun is linked.

4.4.1.2 TED Talks

Anaphoric pronouns are sub-classified, using the *Split* attribute, as:

- *Simple antecedent*: Referring to a single NP present in text (default value)
- *Split reference*: Referring to multiple NPs (and linked to each NP)
- *No explicit antecedent*: Pronoun is clearly anaphoric but its antecedent is not explicitly mentioned in the text

The addition of marking whether the antecedent is a *simple antecedent* or an instance of *split reference* simply acts as a clarification of the number of antecedents to which a pronoun is linked. The requirement to mark split reference is not a part of the MUC-7 guidelines but neither does it constitute a change to those guidelines.

4.4.1.3 Special Cases

The following rules were added to the guidelines to cover specific cases and go beyond what is specified in the MUC-7 guidelines:

- When the pronoun "they" (and its equivalents in German) is used to refer to a collective noun (e.g. "the government"), it should be treated as a plural pronoun
- The use of singular "they" in English is not very common in the ParCor texts. Instances such as "he or she" and "s/he" also exist in place of un-gendered pronouns. These should be treated as a single pronoun
- If a pronoun refers to a modifier in an NP, the pronoun should be linked to the modifier if no other suitable antecedent can be found
- TED only: Third-person reflexive pronouns such as "himself/itself" are labelled in the same way as any other third-person pronoun. In cases like "Here comes the man **himself**", the token "himself" is not considered a pronoun, and if marked as a pronoun by the automated pre-processing pipeline, should be corrected by the annotator
- EU Bookshop only: Pronominal adverbs (e.g. "therefore", "herewith" etc.) are marked. These may be marked as *anaphoric* or *event* depending on their function within the text

4.4.2 Event Reference

The *event reference* category is used for pronouns that refer to propositions, facts, states, situations, opinions, etc. It is used in examples like Ex. 4.5, where the pronoun "this" refers to the action of John arriving late (a verb phrase), and not an NP.

(4.5) John arrived late; **this** annoyed Mary

Due to the differences between the TED Talks and EU Bookshop genres, event reference pronouns may be used in other ways, beyond this simple case. An event reference pronoun may be used to refer back to a whole sentence or section of text, or a concept evoked by the text. For example, the speaker might

say "**This** got me thinking", where "this" refers to a story they just told. In this case "this" should be labelled as event. Event reference pronouns may refer to concrete events, or to hypothetical events as in "...spot prices could decrease and remain low... **This**...", where "This" refers to the hypothetical scenario in which spot prices decrease and remain low.

Event reference pronouns are simply labelled as *event*. From the perspective of using the corpus as a gold-standard test set for SMT, no other information is required. In English-to-German translation, when they co-refer with events, "it", "this" and "that" are typically translated as "es", "dies" and "das" respectively. Unlike anaphoric pronouns which refer to nouns/NPs and for which number and gender agreement must hold in some languages (e.g. German or French), event reference pronouns get their antecedents from verbs, verb phrases, clauses, sentences, and larger spans of text. Agreement with the main verb is not required. By identifying instances of the English pronouns "it", "this" and "that" as event reference pronouns, these can be removed from the set of possible distractors when considering the translation of anaphoric pronouns with the same surface form. In translating between English and German (either direction), event reference pronouns likely pose little challenge, provided their *type* is known. This is because there are few event reference pronouns in German ("dies" and "das") and in English ("this", "that" and "it"). Differences in the use of the event reference pronouns is subtle, and the pronouns are exchangeable to some extent. For example in English we might hear a speaker say any of the following: "**This/That/It** got me thinking".

In monolingual coreference-annotated corpora, event reference is often ignored and it is probably for this reason that event reference pronoun resolution is excluded by coreference resolution systems. Identifying events poses a unique challenge, different to the standard task of identifying anaphoric pronouns and their referents. Also, as Pradhan et al. (2011) state, instances of event reference pronouns tend to be relatively rare when compared to the number of anaphoric pronouns marked in the data typically used to train coreference resolution systems.

4.4.2.1 TED Talks

Two event reference pronouns that refer to the same event are linked together.

4.4.2.2 EU Bookshop

Unlike in the TED Talks corpus, event reference pronouns are not linked together, although in principe they could have been. Pronoun density is much lower in the EU Bookshop documents than in the TED Talks (see Section 4.9). As a result cases in which two event pronouns could be linked together do not occur. This was confirmed via manual examination of the texts.

4.4.3 Extra-textual Reference

The *extra-textual reference* category is used for pronouns whose reference is fixed through the context of the utterance. It was first introduced by Halliday and Hasan (1976) as *exophoric reference*. The extra-textual reference category is used only for *deictic* pronouns. It is commonly used when the speaker refers to items physically present in the room, such as slides, which do not form part of the TED Talk transcription. For example, the speaker might say "The house looked like **this**" whilst pointing at a photo that a member of the audience can see. As a special case, this category may also be used within quoted text when referring to a third-person, e.g. the "He" in Ex. 4.6. This scenario is different to the annotation of speaker reference, which is used only for first-person personal pronouns.

(4.6) People when they see me say, '**He**'s a nice guy'

This category is used only in the TED Talks corpus.

4.4.4 Pleonastic

The *pleonastic* category is used for pronouns that are required by syntax but have no semantic content. They are also commonly call *dummy* or *expletive* pronouns. Pleonastic pronouns are found in both English and German. For example the "It" in "**It** is raining", and "Es" in the equivalent German phrase "**Es** regnet" are both examples of pleonastic pronouns.

As with event reference pronouns, pleonastic pronouns are typically not marked in monolingual coreference-annotated corpora. There is no provision for the handling of pleonastic pronouns in the MUC-7 guidelines and they are not marked

in the OntoNotes (Weischedel et al., 2011) corpora or the BBN Pronoun Coreference and Entity Type corpus (Weischedel and Brunstein, 2005). Pleonastics, are however, marked in the Tüba-D/Z corpus (Naumann and Möller, 2007).

The identification and exclusion of instances of pleonastic "it" has been implemented in a number of coreference resolution systems including the sieve-based Stanford Coreference Resolution System (Lee et al., 2011). Other tools and methods have been developed for the specific purpose of identifying non-referential "it" (Bergsma and Yarowsky, 2011; Boyd et al., 2005). By marking pleonastic pronouns, as with marking event reference pronouns, these may be removed from the set of distractors when considering the automated translation of anaphoric instances of the pronoun "it".

4.4.5 Addressee Reference

Addressee reference pronouns are used to refer to the person being addressed. They include second-person personal pronouns of formal and informal address.

4.4.5.1 EU Bookshop

In the EU Bookshop corpus the addressee reference category is typically used to to label pronouns that refer to the reader of the document. For example, the second-person pronouns "you" and "your" (and their German equivalents) may be used to refer to the reader. Addressee reference pronouns are simply labelled as such, with no additional attributes recorded.

4.4.5.2 TED Talks

In the TED Talks, the addressee is either another person on stage with the speaker, a specific member of the audience or the audience in general. The speaker may also use addressee reference pronouns in imaginary or recounted dialogues — dialogues placed in narrative space, with narrative agents (cf. Section 4.4.10). Second-person pronouns are always labelled as addressee reference. They are then sub-classified as generic or deictic using the *audience* attribute:

- *none* (default value, indicating no annotation has been provided)
- *Deictic "you"*: The speaker refers to the audience or a specific person

- *Generic "you"*: The speaker uses the pronoun generically, e.g. "In England, if **you** own a house **you** have to pay taxes"

N.B. Although the attribute labels include a reference to "you", this includes all addressee reference pronouns in English and their German equivalents. For example, it covers the closed set of English pronouns: "you", "your" and "yours".

Agreement is also recorded for instances of "you", so as to distinguish singular and plural use. In the case of the EU Bookshop documents, no distinction is made between singular and plural "you" as the reader will presumably, always be the addressee (i.e. singular "you"). That is, the audience distinction in TED Talks where the speaker could be addressing a specific person or the audience as a whole, does not apply in the case of the EU Bookshop documents.

When a speaker uses a deictic instance of "you", addressing a whole audience, that instance is always marked as plural, even in cases such as Ex. 4.7.

(4.7) Imagine **you**'re walking alone in the woods

4.4.6 Speaker Reference

Speaker reference pronouns are used to refer to the speaker / writer. For example,"I" and "my" in Ex. 4.8 are both examples of speaker reference pronouns.

(4.8) **I** decided that releasing fireflies would be **my** contribution to the environment here this year.

4.4.6.1 EU Bookshop

In the EU Bookshop corpus the speaker reference category is used to label pronouns that refer to the speaker (first-person pronouns). This includes singular pronouns which exclude the addressee and plural pronouns ("we", "us" and "our", and their German equivalents) which may also include the addressee. In these texts singular first-person pronouns are rare, but are marked when they do occur.

4.4.6.2 TED Talks

In the TED Talks corpus singular first-person pronouns do not require manual annotation, with the exception of those in quoted text (Section 4.4.10). They are automatically identified and labelled as speaker reference as a post-annotation

step (see Section 4.8). Plural first-person pronouns require manual annotation as they also need to be sub-classified as exclusive, co-present or all-inclusive using the *audience* attribute:

- *none* (default value, indicating no annotation has been provided)
- *Exclusive "we"*: Includes the speaker and his/her group, but not the audience
- *Co-present "we"*: Includes the speaker and everyone physically present in the same room
- *All-inclusive "we"*: Incorporates everything else

N.B. Although the attribute values only include a reference to "we", they apply to all plural speaker reference pronouns in English and their German equivalents.

4.4.7 Generic

The *generic* category is used for pronouns that refer to an unspecified person. For example, the instances of "you" in Ex. 4.9 and "man" in Ex. 4.10 do not refer to a specific person, but to people in general.

(4.9) In England, if **you** own a house **you** have to pay taxes.

(4.10) In England muss **man**, wenn **man** ein Haus besitzt, Steuern bezahlen.

In English, "you" and "one" may be used as generic pronouns. German has only one generic pronoun, "man". The token "man" does not serve any function other than as a generic pronoun.

With respect to other coreference annotation efforts, generic pronouns are not marked in the OntoNotes (Weischedel et al., 2011) or BBN Pronoun Coreference and Entity Type (Weischedel and Brunstein, 2005) corpora and there is no provision for the annotation of generic pronouns in the MUC-7 guidelines (Chinchor and Hirschman, 1998). However, that is not to say that generic pronouns are always ignored. The generic German pronoun "man" is labelled in the Tüba-D/Z corpus (Naumann and Möller, 2007), however, it is labelled as *indefinite* and not *generic*. The *indefinite* category is rather broad, but the *generic* category

provides a narrower classification that allows for the distinction to be made between generic and referential instances of the ambiguous English pronoun "you". When translating in the English-to-German direction, this will be useful for distinguishing which instances of "you" should be translated as "Sie/du" (addressee reference) and which should be translated as "man" (generic).

The *generic* category is only used in the EU Bookshop corpus. In the TED Talks corpus, the *audience* attribute captures whether an addressee reference pronoun is generic.

4.4.8 Other Function

The default *pronoun* label is used for words that are clearly pronouns but do not belong to any of the above categories. This includes indefinite pronouns (e.g. "anyone") and some numbers/quantifiers that are used as pronouns but are not themselves bare pronouns (e.g. "others", "each", "both"). Instances such as these are labelled as *pronoun*, with no additional features recorded.

An example of a pronoun belonging to the *other function* category is "others" in Ex. 4.11, referring to people other than the conference participants.

(4.11) During the conference, participants and **others** following the debate through social networking sites worked to put together a roadmap, charting future action.

Indefinite pronouns are also marked in the Tüba-D/Z corpus (Naumann and Möller, 2007), but do not form part of OntoNotes (Weischedel et al., 2011) or BBN Pronoun Coreference and Entity Type corpus (Weischedel and Brunstein, 2005) annotations or the MUC7-guidelines (Chinchor and Hirschman, 1998).

This category is only used in the annotation of the EU Bookshop corpus.

4.4.9 Dealing with Functional Ambiguity of Pronouns

Decisions regarding the labelling of pronoun types are not always straightforward and there may be multiple possible ways to read/interpret a sentence or section of text. In particular the primary annotator of the English EU Bookshop texts identified a number of instances for which both event reference or anaphoric readings are possible and would make sense. (See Section 4.7 for examples.) In these cases, the pronoun is marked as anaphoric. In other cases, other ambiguities

arose, or it was impossible to determine the pronoun type. In these cases, the pronoun was labelled as *unsure* (originally "help! not sure"). Such problems were not raised for the German translations or in the annotation of the TED Talks. One possible explanation for this is that in some cases the translators, for better or worse, consciously or unconsciously, attempted to disambiguate those instances that were ambiguous in the original text. However, confirming this and quantifying the extent to which disambiguation occurred during translation would be an extremely difficult task. Any further exploration of this topic is deemed to be outside of the scope of this work.

4.4.10 Dealing with Pronouns in Quoted Text

Annotating pronouns in direct quotes is more complex than when they occur outside of quotes. The use of direct quotes is infrequent in the corpus texts and to simplify their annotation, the following guidelines were used for first and second-person pronouns. (Third-person pronouns were marked as normal in accordance with the annotation guidelines in the relevant sections.)

4.4.10.1 EU Bookshop

All first and second-person pronouns were labelled as *pronoun* to indicate that they have been seen by the annotator. The surface form of such pronouns is sufficient to determine whether the pronoun is an instance of speaker or addressee reference. In some cases, the text may have the form of an interview (with question and answer sections), with quote marks absent from the text. In this scenario, the text *was not* treated as quoted text; instead speaker/addressee reference pronouns were annotated as normal.

4.4.10.2 TED Talks

It is common in TED Talks for a speaker to illustrate a point by recounting a conversation that they had with someone in the past. Such conversations appear in the transcriptions and translations as quoted text. In this scenario, pronouns in quoted text are annotated from the point-of-view of the quoted speaker, rather than that of the (TED) speaker who quotes the utterance.

First-person pronouns are always labelled as instances of *speaker reference* and second-person pronouns as instances of *addressee reference*. Coreference re-

lations between a first-person or second-person pronoun inside quoted speech and a pronoun outside the quoted speech passage are not marked. For example in "**He** said, '**I** do."', where the pronouns "He" and "I" could arguably be marked as coreferent, these pronouns would not be linked. The focus of the annotation guidelines is the annotation of pronouns and their nearest non-pronominal antecedent.

4.4.11 Exclusions

In designing ParCor, a number of elements that are commonly included in coreference annotated corpora have been excluded. Reflecting the interest in pronoun translation full coreference chains/sets are not provided. The benefit of using complete coreference sets in pronoun translation is not known and falls outside of the scope of the work presented in this thesis.

Apposition is not annotated for NPs. That is, where an NP represents an appositive, we do not further annotate the head and the attribute of the span. The annotation of appositives is commonly included in annotation guidelines and annotated corpora such as the MUC-7 guidelines and OntoNotes. However, the separation of appositives into head and attribute components is not necessarily useful for SMT where head finding techniques will be required for all antecedents of coreferential pronouns (not just appositives) to ensure that agreement holds between the pronoun and head noun.

Implicit pronouns are not annotated – that is, the annotation scheme for the ParCor corpus follows the MUC-7 guidelines in assuming that English has no zero pronouns. This assumption is also extended to German. In practice, this means that the empty string is not considered to be a markable in MMAX-2.

4.5 MMAX-2 Projects

MMAX-2 projects consist of a collection of XML-format files listing the tokens for each text and describing the annotation layers (i.e. markables), the annotation scheme for each annotation layer, and the visual appearance of the text and its annotations (customisations and styles).

Each project has two annotation layers: The *coreference* annotation layer in which pronouns, NPs and the links between them are labelled, and the *sentence*

layer which defines sentence boundaries. The EU Bookshop documents are typically much longer than the TED Talks texts and so an additional *paragraph* layer is provided to insert paragraph breaks, making the text easier to read.

The manual annotations made according to the ParCor annotation guidelines (see Section 4.4) are applied to the coreference annotation layer. The human annotator is guided in annotating the texts by the *schemes* defined in MMAX-2. The coreference scheme file defines a filtered list of annotation fields representing the features required for each pronoun type. It is used by the annotation interface to display the list of pronoun types and to ensure that all relevant information is captured for the selected pronoun type. As annotations over the sentence and paragraph layers are not required, schemes are not provided for these layers.

Customisations ensure that pronouns stand out from the remainder of the text. Pronouns are displayed in bold text. Automatically identified pronouns and manually created markables initially appear with coloured highlights. These disappear when their (pronoun) *type* or *mention type* is set, signalling that annotation of these elements is complete. *Styles* are used to apply *handles* to each markable – brackets define the span of each markable, helping the annotator to distinguish between nested or overlapping markables.

Prior to the creation of a new project, the text is split into sentences and tokenised. Both sentence splitting and tokenisation are achieved using scripts provided with the Moses toolkit (Koehn et al., 2007). The process by which the texts are annotated is described in the following section.

4.6 Annotation Process

The annotation process consists of two main phases. In the first phase, a preprocessing pipeline is used to automatically identify pronoun and NP markables and to generate the *coreference* annotation layer XML file containing these markables. In the second phase, a human annotator manually labels the pronouns that have been identified, according to the annotation scheme detailed in Section 4.4. These automated and manual annotation phases are described in Sections 4.6.1 and 4.6.2 respectively.

4.6.1 Automated Pre-processing

The aim of automated pre-processing is to identify all pronouns and NPs in the English and German texts and use this as a starting point for annotation, thereby reducing annotator effort and improving inter-annotator agreement. Those pronouns and NPs that are identified by the pipeline(s) are included as markables on the *coreference* annotation layer. Marking the pronouns automatically reduces the effort of the annotator in identifying and marking the pronouns manually, therefore allowing them to focus on the task of labelling the pronouns – something that is difficult (and in some cases, impossible) to do automatically. Marking the NPs automatically provides the annotator with a set of candidate antecedents for anaphoric and cataphoric pronouns. This not only reduces annotator effort in terms of manually marking the NPs, but also in terms of defining what the span of the NP should be. If the parse is of good quality we can be confident that the complete NP is provided as a markable. The provision of good quality spans should also have the added effect of improving consistency in antecedent choice between multiple annotators.

Separate pipelines were used for the English and German texts. Each relies on a parser as a central component and each outputs an MMAX-2 format XML file representing the coreference annotation layer in which pronouns and NPs are included as markables.

The English pipeline, used to process the EU Bookshop texts, starts by defining markables. It uses the Berkeley Parser (Petrov et al., 2006) to identify NPs and pronouns. NADA (Bergsma and Yarowsky, 2011) is then used to identify instances of pleonastic "it" (no equivalent system exists for German) and the Stanford Dependency Parser (de Marneffe et al., 2006) is used to identify whether instances of "it" are in subject or non-subject position. A number of other markables are identified using pre-defined lists. These include pronominal adverbs (e.g. "thereafter", "herein", etc.) and un-gendered third-person pronoun expressions (e.g. "he or she", "his or her(s)", "him or her" and "s/he") which are treated as a single pronoun under the annotation scheme (cf. Section 4.4.1.3). As pronouns used as speaker and addressee reference are unambiguous in English, pre-defined lists are later used to automatically identify and set the pronoun *type* for these pronouns (see Section 4.8). The English TED texts were processed and annotated by the team at the University of Uppsala. Their pipeline is not described here.

The German pipeline, used to process both the EU Bookshop and TED Talks texts, is described in full in Broscheit et al. (2010) and Versley et al. (2010). The German pipeline also starts by parsing the texts using the Berkeley Parser. It then extracts nominal (both minimal and maximal noun projections) and pronominal mentions from the parse trees. In the next step, a morphological tagger (described in Broscheit et al. (2010)) is used to provide number and gender information as well as the mention type (definite/indefinite noun phrase, name, personal/relative/reflexive pronoun). The pipeline outputs a MMAX-2 XML-format markables file containing pronoun and NP markables, from which a new markables file is constructed to match the format required for the ParCor annotation scheme. The section of the German pipeline used in this work does not include coreference resolution, rather the output of the pipeline could form the input to a coreference resolution system.

An alternative approach would have been to include state-of-the-art coreference resolution systems within the automated pre-processing pipelines. This approach is rejected due to the inaccuracies of even state-of-the-art coreference resolution systems. Manual annotation would still be required to correct the output and add missing pronouns, antecedents and the links between them. There is also the risk that the human annotators would be biased by the automated coreference link annotations. Lapshinova-Koltunski and Kunz (2014) also reject the use of coreference resolution systems in the automated annotation of their GECCo corpus (cf. Section 2.6.4).

4.6.2 Manual Annotation

There are limitations as to what can be annotated using the automated pre-processing pipelines. For example, at the time of corpus construction, no standard tools existed for the tasks of determining generic vs. deictic use of the (addressee reference) pronoun "you", for identifying audience involvement when second-person plural pronouns are used, or for identifying event reference pronouns. Furthermore, even those decisions that can be made by the automated pre-processing pipelines may be incorrect. An element of manual annotation is therefore necessary.

Given the output of the automated pre-processing pipeline, the annotator's task is simplified to one of labelling the auto-marked pronouns (or providing cor-

rections where automatically generated labels are incorrect) and linking anaphoric and cataphoric pronouns to their antecedent(s). In some cases, the annotator will also be required to create markables for those pronouns and/or NPs that were not identified by the automated pre-processing pipeline. However, the extent of this effort is greatly reduced when compared to the effort required to create the complete set of pronoun and NP markables from scratch.

The annotation guidelines used in the manual annotation phase are based on the annotation scheme defined in Section 4.4. The complete set of annotation guidelines is provided as part of the ParCor 1.0 release[7] and is included in Appendix A.

4.7 Inter-Annotator Agreement

Inter-annotator agreement (IAA) scores were calculated using Cohen's Kappa (Cohen, 1960). The purpose of collecting these scores was to ensure the quality of the annotation guidelines and the consistency of annotation of one or more documents by two annotators, before a single annotator completed the annotation of the remaining documents (in the relevant language / corpus).

Kappa scores were calculated (separately) for the following attributes:

- Pronoun functional *type*: All pronouns
- *Agreement*: Anaphoric pronouns
- *Position*: Anaphoric pronouns, English only
- *Case*: Anaphoric pronouns, German only
- *Audience*: Speaker/Addressee reference pronouns, TED Talks only

Kappa scores are computed for pronouns annotated by both annotators, and do not include those pronouns marked by only one annotator.

Since antecedents are spans, IAA considers both exact and partial matches between two annotations. A partial match is defined as one in which string A is a sub-string of string B (or vice versa). Agreement is measured in terms of the number of complete and partial antecedent matches.

[7] ParCor 1.0: http://opus.lingfil.uu.se/ParCor/

Category	**Pronouns**	**Disagree**	**Kappa**
ENGLISH: MJ3011331			
Type	138	13	0.85
Agreement	73	0	1.00
Position	73	5	0.82
Antecedent	73	13	N/A
GERMAN: MJ3011331			
Type	205	4	0.96
Agreement	136	4	0.96
Case	136	11	0.85
Antecedent	136	9	N/A
GERMAN: QE3011322			
Type	319	14	0.90
Agreement	224	8	0.95
Case	224	15	0.89
Antecedent	224	3	N/A

Table 4.5: IAA Scores for English and German EU Bookshop documents

As the annotation of the German EU Bookshop documents preceded that of the English documents, IAA was calculated for two German documents to assure the quality of the annotation guidelines. By the time annotation of the English documents began, the annotation guidelines had been fixed. IAA for the English side of the EU Bookshop corpus was therefore only computed for a single document. See Table 4.5 for English and German EU Bookshop IAA scores.

IAA scores for the TED Talks corpus are provided only for English, for the following reasons. Firstly, annotation of the TED Talks corpus followed that of the EU Bookshop corpus, hence the annotation scheme was largely stabilised with the exception of a few genre-specific changes. Secondly, the German annotator was already familiar with the annotation guidelines used in the EU Bookshop annotation. Computing IAA scores for two English TED Talks therefore serves to ensure that the changes to the annotation guidelines do not adversely affect the quality of the annotations. Furthermore, as the primary annotator of the English TED Talks had not previously worked on the English EU Bookshop annotation

Category	**Pronouns**	**Disagree**	**Kappa**
TED Talk: 785			
Type	191	27	0.81
Agreement	50	6	0.78
Position	50	1	0.97
Antecedent	50	5	N/A
Audience	99	13	0.82
TED Talk: 824			
Type	363	37	0.85
Agreement	133	6	0.90
Position	133	2	0.98
Antecedent	133	10	N/A
Audience	163	22	0.75

Table 4.6: IAA Scores for English TED Talks

task, calculating IAA for two English TED Talks ensured the suitability of the new annotator for the task. See Table 4.6 for English TED Talks IAA scores.

IAA was not measured for the TEDx Talks. The reason for this is two-fold. Firstly, the same annotation guidelines used in the annotation of the TED Talks corpus were used for the TEDx Talks, with no changes. Secondly, the two human annotators who annotated the TEDx Talks were already familiar with the annotation task. The German annotator had served as the primary German annotator for both TED Talks and the EU Bookshop documents. The English annotator had served as the primary English annotator for EU Bookshop documents and as the secondary annotator for the English TED Talks.

Whilst the patterns of disagreement are less clear for pronoun types in German and for other attributes in both languages, common disagreements in the annotation of English TED Talks arose from the labelling of instances of "it" as anaphoric vs. event reference and instances of "you" as generic vs. deictic.

In English TED Talk 785, 18 of the 27 disagreements for pronoun *type* are due to differences in opinion as to whether an instance of "it" is anaphoric or event reference. Consider the following examples:

(4.12) I said, "We're going to dive to **the wreck**. We're going to film **it** for real.

(4.13) I didn't really learn about leadership until I did **these expeditions**. Because I had to, at a certain point say, "What am I doing out here? Why am I doing this? What do I get out of **it**?

In Ex. 4.12 one annotator labelled the highlighted pronoun "it" as anaphoric referring to "the wreck". The other annotator labelled the pronoun as event reference, presumably referring to the event of diving to the wreck, although the events of event reference pronouns are not marked so we cannot be certain. A similar observation is made for Ex. 4.13 in one annotator labelled the highlighted pronoun "it" as anaphoric referring to "these expeditions", even though the use of a singular pronoun to refer to multiple expeditions contravenes pronoun-antecedent agreement. The other annotator labelled the pronoun as event reference and one might assume that "it" refers to the act of going on the expeditions.

In the same English TED Talk (785), 10 of the 13 disagreements for the *audience* attribute are due to differences in opinion as to whether an instance of "you" is generic or deictic. For example, in Ex. 4.14, one annotator labelled all instances of "you/yourself" as generic and the other labelled all instances as deictic.

(4.14) **You** get in this capsule, **you** go down to this dark hostile environment where there is no hope of rescue if **you** can't get back by **yourself**.

Even for those attributes for which agreement is very high, there may be some cases that humans find difficult to annotate or where they make mistakes. For example, the annotators disagreed on whether the *position* of the instances of "it" in Ex. 4.15 should be labelled as subject or object position.

(4.15) And some of **it** worked and some of **it** didn't.

4.8 Automatic Insertion of Unambiguous Pronouns

First-person singular pronouns in the TED Talks corpus are automatically annotated following manual annotation (see Section 4.4.6.2). With the exception of those in quoted text, which will already have been annotated according to the specific guidelines in Section 4.4.10, first-person singular pronouns are unambiguous in both English and German. They can therefore be automatically labelled as speaker reference pronouns with a high degree of accuracy. Given that

first-person singular pronouns occur frequently in the TED Talks corpus, their automated annotation reduces the manual annotation effort. A small number of first-person singular pronouns were accidentally omitted during manual annotation of the EU Bookshop corpus and discovered during automated checks. These were automatically annotated (12 for English and 13 for German).

The unambiguous generic German pronoun "man" (equivalent to "one" in English) was also automatically annotated following manual annotation, for both the EU Bookshop and TED Talks corpora.

4.9 Corpus Statistics: TED and EU Bookshop

Corpus counts for pronouns by functional *type* and *form* are provided in Tables 4.7 and 4.8 respectively[8]. The pronouns and the *form* category (personal, possessive, reflexive etc.) that they belong to are listed in Appendix B. Because there are differences in some of the pronoun types annotated in the TED Talks and EU Bookshop corpora, some type categories are marked as not applicable in the tables. The corpus annotation differences are explained in Section 4.4.

The TED Talks corpus has a greater pronoun density than the EU Bookshop corpus. The EU Bookshop corpus contains approximately 30.49 English pronouns per 1,000 tokens and 35.20 German pronouns per 1,000 tokens. The TED Talks corpus contains approximately 98.62 English pronouns per 1,000 tokens and 125.95 German pronouns per 1,000 tokens.

The differences in pronoun counts between English and German suggest differences in the use of pronouns in these languages (see Tables 4.7 and 4.9). More anaphoric pronouns and many more pleonastic pronouns are marked in the German sides of each corpus. However, a true comparison would require further analysis of the data to determine the extent to which this is a result of systematic differences between English and German versus the effect of *translationese* and translation direction. The analysis of manual translation in Chapter 5 discusses some concrete reasons for differences in anaphoric and pleonastic pronoun use between the English and German texts. The effects of translationese on pronoun translation are not considered. Such an analysis would require the definition of translationese patterns and their manual or automatic identification.

[8]Some counts differ from those in Guillou et al. (2014) due to minor changes/corrections prior to corpus release and the automatic addition of first-person pronouns and German "man".

Pronoun Type	**TED Talks**				**EU Bookshop**			
	English		**German**		**English**		**German**	
Anaphoric	*886*	*(27.71)*	*1,228*	*(40.52)*	*2,767*	*(20.32)*	*3,036*	*(22.72)*
Anaphoric (pronominal adverb)	N/A		N/A		70	(0.51)	84	(0.63)
Cataphoric	5	(0.16)	16	(0.53)	67	(0.49)	19	(0.14)
Event reference	264	(8.26)	331	(10.92)	239	(1.76)	255	(1.91)
Event (pronominal adverb)	N/A		N/A		0	(0.00)	78	(0.58)
Extra-textual reference	52	(1.63)	26	(0.86)	N/A		N/A	
Pleonastic (non-referential)	*61*	*(1.91)*	*224*	*(7.39)*	*191*	*(1.40)*	*391*	*(2.93)*
Addressee reference	499	(15.61)	525	(17.32)	112	(0.82)	76	(0.57)
Speaker reference	1,386	(43.35)	1,467	(48.41)	548	(4.02)	580	(4.34)
Generic	N/A		N/A		9	(0.07)	58	(0.43)
Pronoun (other)	N/A		N/A		135	(0.99)	126	(0.94)
Pronoun (unsure)	N/A		N/A		14	(0.10)	0	(0.00)
Total	3,153	(98.62)	3,817	(125.95)	4,152	(30.49)	4,703	(35.20)

Table 4.7: Pronoun **type** counts for English (source) and German (translation) texts in ParCor. Counts per 1,000 tokens are provided in parentheses. *N/A* indicates that the type is not marked for one of the corpora

Pronoun Form	**TED Talks**				**EU Bookshop**			
	English		**German**		**English**		**German**	
First-person personal	1,181	(36.94)	1,259	(41.54)	431	(3.17)	461	(3.45)
Second-person personal	454	(14.20)	525	(17.32)	80	(0.59)	76	(0.57)
Third-person personal	949	(29.68)	881	(29.07)	1,492	(10.96)	1,516	(11.35)
Possessive	326	(10.20)	293	(9.67)	1,003	(7.37)	906	(6.78)
Relative/Demonstrative	208	(6.51)	771	(25.44)	986	(7.24)	1,547	(11.58)
Reflexive	35	(1.09)	84	(2.77)	0	(0.00)	3	(0.02)
Pronominal Adverbs	N/A	(0.00)	N/A	(0.00)	71	(0.52)	164	(1.23)
Other	0	(0.00)	4	(0.13)	89	(0.65)	30	(0.22)
Total	3,153	(98.62)	3,817	(125.95)	4,152	(30.49)	4,703	(35.20)

Table 4.8: Pronoun **form** counts for English and German texts in the TED Talks and EU Bookshop portions of the corpus. Counts per 1,000 tokens are provided in parentheses

N.B. The category *Pronoun (unsure)* marks those pronouns that are truly ambiguous and for which the type cannot be determined (see Section 4.4.9). For example, the primary English EU Bookshop annotator reported a number of instances where the pronoun could be anaphoric or event reference. This problem did not arise in the annotation of the German translations. It is possible that some disambiguation, for better or worse, takes place during the translation process.

All documents were originally written or spoken in English, with the possible exception of one EU Bookshop document (MI3112464, "Road Transport ") for which the source language could not be identified[9].

The counts displayed in the tables are merely net differences: They do not take into consideration the number of sentences for which differences exist in terms of pronoun use (i.e. the addition of a pronoun in one translated sentence and the omission of another from a different sentence will cancel out). A more complete analysis is provided in Chapter 5.

4.10 Corpus Statistics: TEDx

Corpus counts for the TEDx corpus, by pronoun *type* and *form*, are provided in tables 4.9 and 4.10 respectively. The pronoun density for TEDx Talks is similar to that for TED Talks, with 106.36 English pronouns per 1,000 tokens, and 118.36 German pronouns per 1,000 tokens.

4.11 Discussion

The two-phase annotation approach is similar to that employed by Lapshinova-Koltunski and Kunz (2014), although the aims of their annotation project are different from the work described here. Lapshinova-Koltunski and Kunz (2014) also observed positive results with regards to both reduced manual annotation time and improved inter-annotator agreement. The aim of their GECCo corpus is to serve as a resource for conducting comparative studies between two languages, in both manual and automated translation. Chapter 5 presents a similar analysis using the ParCor corpus.

[9]Information regarding the original language of EU Bookshop documents is not provided in the online archive and the administration team was unable to provide this information for all documents.

Pronoun Type	**TEDx Talks**			
	English		**German**	
Anaphoric	*193*	*(35.89)*	*173*	*(33.51)*
Cataphoric	4	(0.74)	5	(0.97)
Event reference	40	(7.44)	71	(13.75)
Extra-textual reference	33	(6.14)	8	(1.55)
Pleonastic (non-referential)	*13*	*(2.42)*	*34*	*(6.59)*
Addressee reference	120	(22.31)	157	(30.41)
Speaker reference	169	(31.42)	163	(31.58)
Total	572	(106.36)	611	(118.36)

Table 4.9: Pronoun **type** counts for German (source) and English (translation) texts in the TEDx Talks. Counts per 1,000 tokens are provided in parentheses

Pronoun Form	**TEDx Talks**			
	English		**German**	
First-person personal	148	(27.52)	148	(28.67)
Second-person personal	115	(21.38)	157	(30.41)
Third-person personal	154	(28.64)	91	(17.63)
Possessive	51	(9.48)	17	(3.29)
Relative/Demonstrative	88	(16.36)	169	(32.74)
Reflexive	4	(0.74)	17	(3.29)
Other	12	(2.23)	12	(2.32)
Total	572	(106.36)	611	(118.36)

Table 4.10: Pronoun **form** counts for German and English texts in the TEDx Talks. Counts per 1,000 tokens are provided in parentheses

4.12 Conclusion

This chapter presented the ParCor corpus which was constructed with the dual aims of providing a resource from which to learn systematic differences in pronoun use between two languages and as a gold-standard test set for use in developing and testing SMT systems. The application of the corpus for these purposes, is demonstrated in the following chapters.

Possible extensions to the ParCor corpus include the addition of more language pairs and/or text genres, and the annotation of full coreference chains/sets.

Chapter 5

Pronoun-focussed Analysis of Manual Translation

This chapter describes a corpus analysis of differences in pronoun use between original English texts and their German translations by human translators. The analysis makes use of the English-German texts in the ParCor corpus (Guillou et al., 2014) and the pronoun type (i.e. function) labels provided by the manual annotations.

The corpus analysis aims to answer a number of specific questions, including whether there are systematic differences in pronoun use between English and German, and if so, what implications these differences have for the design of SMT systems. The outcome of the analysis is the discovery that the German translations in ParCor contain many more anaphoric and pleonastic pronouns than the original English texts. This discovery motivates the detailed analysis of the performance of state-of-the-art SMT systems on the translation of anaphoric pronouns in Chapter 6.

The contributions of this chapter include details of the differences in pronoun use between the English and German texts in ParCor, insights into why these differences arise, and the recommendations with respect to the design of SMT systems and evaluation methods for pronoun translation. The main contribution, however, is the methodology behind the analysis, which may be applied to other language pairs for which ParCor annotations exist. Indeed future effort in terms of corpus analysis is to be encouraged as a means to better understanding the problems involved in translating different groups of pronouns, for different language pairs and different text genres.

The content of this chapter is based on a paper published at the 2nd Workshop on Discourse in Machine Translation (DiscoMT) at EMNLP 2015 (Guillou and Webber, 2015). Extensions and additions include further detailed sentence-level analysis of the data. My co-author, Bonnie webber, contributed to this work in terms of high-level discussions of the analysis and more detailed discussions of specific examples and findings.

5.1 Overview

Previous work on pronouns in SMT has focussed on third-person pronouns, with some approaches treating them all as anaphoric. With the exception of some consideration paid to the use of pleonastic "it", little attention has been paid to other functions or other groups of pronouns. In the most comprehensive study to date, Hardmeier (2014) concludes that current models for pronoun translation are insufficient. To address this, he suggests that *"...future approaches to pronoun translation in SMT will require extensive corpus analysis to study how pronouns of a given source language are rendered in a given target language"*. The analyses of pronouns in English-German manual translation by Becher (2011) and Kunz and Lapshinova-Koltunski (2015) that are introduced in Section 2.6.3 provide useful insights into a number of differences in pronoun use but do not have a specific focus on machine translation. The work in this chapter aims to address this gap, presenting a corpus analysis of original English texts and their manual German translations and highlighting some of the problems that hinder progress in improving pronoun translation in SMT.

To investigate similarities and differences in pronoun use across languages, the ParCor corpus of pronoun annotations over a set of parallel English-German texts is used. For details of the ParCor corpus annotation scheme, see Chapter 4.

The manual annotations of pronoun type in the ParCor corpus allows for the corpus analysis to be carried out automatically once the parallel texts have been word–aligned. The annotations also allow for the separation of ambiguous pronouns such as "it" which may serve as an anaphoric, event reference or pleonastic pronoun. This allows for a more granular analysis than has been provided in other similar studies.

5.2 Corpus Analysis

The aim of the corpus analysis of manual translation is to identify and understand systematic differences in pronoun use between a pair of languages, with the ultimate aim of informing the design of SMT systems. Original English texts and their human-authored German translations in the ParCor corpus are compared at the corpus, document and sentence levels. Analysis at the corpus level reveals net differences in pronoun use between a pair of languages. Generalisations may be made if the same differences observed at the corpus level are also observed at the document level, when differences in authors/speakers and translators are taken into consideration. At the sentence level, we may identify more fine-grained patterns of pronoun use which help to explain some of the differences observed at the corpus and documents levels.

Separate analyses are conducted for the TED Talks and EU Bookshop corpora, so as to ascertain whether differences in pronoun use are specific to genre.

5.2.1 Corpus Level

Corpus-level comparison reveals the first differences between pronoun use in the two languages. (See Table 5.1.) Specifically, the German translations contain more anaphoric and pleonastic pronouns than the original English texts. Paired t-tests show that this difference is statistically significant for pleonastic pronouns in both the TED Talks corpus, $t(10)$=-5.08, $p < .01$, and the EU Bookshop corpus, $t(10)$=-3.68, $p < .01$. The difference in anaphoric pronoun use is statistically significant for the TED Talks corpus, t(7)=-3.52, p < .01, but not the EU Bookshop corpus, t(7)=-1.09, (p=0.31).

That the German translations contain more pronouns than the original English texts at first appears to go against the finding of Koppel and Ordan (2011), that pronouns are under-represented in translation[1]. They attribute this under-representation to two possible causes. The first is *explicitation*, in which pronouns are replaced with noun phrases. The second is *simplification* (Laviosa, 2002), whereby the translator simplifies the message of the original text, the language, or possibly both. Both of these reasons likely hold true for the manual translations of the ParCor texts. However, the sentence-level corpus analysis described in Section 5.2.3 reveals other, more specific reasons for the differences in

[1]English-German is considered among other language pairs.

Pronoun Type	TED Talks				EU Bookshop			
	English		German		English		German	
Anaphoric	*886*	*(27.71)*	*1,228*	*(40.52)*	*2,767*	*(20.32)*	*3,036*	*(22.72)*
Anaphoric (pronominal adverb)	N/A		N/A		70	(0.51)	84	(0.63)
Cataphoric	5	(0.16)	16	(0.53)	67	(0.49)	19	(0.14)
Event reference	264	(8.26)	331	(10.92)	239	(1.76)	255	(1.91)
Event (pronominal adverb)	N/A		N/A		0	(0.00)	78	(0.58)
Extra-textual reference	52	(1.63)	26	(0.86)	N/A		N/A	
Pleonastic (non-referential)	*61*	*(1.91)*	*224*	*(7.39)*	*191*	*(1.40)*	*391*	*(2.93)*
Addressee reference	499	(15.61)	525	(17.32)	112	(0.82)	76	(0.57)
Speaker reference	1,386	(43.35)	1,467	(48.41)	548	(4.02)	580	(4.34)
Generic	N/A		N/A		9	(0.07)	58	(0.43)
Pronoun (other)	N/A		N/A		135	(0.99)	126	(0.94)
Pronoun (unsure)	N/A		N/A		14	(0.10)	0	(0.00)
Total	3,153	(98.62)	3,817	(125.95)	4,152	(30.49)	4,703	(35.20)

Table 5.1: Pronoun **type** counts for English (source) and German (translation) texts in ParCor. Counts per 1000 tokens are provided in parentheses. *N/A* indicates that the type is not marked for one of the corpora

anaphoric and pleonastic pronoun use.

5.2.2 Document Level

Again, at the document level the German translations contain more pronouns than the original English texts, with the exception of a single document in the EU Bookshop corpus (NA3211776). In terms of pronoun type, the German translations typically contain more anaphoric and pleonastic pronouns than the original English texts. See tables 5.2 and 5.3 for the document-level pronoun counts for the TED Talks and EU Bookshop texts respectively.

Similar trends for anaphoric and pleonastic pronoun use were observed for many of the documents in the corpus, which suggests that this is not simply a consequence of stylistic differences over authors or speakers.

Documents in ParCor were originally produced in English and then translated into German. To ascertain whether similar patterns of pronoun use can be ob-

Document	**767**		**769**		**779**		**783**		**785**		**790**		**792**		**799**		**805**		**824**		**837**	
Pronoun Type	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**
Anaphoric	121	189	99	139	97	151	19	34	53	86	135	180	39	78	91	109	47	66	138	144	47	52
Cataphoric	0	2	1	0	1	3	0	3	0	3	2	2	1	3	0	0	0	0	0	0	0	0
Event reference	49	59	12	14	24	32	12	9	35	44	14	28	22	12	33	38	3	3	46	80	14	12
Extra-textual ref.	5	6	3	1	3	2	2	0	2	0	2	1	17	9	16	7	0	0	2	0	0	0
Pleonastic	8	54	9	24	11	38	1	1	3	11	4	12	2	10	12	29	1	3	7	31	3	11
Addressee ref.	102	91	34	44	82	89	20	18	47	37	29	36	49	51	66	80	5	5	49	59	16	15
Speaker ref.	156	163	198	216	112	124	69	72	175	201	110	122	97	102	195	195	45	41	208	208	21	23
Total	441	564	356	438	330	439	123	137	315	382	296	381	227	265	413	458	101	118	450	522	101	113

Table 5.2: Pronoun **type** counts for TED Talk texts (en=English;de=German)

served for the opposite translation direction, two German TEDx talks and their English translations were also annotated, again using the guidelines described in Guillou et al. (2014).

Similar patterns in pronoun use are observed for the TEDx Talks (see Table 5.4), with more pleonastic pronouns used in German than in English (19 vs. 11 pleonastic pronouns in one document, and 15 vs. 2 in the other). For anaphoric pronouns, one document has 119 in the German original and 140 in the English translation, with near equal numbers (54 vs. 51) in the other document. With only two documents it is not possible to confirm whether German systematically makes use of more anaphoric and pleonastic pronouns, but cf. Becher (2011) who points to several patterns, in particular the insertion of explicit possessive pronouns in German-to-English translation and pronominal adverbs in the opposite direction. Kunz and Lapshinova-Koltunski (2015) report that English uses more personal pronouns and demonstrative determiners than German and that German uses more demonstrative pronouns and pronominal adverbs than English. The observation that more pronominal adverbs are used in German than in English is mirrored in the pronoun type counts for the EU Bookshop corpus (see Table 5.1). In the EU Bookshop corpus, there are more anaphoric pronominal adverbs in the German translations than the English original texts (84 in German; 70 in English) and many more event pronominal adverbs (78 in German; none in English). With respect to the other claims, the pronoun type categories in ParCor are more difficult to align, especially given that the anaphoric category covers personal, possessive and demonstrative pronouns.

Document	**KEBC11002**		**KEBC12001**		**KH7911105**		**MI3112464**		**MJ3011331**		**NA3211776**		**QE3011322**		**QE3211790**	
Pronoun Type	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**	**en**	**de**
Anaphoric	445	653	534	443	86	94	79	80	92	141	455	437	208	236	868	952
Anaphoric (pro. adv.)	20	32	21	26	0	0	4	7	3	3	9	7	7	6	6	3
Cataphoric	9	0	5	0	1	0	4	0	0	1	19	3	13	9	16	6
Event reference	28	40	56	71	7	9	5	10	17	18	42	28	18	14	66	65
Event (pro. adv.)	0	8	0	30	0	1	0	8	0	16	0	2	0	4	0	9
Pleonastic	33	83	41	88	2	12	2	5	11	25	18	36	12	21	72	121
Addressee ref.	19	18	10	9	2	4	3	2	3	2	37	30	6	4	32	7
Speaker ref.	50	48	68	69	11	11	8	9	22	24	7	16	46	46	336	357
Generic	0	5	0	5	0	1	0	1	0	0	4	1	0	3	5	42
Pronoun (other)	25	4	49	29	3	2	0	0	2	0	8	6	11	16	37	69
Pronoun (unsure)	4	0	2	0	1	0	0	0	5	0	0	0	0	0	2	0
Total	633	891	786	770	113	134	105	122	155	230	599	566	321	359	1,440	1,631

Table 5.3: Pronoun **type** counts for EU Bookshop texts (en=English;de=German)

Document	Komfortzone		Wheelmap	
Pronoun Type	**de**	**en**	**de**	**en**
Anaphoric	119	142	54	51
Cataphoric	5	4	0	0
Event reference	63	35	8	5
Extra-textual ref.	8	33	0	0
Pleonastic	19	11	15	2
Addressee ref.	131	99	26	21
Speaker ref.	108	117	55	52
Total	453	441	158	131

Table 5.4: Pronoun **type** counts for TEDx Talk texts (de=German;en=English)

5.2.3 Sentence Level

Pronoun counts at the corpus and document levels are simply raw counts. They do not tell us anything about cases in which a pronoun is used in the original text and dropped from the translation (*deletions*), or is absent from the original text but present in the translation (*insertions*). To discover this, it is necessary to drill down to the sentence level.

The sentence–aligned parallel texts provided as part of the ParCor release provide the starting point for this part of the corpus analysis. Word alignments are used to identify the German translation of each pronoun in the original English text. The word alignments are computed using Giza++[2] with *grow-diag-final-and* symmetrisation. To ensure robust alignments, the ParCor texts are concatenated with additional data – specifically the IWSLT 2013 shared task training data (for TED and TEDx) and Europarl data (for EU Bookshop). An English and a German pronoun are considered equivalent if the following conditions hold: (a) a word alignment exists between them, and (b) they share the same pronoun type label in the ParCor annotations. The number of pronouns that meet these conditions (and those that do not) are computed automatically.

As the automatic comparison hinges upon word alignments, an assessment of the quality of these alignments is necessary. Whilst automatically generated word alignments typically suffer from problems whereby punctuation marks (especially the sentence-final period) can be aligned to just about anything, the alignment of pronouns appears to be reasonable. The word alignment quality was manually

[2]Giza++: https://code.google.com/p/giza-pp/

assessed for a random sample of 100 sentences from the TED Talks corpus. These 100 sentences contain 213 English and 241 German pronouns. A bad alignment is defined as one where a pronoun is aligned to something that is not the corresponding pronoun in the other language, or should be unaligned but is not. 14 (6.57%) English and 22 (9.12%) German pronouns in the sample are part of a bad alignment. It should be noted that alignment quality may vary depending on the pronoun in question and the frequency of the pronoun in the corpus may be a contributing factor. For a complete understanding of the problem of pronoun alignment, a more detailed study would be required.

Bad alignments may arise for a number of reasons. For example, restructuring of the original sentence in translation, or as a result of a reflexive verb being used in German and not in English (which can result in the use of two German pronouns corresponding to a single pronoun in English). It is perhaps because pronouns in both English and German are overt that the alignment quality is reasonably good. Pro-drop, which is seen in other languages (Spanish, Czech etc.) and could contribute to misalignments through aligning an overt English pronoun with a non-pronoun word in the Spanish/Czech/etc., is not an issue. The effect of bad word alignments on the analysis is mitigated, to some extent, by considering both the alignment and pronoun functional type information.

Those pronouns for which the equivalence conditions (i.e. both pronouns share the same type label and a word-alignment exists between them) hold are considered to be *matches*, with all other pronouns, in the original text or its translation considered *mismatches*. Mismatches are split into:

- Deletions: A pronoun is present in the original English text that is missing from the German translation
- Insertions: A pronoun is present in the German translation but not in the original English text

An example of a mismatch is shown below. In Ex. 5.2 the pronoun "Ihnen" ("you") is inserted into the German translation but is not present in the original English sentence (Ex. 5.1). (N.B. A translation without "Ihnen" is also possible.)

(5.1) I'm going to talk today about energy and climate .

(5.2) Heute spreche ich zu **Ihnen** über Energie und Klima .
(Lit: Today I will speak to **you** about energy and climate)

Pronoun Type	**English (deletion)**	**German (insertion)**
Anaphoric	*49*	*117*
Cataphoric	0	2
Event reference	26	36
Extra-textual ref.	4	5
Pleonastic	*3*	*49*
Addressee reference	31	20
Speaker reference	30	37
Total	143	266

Table 5.5: Sentence-level pronoun **type + alignment** mismatches for TED Talk 767

Other examples of mismatches involving anaphoric and pleonastic pronouns are shown in Sections 5.2.3.1 and 5.2.3.2 respectively.

Taking TED Talk 767 as an example and using the combination of pronoun type and alignments to identify a source-target pronoun match, many mismatches are observed. Table 5.5 shows that 409 pronouns are unique to either the English text or the German translation. This leaves only 298 matching English-German pronoun pairs. The largest absolute difference lies in the number of anaphoric pronouns in the target-language text for which there is no comparable pronoun in the source-language text (*anaphoric* insertions), followed by *pleonastic* insertions.

The process for automatically comparing pronouns produces not only document and sentence-level statistics but also a file containing aligned sentence pairs in which the following is annotated:

- Equivalent source and target-language pronouns are indexed, such that both pronouns in the pair are allocated the same numeric ID
- Anaphoric pronouns only:
  - Antecedent spans are marked using [ ] and are given an index value
  - Pronouns are marked with the index of their antecedent (in addition to the index which links them to the equivalent pronoun in the other language)

Ex. 5.3 from TED Talk 785 (sentence 98) and its German translation (Ex. 5.4) shows the pronoun labelling:

(5.3) And [**we**]1 wound up going to the Bismark , and exploring [**it**]2 with robotic vehicles .

(5.4) [**Wir**]1 tauchten zur " Bismarck " , und erforschten [**sie**]2 mit Roboterfahrzeugen .
(Lit: [**We**]1 dove to the "Bismarck" and explored [**it**]2 with robotic vehicles.)

These annotated files may then be used to conduct detailed manual analyses to better understand the reasons behind insertions and deletions during manual translation.

5.2.3.1 Anaphoric Insertions and Deletions

There is no single reason for *anaphoric* deletions: Anaphoric pronouns may be omitted from the German output for stylistic reasons, as a result of paraphrasing or possibly to conform with language-specific constraints. Specific reasons may include translating an active sentence as a passive one or replacing a pronoun with a full referring expression. The opposite may also happen, leading to the insertion of pronouns in German.

With respect to *anaphoric* insertions, intra-sententially, many correspond to relativizers in English. That is, while in English a relative clause is introduced with a *that-*, *wh-* or *null-relativizer*, an anaphoric pronoun serves as a relativizer in German.[3] For example, "that" in "The house **that** Jack built" is a relativizer and the corresponding "das" in "Das Haus, **das** Jack gebaut hat" is a relative pronoun. As with other anaphoric pronouns in German, a relative pronoun must agree with its antecedent in terms of number and gender. Here, the antecedent is the translation of the head noun of the NP that appears immediately before the relativizer in the original English text. In the previous example the antecedent is "Haus" ("house", neuter).

The following example from the TED Talks corpus illustrates a mismatch in pronoun use arising from the insertion of a relative pronoun in the German translation. The relative pronoun "die" which refers to "Dienste" ("services" [pl.])

[3]The ParCor corpus has not marked instances of *that* when used as a relativizer in English.

ID	English		German	
	Words	Pleonastic pronouns	Words	Pleonastic pronouns
767	5,371	8	4,775	54
769	3,414	9	3,430	24
779	3,564	11	3,566	38
783	1,280	1	1,163	1
785	3,265	3	3,054	11
790	2,988	4	2,921	12
792	1,788	2	1,768	10
799	4,354	12	3,947	29
805	1,002	1	989	3
824	3,644	7	3,531	31
837	1,301	3	1,161	11

Table 5.6: Pleonastic pronouns marked in the TED Talks

in Ex. 5.6 corresponds to the null-relativizer (∅) in the original English sentence (Ex. 5.5).

(5.5) The second factor is the services ∅ we use .

(5.6) Der zweite Faktor sind die Dienste , **die** wir nutzen .

Manual analysis of the German translation for TED Talk 767 identified 42 cases where an anaphoric pronoun was inserted as a relative pronoun corresponding to a relativizer in English. 35 of the relative pronouns inserted in the German translation correspond to a that-relativizer in English. While this does not explain all of the *anaphoric* insertions, it is frequent enough to deserve further attention.

5.2.3.2 Pleonastic Insertions and Deletions

The pleonastic pronoun counts in the TED Talks corpus are displayed in Table 5.6.

Fixed expressions in English appear to trigger *pleonastic* insertions in German. A commonly observed pair is "There *+be*"/"Es gibt". These *existential there* constructions, used in *existential clauses* or *sentences*, are not annotated in ParCor, but their presence accounts for some (not all) of the insertions of pleonastic pronouns in German.

For example:

(5.7) **There are** some innovations in nuclear : modular , liquid .

(5.8) **Es gibt** einige Innovationen im Nuklearbereich ; modular , flüssig .

Manual analysis of TED Talk 767 identified 54 pleonastic pronouns in the German translation:

- 5 cases where the pleonastic pronoun in German corresponds to a pleonastic pronoun in English
- 25 cases where the pleonastic pronoun in German **does not correspond** to a pleonastic pronoun in English (and no further pattern is identified)
- 24 cases where a pleonastic pronoun was inserted into the German translation, corresponding to a "there *+be*" expression in the English text

The TED Talks were delivered by different speakers and translated into German by different translators. By examining other English texts and their translations, we can see that the correlation between "there *+be*" expressions in English and the insertion of pleonastic pronouns in German is not just a result of the actions of a single translator. A similar pattern is observed for TED Talk 799. The German translation contains 29 pleonastic pronouns:

- 7 cases where the pleonastic pronoun in German corresponds to a pleonastic pronoun in English
- 14 cases where the pleonastic pronoun in German **does not correspond** to a pleonastic pronoun in English (and no further pattern is identified)
- 8 cases where a pleonastic pronoun was inserted into the German translation, corresponding to a "there *+be*" expression in the English text

Again, the use of "there *+be*" expressions in English does not explain all of the *pleonastic* insertions in German. As the fixed expressions are short and occur frequently, phrase-based systems could be expected to provide accurate translations. This was confirmed by a manual inspection of 100 instances of "there *+be*" expressions in the TED Talks corpus and their German translations by the state-of-the-art phrase-based system employed in the study in Chapter 6. 85%

(85/100) "there *+be*" expressions were successfully translated using an equivalent German construction continuing a pleonastic pronoun (e.g. "es gibt" or "es gab"). Further study of pleonastic pronouns is therefore excluded from the analysis of state-of-the-art SMT output in Chapter 6.

What about the remaining pleonastic pronoun insertions? One possibility is that the increase in German pleonastic pronouns is due to passivisation of the original English sentence. In particular, this may be achieved via the use of impersonal constructions: e.g. "**It** is said that..." rather than "People say that...". For example the "es" in Ex. 5.10 (from TED Talk 767) is introduced as part of the impersonal "gibt es" construction, corresponding to the personal "we ... have" construction in English:

(5.9) Usually , **we don 't have a deadline** , where you have to get the miracle by a certain date .

(5.10) Normalerweise **gibt *es* keine Deadline** , dass man ein Wunder bis zu einem bestimmten Datum braucht .

Again, these expressions may be short, which means that a phrase-based SMT system ought to stand a good chance of producing a correct translation.

Of course not all pronouns will be introduced into a translation for one the of the above reasons. Pronouns may be inserted, deleted or substituted at the whim of the translator (cf. Becher (2011)). The aim of this analysis is to try to identify majority cases.

5.3 Implications for SMT

That the human translators do not always follow the same use of pronouns exhibited in the source-language text, has serious implications for SMT systems and automated evaluation methods for pronoun translation.

5.3.1 SMT System Design

Since SMT systems are trained on parallel data similar to that in ParCor, it is important to be aware that content words such as nouns and verbs are more likely to be faithfully translated as there are fewer ways to convey the same meaning. On the other hand, there is more variation in the translation of function words such as

pronouns — for example in active to passive conversions (and vice versa). Where there is a lot of variation, the SMT system may not be able to learn accurate translation mappings. Whilst word alignment quality is reasonably good for pronouns (see Section 5.2.3) poor alignments still arise. These poor alignments may contribute to translation errors. However, even with perfect alignments, pronouns may be translated incorrectly if insufficient context is provided – for example, the problem of translating an anaphoric pronoun without knowledge of its antecedent, does not go away. The frequent insertion of pronouns in the target language where none were present in the source-language text, and the deletion of pronouns from the target language where one was present in the source-language text both give rise to a number of questions:

- Where and when is it **appropriate** or **necessary** to insert/delete a pronoun? I.e. is the translation acceptable, even if it is not *natural*? Is a pronoun required to produce an accurate translation, or to adhere to constraints of the target language?
- Does the answer differ depending on whether the pronoun in question has a specific referent or if the pronoun is indefinite or non-referential?

These questions should be addressed if the aim is to provide a comprehensive solution to pronoun translation. Previous work has not addressed these problems directly, and the insertion of dropping of pronouns by SMT systems has been by accident rather than design. This thesis raises the questions, but does not seek to address them. A wider question might be: What is the aim of pronoun translation in SMT? Is it to provide a sufficiently accurate translation such that a human reader of the original source-language text believes the same thing(s) as a different human reader of the translated text? Or is it to provide a natural translation similar to one that a human translator might produce? Here we can draw a parallel with the qualities of *adequacy* and *fluency* that SMT output may be measured against at a more general level. Such issues have not yet been discussed within the field, but ought to be addressed in the future.

5.3.2 Evaluation

Automatic evaluation methods that rely on reference translations to assess pronoun translation will not be able to provide an adequate evaluation when the

reference translation departs from the original source-language text. Even when multiple references are available, it may not be possible to capture all of the possible variations in pronoun use. This raises questions surrounding the design of automated evaluation metrics, in particular how to score pronoun translations where:

- A pronoun in the source-language text is translated, but no corresponding pronoun is present in the reference
- A pronoun that is not present in the source-language text is inserted in the translation (both if a corresponding pronoun is present in the reference or not)

The problems presented in the above scenarios apply to all pronoun types. Measuring the translation accuracy of anaphoric pronouns, presents an additional set of problems. Firstly, an automatic evaluation metric that relies solely on the reference translation, may make poor decisions. Consider the following scenarios:

1. The SMT output and the reference translation contain **different translations** of a given source-language pronoun
2. The SMT output and the reference translation contain **the same translation** of a given source-language pronoun

Now consider the following example:

(5.11) I have a **box**. **It** is large.

(5.12) Ich habe eine **Box**. **Sie** ist groß. [Reference]

(5.13) Ich habe einen **Karton**. **Er** ist groß. [SMT output]

The reference translation (Ex. 5.12) translates "box" in the original English input as "Box" (feminine) and the anaphoric pronoun "it" as "sie" (feminine). The SMT system (Ex. 5.13) translates "box" as "Karton" (masculine) and the anaphoric pronoun "it" as "er" (masculine). Pronoun-antecedent agreement requirements are met in both the reference translation and the SMT output, and both translations are valid. However, an automatic evaluation metric that relies solely on the reference translation to assess the accuracy of pronoun translation (and considers only the translation of the pronoun) would detect that the SMT

output contained a different translation of the pronoun to the reference and (in error) mark the translation as incorrect. The reverse is also true. The same metric would also mark a pronoun as correct if it matched the one used in the reference translation, even if pronoun-antecedent agreement did not hold.

This problem whereby the SMT output and reference may contain different translations of the same antecedent, each with a different number and/or gender, has also been highlighted by Hardmeier (2014). The implication for SMT is that automatic evaluation metrics, which are necessary for further progress, cannot simply rely on pronoun-translation comparisons with the reference translation. They must, as the case of anaphoric pronouns demonstrates, incorporate additional contextual information (i.e. the translation of the pronoun's antecedent). The information required may vary depending on the pronoun type in question. A method for the semi-automatic evaluation of pronouns is presented in Chapter 8.

5.4 Conclusion

This chapter described a corpus analysis of manual translation, using the annotations provided in the ParCor corpus. The analysis revealed that pronouns are frequently dropped and inserted by human translators and that there are many more anaphoric and pleonastic pronouns in the German translations as compared to the English original texts. These differences in pronoun use can have serious implications both for the design of SMT systems and pronoun-specific automated evaluation metrics.

Chapter 6

Pronoun-focussed Analysis of State-of-the-Art Statistical MT

This chapter describes an analysis of the ability of two state-of-the-art SMT systems to translate different categories of pronoun. It follows on from the analysis of manual translation in Chapter 5, which revealed that the German translations in ParCor contained more anaphoric pronouns than the original English texts. The analysis aims to answer the questions: Given this finding, how well do current state-of-the-art SMT systems perform at translating anaphoric pronouns, and what, if any, implications does this have for SMT system design. The anaphoric category is further sub-divided, reflecting differences in terms of the translation requirements for different types of anaphoric pronoun. The pronouns selected for investigation are anaphoric "it" and "its", and English relativizers (which may be translated in German as relative pronouns).

The analysis revealed that biases in the training data stemming from the common alignment of "it" and "es" and the selection of the correct base form for possessive pronouns (i.e. "ihr" vs. "sein" for "its" as determined by the number/gender of the antecedent or "possessor") are both problems which SMT systems must overcome. For relative pronouns selecting the correct preposition is also important as it influences the case of the pronoun. Whilst these findings represent a contribution in themselves, the main contribution of this chapter is to highlight the need to further sub-categorise pronouns. Considering pronoun function in the source language is not enough, we must also consider the target-language translation requirements.

The content of this chapter is based on a paper published at the 2nd Workshop

on Discourse in Machine Translation (DiscoMT) at EMNLP 2015 (Guillou and Webber, 2015). Extensions to the paper include (manual) translation distributions for the third-person pronouns “he”, “she”, “it” and “they”, and possessive “its”, and the inclusion and discussion of additional pronoun translation examples. Again, my co-author, Bonnie webber, contributed to this work in terms of high-level discussions of the analysis and more detailed discussions of specific examples and findings.

6.1 Overview

Analyses of the output of state-of-the-art SMT systems provide an indication of how well current systems are able to translate pronominal coreference — what they are good and bad at. Analyses provided as part of previous research were each conducted for a single system and have either focussed on a single source-language pronoun with a small number of instances (~100) (Novák, 2011) or have considered a larger, but mixed set of pronouns such that the number of instances for each source-language pronoun form is small (~100) (Hardmeier and Federico, 2010; Weiner, 2014).

The analysis of manual translation in Chapter 5 revealed differences between English and German for pleonastic and anaphoric pronouns. Following on from this work, this chapter describes an analysis of English-to-German translation for three different sub-types of anaphoric pronoun: “it”, “its” (possessive), and those relative pronouns triggered by the use of relativizers in English, which can also be considered anaphoric. While reflexive pronouns are also anaphoric, as they occur only infrequently in the ParCor corpus, they are excluded from the study. Given the high cost of manual evaluation, this analysis follows previous analyses in studying a small number of pronouns. In an attempt to maximise the usefulness of the evaluation, only a small subset of English pronouns (“it” and “its”) and relativizers (“that” and “null”) are included, and the study is carried out for two SMT systems.

The state-of-the-art systems are two English-to-German SMT systems from the IWSLT 2014 shared task in machine translation (Birch et al., 2014). The first is a phrase-based system that incorporates factored models for words, part-of-speech tags and Brown clusters. The second is a syntax-based, string-to-tree, system. Both systems were trained using a combination monolingual and parallel

data taken from the following corpora: TED data from WIT[3] (Cettolo et al., 2012), Europarl (Koehn, 2005), MultiUN (Eisele and Chen, 2010), Gigaword as provided by the Linguistic Data Consortium[1], the German Political Speeches Corpus (Barbaresi, 2012), and the corpora provided for the WMT 2014 shared translation task (Bojar et al., 2014). Both systems were tuned using the *dev2010* data provided for the WMT shared translation task. Here, TED Talks are considered to be in-domain, with the EU Bookshop texts considered out-of-domain. The different architecture of the two systems makes direct comparisons between them unfair. However, similarities in the translation accuracy of two systems can show that the findings outlined in this chapter are not specific to a single system or type of system.

For manual translation, one can assume that a pronoun is accurately translated, inserted or dropped, as part of a close translation of the original sentence or an acceptable paraphrase. As such, it is reasonable to use automated analysis based on the ParCor annotations and alignments between the texts. With automated translations, however, there is no guarantee that a source-language pronoun is translated correctly by the system. We must therefore rely more heavily on manual analysis.

However, manual analysis can be aided by some automated pre-processing steps, to help select pronouns for further study. Using the source-language text and its translation together with word alignments output by the SMT systems, it is possible to investigate which pronouns may be more difficult to translate than others – i.e. using frequency distributions of the translations produced for each source-language pronoun surface form (split by pronoun type).

6.2 Identifying Pronouns for Analysis

Examining the translation frequency distributions in the human authored reference translations in ParCor (see Tables 6.1 and 6.2), the following may be observed. First, the pronoun “it” may be translated into German, depending on its function as one of he following third-person pronouns: “er” (masculine singular (sg.)), “sie” (feminine sg.) or “es” (neuter sg.), or “sie” (plural). As plural pronouns are not gendered, “they” has fewer possible translation options. The possessive pronoun “its” has additional possible translation options due its

[1]Linguistic Data Consortium (LDC): http://www.ldc.upenn.edu

Source pronoun	Translation	Count
he	er	94
	untranslated	3
	other pronoun	3
	other POS	2
she	sie	3
it	er	15
	sie	58
	es	108
	pronominal adverb	23
	other pronoun	57
	untranslated	37
	other POS	25
they	sie	156
	other pronoun	24
	untranslated	18
	other POS	14

Table 6.1: Translation distributions for "he/she/it/they" in the ParCor TED Talks corpus

multiple dependencies. That is, possessive pronouns in German must agree in number/gender with both the possessor and the object that is possessed. Different base forms are used depending on whether the possessor is feminine/plural ("ihr") or masculine/neuter ("sein"). Other anaphoric pronouns such as "he" and "she" have far fewer translation options and are therefore less interesting. The counts in Table 6.1 for "he/she/it/they" are taken from the TED Talks corpus. Similar distributions may be observed for the EU Bookshop corpus. The possessive pronoun "its" is uncommon in the TED Talks corpus, and so distributions are reported for the EU Bookshop corpus (see Table 6.2).

Based on the possible translation options, I have selected (anaphoric) "it" and "its" for further analysis.

The analysis of manual translation (Chapter 5) showed that relativizers in English often corresponded to a relative pronoun inserted in the German translation. To see how well SMT systems handle the translation of relativizers, that-

Source pronoun	Translation	Count
its	ihr	6
	ihre	42
	ihren	15
	ihrem	5
	ihrer	26
	ihres	7
	sein	6
	seine	41
	seinen	8
	seinem	4
	seiner	16
	seines	1
	other pronoun	70
	untranslated	11
	other POS	21

Table 6.2: Translation distributions for "its" in the ParCor EU Bookshop corpus

relativizers (explicit in English text) and null-relativizers (implicit) were added to the set of pronouns for analysis. Wh-relativizers, also explicit, but with many forms (what, who, etc.), are excluded in order to reduce the annotation effort. Assessing their translation is left for future work.

6.3 Pronoun Selection Task

The manual analysis of pronoun translation is framed as a *pronoun selection task*. In this setting a human annotator is asked to identify which pronoun(s) could validly replace a placeholder masking a pronoun at a specific point in the SMT output. Masking the pronoun removes the risk that the annotator is biased by the pronoun present in the SMT output. The annotator's selections may then be compared with the pronouns produced by the SMT system in order to assess translation accuracy.

The tool developed by Hardmeier (2014) is used for the pronoun selection task. The interface in Figure 6.1 presents the annotator with the source-language sentence and its translation plus up to five previous sentences of history, as well

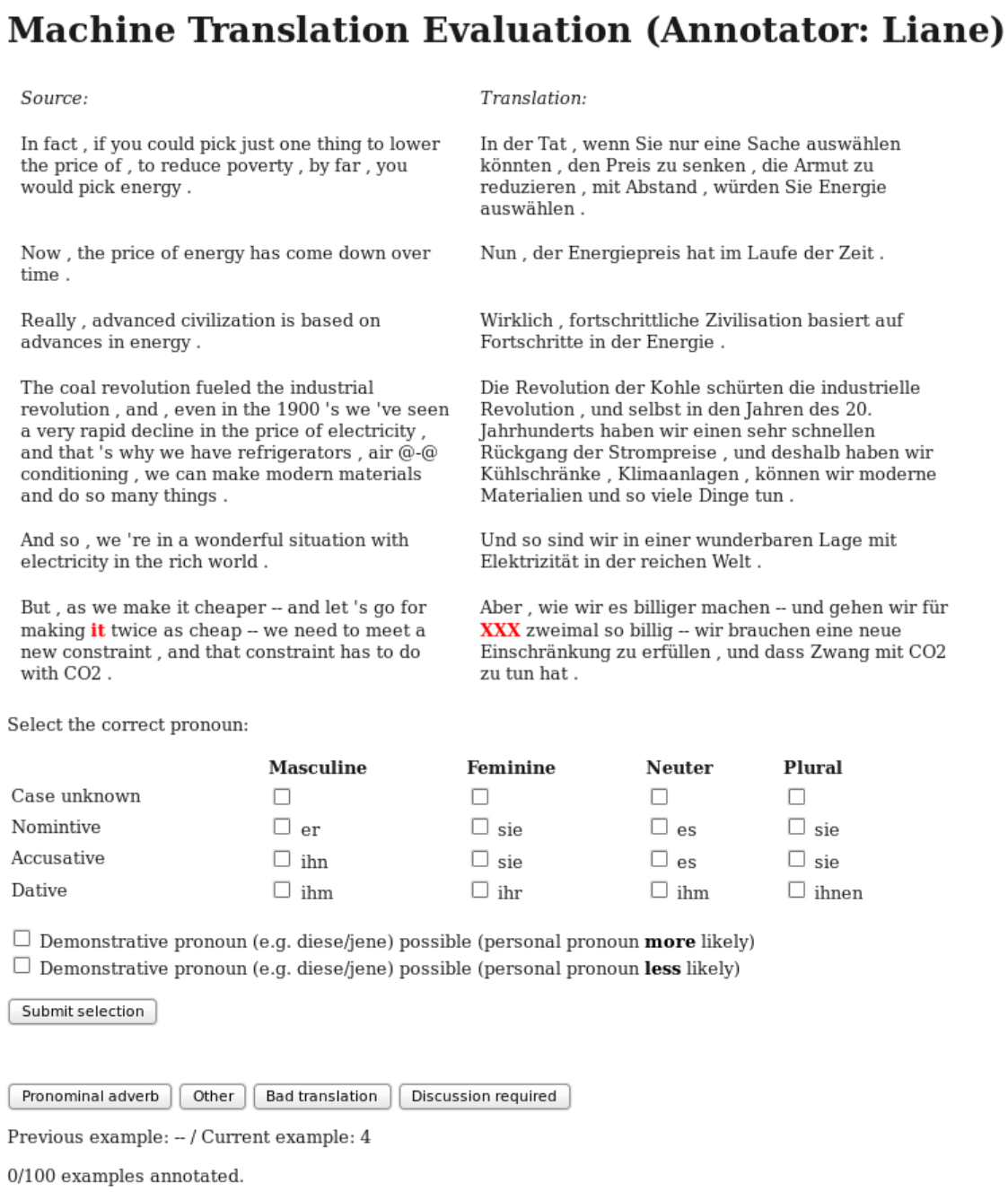
as a number of pronoun options. The source-language pronoun in the final sentence of each example block is highlighted and its translation is replaced with a placeholder.

The manual annotations in ParCor were used to determine how many sentences of history to present to the annotator (to help them identify the antecedent of an anaphoric pronoun). By calculating two standard deviations from the mean number of sentences between pronoun and antecedent, we can find the pronoun-antecedent distance that will account for 95% of the pronouns. (Intra-sentential pronouns have a distance of zero.) For the TED Talks corpus the mean distance between pronoun and antecedent is 1.33 sentences, and two standard deviations from the mean is 4.95 sentences. For the EU Bookshop corpus (in which sentences are longer), the distances between pronoun and antecedent are typically shorter, with a mean distance of 0.67 sentences and two standard deviations from the mean at 3.57 sentences. The interface is therefore configured to allow for up to five previous sentences of history for each example (fewer for examples in the first five sentences of a document), regardless of genre.

Inter-annotator agreement scores for the pronoun selection task for English-to-French translation are presented in Hardmeier (2014) and are described as being acceptable for the task. Agreement for the pronoun selection task for English-to-German translation are not computed, owing to the cost of additional manual evaluation and because agreement had already be calculated for a very similar task.

6.3.1 Guidelines

The following guidelines were adapted from those used by Hardmeier (2014) in order to cater for the requirements of English-to-German translation. Due to the inflection of pronouns in German there are more translation options for English-to-German than for English-to-French (the language pair that the tool was originally used for). In addition, the same German pronoun forms may be used for pronouns of different gender, number or case. For example, “sie” (“she”, “they”) may be used as both a third-person singular and plural pronoun, in both the nominative and accusative cases. In place of the original set of buttons labelled with pronoun surface forms, a grid of checkboxes is used to represent the different options which are expressed in terms of gender, number and case. The annotation



Machine Translation Evaluation (Annotator: Liane)

Source:	*Translation:*
In fact , if you could pick just one thing to lower the price of , to reduce poverty , by far , you would pick energy .	In der Tat , wenn Sie nur eine Sache auswählen könnten , den Preis zu senken , die Armut zu reduzieren , mit Abstand , würden Sie Energie auswählen .
Now , the price of energy has come down over time .	Nun , der Energiepreis hat im Laufe der Zeit .
Really , advanced civilization is based on advances in energy .	Wirklich , fortschrittliche Zivilisation basiert auf Fortschritte in der Energie .
The coal revolution fueled the industrial revolution , and , even in the 1900 's we 've seen a very rapid decline in the price of electricity , and that 's why we have refrigerators , air @-@ conditioning , we can make modern materials and do so many things .	Die Revolution der Kohle schürten die industrielle Revolution , und selbst in den Jahren des 20. Jahrhunderts haben wir einen sehr schnellen Rückgang der Strompreise , und deshalb haben wir Kühlschränke , Klimaanlagen , können wir moderne Materialien und so viele Dinge tun .
And so , we 're in a wonderful situation with electricity in the rich world .	Und so sind wir in einer wunderbaren Lage mit Elektrizität in der reichen Welt .
But , as we make it cheaper -- and let 's go for making **it** twice as cheap -- we need to meet a new constraint , and that constraint has to do with CO2 .	Aber , wie wir es billiger machen -- und gehen wir für **XXX** zweimal so billig -- wir brauchen eine neue Einschränkung zu erfüllen , und dass Zwang mit CO2 zu tun hat .

Select the correct pronoun:

	Masculine	**Feminine**	**Neuter**	**Plural**
Case unknown	☐	☐	☐	☐
Nomintive	☐ er	☐ sie	☐ es	☐ sie
Accusative	☐ ihn	☐ sie	☐ es	☐ sie
Dative	☐ ihm	☐ ihr	☐ ihm	☐ ihnen

☐ Demonstrative pronoun (e.g. diese/jene) possible (personal pronoun **more** likely)
☐ Demonstrative pronoun (e.g. diese/jene) possible (personal pronoun **less** likely)

Submit selection

Pronominal adverb | Other | Bad translation | Discussion required

Previous example: -- / Current example: 4

0/100 examples annotated.

Figure 6.1: Pronoun Selection Task Tool. The placeholder masks the pronoun "sie" which matches the human judgement that an accusative feminine pronoun should be used

guidelines are extended from Hardmeier (2014) to include the annotation of case.

1. Select the pronoun that will create the most fluent translation, while preserving the meaning of the English sentence as much as possible. The latter means assigning correct number/gender to the pronoun that replaces the placeholder: Its case may be left "unknown".

    - If the SMT output is sufficiently fluent to be able to determine the case of the pronoun, select the appropriate check-box.
    - Use the plural options if the antecedent is translated as a plural, or in any other scenarios in which a plural might seem appropriate.
    - If different, equally grammatical options are available, select all appropriate check-boxes.

2. Alternatively select "Other" if the sentence should be completed with a pronoun not included in the list, "Bad translation" if a grammatical and faithful translation cannot be created without making major changes to the surrounding text, or "Discussion required" if you are unsure what to do.

3. Ignore minor dis-fluencies (e.g. incorrect verb agreement or obviously missing words).

4. Always try to select the pronoun that best agrees with the antecedent in the SMT output, even if the antecedent is translated incorrectly, and even if this forces you to violate the pronoun's agreement with immediately surrounding words such as verbs, adjectives etc.

5. If the translation does not contain a placeholder, but a pronoun corresponding to the one marked in the English text should be inserted somewhere, indicate which pronoun should be inserted.

6. If the SMT output does not contain a placeholder, but already includes the correct pronoun, annotate the example as if a placeholder were present. This will mean selecting the same pronoun that is included in the SMT output.

7. Prefer "Bad translation" over the "Discussion required" option. The "Discussion required" option should be reserved for cases where there is a problem with the guidelines / options provided.

6.3.2 Assessing Correct Translations

The translations produced by the systems are automatically compared with the selections made by the annotator. If the system-generated pronoun matches one of the annotator's selections, there is a *pronoun match*. If it doesn't match any of the annotator's selections or the system did not generate a pronoun, there is a *pronoun mismatch*. Matches are recorded in terms of number/gender and case if the annotator supplied it, or number/gender only, if not.

Pronominal adverb match applies only to the translation of "it", for which the "Pronominal adverb" option exists (cf. Section 6.4.2). For example, if the translation requires the use of a pronominal adverb such as "danach" ("thereafter") or "worauf" ("whereupon"). It is used when the SMT output contains a pronominal adverb and the annotator had indicated that one would be appropriate. As the annotator was not asked to specify the pronominal adverb, no further comparison is made. *Pronominal adverb mismatch* is the opposite; the annotator indicated that a pronominal adverb should be used but the system did not output one.

"Other", "Bad translation" and "Pronoun not required"[2] are used for those pronouns marked as such in the pronoun selection task.

Instead of comparing the systems, the results from both are used to assess how well state-of-the-art systems perform at pronoun translation. The initial observation is that both systems typically produce more incorrect translations than correct ones.

6.4 Anaphoric "it"

6.4.1 Translation Requirements

The anaphoric pronoun "it" can co-refer either *intra-sententially* (i.e. to an antecedent in the same sentence) or *inter-sententially* (i.e. to an antecedent in a different sentence). While coreference imposes number–gender constraints on a pronoun and its antecedent, intra-sentential coreference imposes additional constraints.

In addition to their translation as personal pronouns in German, English

[2]Although the "pronoun not required" option was not initially provided for the "it" task, it was added later when the need arose.

anaphoric pronouns may also be translated as pronominal adverbs[3]. Consider the following example taken from (Gruber and Redeker, 2014):

(6.1) I've got a car, I'm going to Hamburg **with it** today.

(6.2) Ich habe ein Auto, **damit** fahre ich heute nach Hamburg.
(Lit: I have a car, **with it** I am driving to Hamburg today)

The anaphoric "it" in Ex. 6.1 refers to "car" and could have been translated as "es" (to agree with "Auto", neuter). Instead, in Ex. 6.2, the pronominal adverb "damit" ("with it") is used in German.

6.4.2 Pronoun Selection Task

The sample set consists of a random selection of 50 inter- and 50 intra-sentential tokens of "it" labelled as anaphoric in the ParCor annotations. Tokens were selected from the TED Talks corpus, as sentences there are typically shorter than those in the EU Bookshop corpus and hence, potentially easier for the human annotator to work with. Additional guidelines are provided for "it":

- Select "Pronominal adverb" if the most fluent translation would come from using a German pronominal adverb. (Selection of the pronominal adverb is not required.)
- If the use of a demonstrative pronoun (e.g. "diese" or "jene") is possible, select whether it is **more** or **less** likely than the personal pronoun(s).
- Genitive options are not available as these are used for possessives.

The additional options for pronominal adverbs and for expressing a preference for the use of a demonstrative pronoun over a personal pronoun are both extensions to the original guidelines in Hardmeier (2014).

The annotator is presented with a table of options for number/gender and case combinations. The number/gender options are masculine, feminine, neuter and plural. The case options are: "case unknown", and three German cases: Nominative, accusative and dative. See Figure 6.2.

Although the ParCor annotations contain antecedent links for anaphoric pronouns, these were not displayed to the annotator for any of the tasks.

[3]Pronominal adverbs also exist in English (e.g. therefore, wherein, hereafter) but are used more frequently in German.

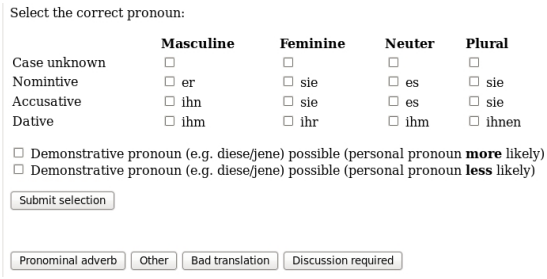


Figure 6.2: Annotator options for "it"

6.4.3 Results

The results of the pronoun selection task for 100 instances of anaphoric "it" are presented in Table 6.3.

Some instances of "it" were initially left for discussion during the pronoun selection task for anaphoric pronouns. These were later assigned one of two new categories: "Anaphoric but could not find antecedent" where the antecedent could not be identified due to insufficient history or "Unsure: may not be anaphoric" where the annotator believed that the pronoun may not in fact be anaphoric, despite being labelled as such in the ParCor corpus.

For most examples the annotator was able to determine the case of the pronoun as well as its number/gender. Recall that the annotator was specifically instructed to only select the case of the pronoun if the SMT output was sufficiently fluent so as to make this possible. It would therefore appear that the assumptions about difficulties in identifying syntactic roles in SMT output (see Section 6.3.1) that were made during the development of the annotation guidelines, are not entirely correct.

Both systems regularly translate "it" as "es": 79% (79/100) of cases for the phrase-based and 78% (78/100) for the syntax-based system. This reflects biases in the training data, where the use of "it" and "es" as both anaphoric and pleonastic pronouns leads to their frequent alignment.

It is often acceptable to translate "it" using either a personal or demonstrative pronoun: 49% (49/100) of cases for the phrase-based and 50% (59/100) of cases for the syntax-based system. However, neither system generated demonstrative pronouns, perhaps due to the bias toward translating "it" as "es".

	Inter-sentential		Intra-sentential	
Result	**Phrase**	**Syntax**	**Phrase**	**Syntax**
Pronoun match (number/gender + case)	20	8	14	15
Pronoun match (number/gender only)	0	1	1	0
Pronoun mismatch	14	28	27	26
Pronoun not translated (mismatch)	1	0	0	0
Pronominal adverb match	5	8	2	2
Pronominal adverb mismatch	2	0	0	1
Other	2	0	2	1
Bad translation	4	1	1	2
Pronoun not required	0	1	0	0
Anaphoric but could not find antecedent	0	1	0	0
Unsure: may not be anaphoric	2	2	3	3
Total	50	50	50	50

Table 6.3: Pronoun selection task results for anaphoric "it"

6.4.4 Discussion

When annotating the English side of ParCor, deciding whether a pronoun was anaphoric, event reference or pleonastic was one of the major causes of annotator disagreement. It is therefore not surprising that problems might arise in identifying the pronoun's antecedent for the pronoun selection task. This ambiguity did not arise for the "its" or relativizers tasks. With "its", events are rarely (if ever) possessors and so rarely serve as antecedents. With relativizers, the relative pronoun and its antecedent (in German) are likely to be very close together, and certainly intra-sentential.

In one case the annotator was unable to identify the pronoun's antecedent and believed that it may have appeared earlier in the text (i.e. outwith the history provided for the example). However, increasing the previous history is not practical due to the resulting increase in text displayed in the tool interface.

In Ex. 6.4 the annotator suggested replacing the placeholder with two pronouns: "**es die**", where "es" is pleonastic, and "die" is a relative pronoun referring to "Idee" ("idea"). Such scenarios were not planned for in this task, and appear to occur rarely, but they should be considered in the design of SMT systems.

(6.3) If I can leave you with one big idea today , **it** 's that...

(6.4) If kann ich Ihnen mitgeben heute eine große Idee , ist **XXX** , dass...

The syntax-based system is much better at translating intra-sentential pronouns than inter-sentential ones (15 pronoun matches for intra-sentential pronouns vs. 9 for inter-sentential pronouns). Although this system contained no such enhancements, one might expect that pronoun-aware syntax-based systems could be designed to leverage the fact that intra-sentential pronouns are syntactically governed, and to produce better translations. One possible option would be to combine two systems: A phrase-based system to translate inter-sentential pronouns, and an enhanced syntax-based system to translate intra-sentential pronouns.

Despite differences in their approach, the findings of previous analyses echo some of the findings for "it" in this study. Weiner (2014) reports translation accuracy for "it" at 47.2% and 47.6% (for TED and news genres) for English-to-German translation (with sample sizes of fewer than 50 pronouns). Novák (2011) reports translation accuracy for "it" at 31.25% for English-to-Czech translation.

6.5 Anaphoric possessive "its"

6.5.1 Translation Requirements

In German, a *dependent* possessive pronoun (i.e. one which precedes a noun) must agree not only with the number/gender of its antecedent (possessor) but also with the number/gender of its object (i.e. the noun that follows the pronoun). For example in: "**Der** Staat und **sein***e Einwohner*" ("The state and **its** inhabitants") the antecedent "Staat" ("state") is masculine (sg.) and so a "sein" base form is required for the possessive pronoun. The ending "e" in "seine" is needed because the noun following the possessive pronoun is plural ("Einwohner/inhabitants").

6.5.2 Pronoun Selection Task

The sample set consists of a random selection of 50 instances of "its" marked as anaphoric in ParCor. As "its" is uncommon in the TED Talks corpus, all 50 instances came from the EU Bookshop corpus.

Result	Phrase	Syntax
Pronoun match (number/gender + case)	15	9
Pronoun match (number/gender only)	8	10
Pronoun mismatch	24	28
Pronoun not translated (mismatch)	0	0
Other	0	0
Bad translation	1	1
Pronoun not required	2	2
Anaphoric but could not find antecedent	0	0
Unsure: may not be anaphoric	0	0
Total	50	50

Table 6.4: Pronoun selection task results for anaphoric possessive "its"

Additional guidelines are provided for "its":

- Select the relevant combination of number/gender of possessor and object. Select the case of the pronoun if the quality of the SMT output permits this.
- Select "Pronoun not required" if the translation does not require a pronoun.

These additional guidelines are both extensions to the original guidelines in Hardmeier (2014).

The annotator is presented with a table of options capturing the number/gender of the possessor vs. the number/gender of the object. To reduce the number of options, a separate set of check-boxes is provided for case options, including "case unknown", nominative, accusative, dative and genitive.

6.5.3 Results

The results of the pronoun selection task for 50 instances of anaphoric possessive "its" are presented in Table 6.4.

One reason for pronoun mismatches is that the systems often select an incorrect base form (determined by the number/gender of the antecedent or "possessor") for the pronoun, i.e. "ihr" when "sein" should be used, and vice versa.

The phrase- and syntax-based systems selected the incorrect base form for 34% (17/50) and 30% (15/50) of instances respectively.

6.5.4 Discussion

Again, the findings of Weiner (2014) echo those for "its" in this study. Weiner reports translation accuracy for "its" at 54.5% (news) and 71.4% (TED) for English-to-German translation, with sample sizes of fewer than 50 pronouns.

6.6 Relativizers

6.6.1 Translation Requirements

English relativizers may be explicit (*that-* and *wh-relativizers*), or implicit (*null-relativizers*). Both explicit and implicit relativizers may be translated as relative pronouns in German. As relative pronouns and their antecedents are often located close together, this represents a *local* sub-problem of the *anaphoric problem*.

6.6.2 Pronoun Selection Task

The sample set consists of a random selection of 50 instances of relativizers from the TED Talks corpus; 25 that- and 25 null-relativizers. The selection was semi-automatic, based on identifying relative clauses in the output of the Berkeley Parser (Petrov et al., 2006) and manually selecting those that contained a that- or null-relativizer.

As null-relativizers are implicit, there are no tokens in the English text to highlight. To keep this task in line with the others, symbols are inserted for the nulls, i.e. the "∅" in "The house ∅ Jack built". These are manually aligned to the corresponding token in the SMT output. (Null-relativizers that are not translated, are left unaligned.) Instead of a pronoun in the English text, the annotator is presented with an instance of "that" or a symbol representing the null-relativizer. Placeholders are included in the translation as normal.

The options table of check-boxes captures pronoun number/gender and case. It is similar to the table for "it", but with relative pronoun forms and options for "case unknown" and all four German cases.

	That		Null	
Result	**Phrase**	**Syntax**	**Phrase**	**Syntax**
Pronoun match (number/gender + case)	14	12	13	12
Pronoun match (number/gender only)	0	0	2	0
Pronoun mismatch	2	3	1	1
Pronoun not translated (mismatch)	4	3	5	6
Other	2	3	3	3
Bad translation	3	4	1	2
Pronoun not required	0	0	0	1
Anaphoric but could not find antecedent	0	0	0	0
Unsure: may not be anaphoric	0	0	0	0
Total	25	25	25	25

Table 6.5: Pronoun selection task results for relativizers

6.6.3 Results

The results of the pronoun selection task for 25 instances of that-relativizers and 25 null-relativizers are presented in Table 6.5.

Here we can observe that both systems are able to insert relative pronouns when a null-relativizer is encountered in the English source text, with a similar accuracy to the translation of that-relativizers. We might expect that translating an explicit source-language token would be easier (and more accurate) than inserting a token in the SMT output which has no explicit representation in the source language. However, this does not appear to be the case. The use of null-relativizers is not uncommon in English and it seems likely that the SMT systems have "learned" mappings between the use of a null-relativizer in English and a relative pronoun in German.

A similar bias to that observed for "it" (commonly translated as "es") is also observed for relativizers. Both that- and null-relativizers are commonly translated using the relative pronoun "die". As "die" is used for both feminine and plural referents, and in both the nominative and accusative case, alignments in the training data between that-relativizers (English) and the German relative pronoun "die" will be more common than alignments between that-relativizers and the other relative pronouns (e.g. "der/das" etc.). The resulting bias toward

translating that-relativizers as "die" is observed for the state-of-the-art systems. Both translate a that-relativizer as "die" in 61.9% (13/21) of instances in which a translation is provided, though not the same 13 of 21 instances (for both systems).

6.6.4 Discussion

When the antecedent is not a noun, i.e. "something" ("etwas"), "anything" ("alles" / "jedes" etc.) or "nothing" ("nichts"), then "was" should be used:

(6.5) Now , when I use the term miracle , I don 't mean something **that** 's impossible.

(6.6) Nun , wenn ich den Begriff Wunder verwenden , ich meine nicht etwas , **XXX** ist unmöglich .

A better translation of Ex. 6.5 would use "... nichts, was ..." instead of "... nicht etwas, was ...", but both options would require "that" to be translated as "was". As "was" is not provided as an option in the pronoun selection task, the annotator marked Ex. 6.6 (and others like it) as "Other". SMT systems must decide whether to use a relative pronoun that conveys the number/gender of the antecedent (i.e. der/die/das) or "was/wer/wo" (if the antecedent cannot be determined / there is no antecedent). As this decision depends on the antecedent, relative pronouns may therefore be treated as a more localised sub-set of anaphoric pronouns.

The translation of relativizers may also require a preposition preceding the relative pronoun:

(6.7) That 's the planet ∅ we live on .

(6.8) Das ist die Welt , **XXX** wir leben .

The correct translation of Ex. 6.7, which contains a null-relativizer (indicated by ∅), would be "Das ist die Welt, **in der** wir leben" (Lit: This is the world, in which we live). However, in the SMT output the preposition "in" is missing, and so the annotator was required to select the correct pronoun as if the preposition had been present.

In German, the preposition determines the case of the relative pronoun. Some prepositions always take the accusative case and others the dative case. *Two-way prepositions* (e.g. "in/auf/an") may take either case depending on whether

they express direction or destination (accusative) or a static position or situation (dative). The choice of preposition and therefore the case of the pronoun are determined by the verb of the clause. SMT systems could therefore also consider the translation of prepositions when translating relative pronouns.

6.7 Implications for SMT

To the problems outlined in Chapter 5, is added the problem of functionally ambiguous pronouns such as "it", for which the anaphoric and pleonastic forms both translate as "es" in German. These frequent alignments in the training data may also bias the likelihood that "it" is incorrectly translated as "es" (neuter), even if a feminine or masculine pronoun is required in German. SMT systems need to disambiguate the function of those pronouns with ambiguous surface forms so that each pronoun may be translated in an appropriate way. In the case of anaphoric pronouns it is necessary to use contextual information to overcome these biases in order to produce accurate pronoun translations.

Different types of pronouns have different translation requirements and therefore not all pronouns should be treated alike. For example, there is only one form for pleonastic pronouns in German ("es") and no agreement requirement, whereas anaphoric pronouns have many forms in German and are subject to pronoun-antecedent agreement constraints. However, the differences do not stop there. As highlighted in Sections 6.4.1, 6.5.1 and 6.6.1, different sub-types of anaphoric pronoun have different translation requirements in German. For example, the German translation of an instance of anaphoric "it" must agree with its antecedent, but an instance of *dependent* possessive "its" (as in "The state and **its** inhabitants") must agree with both the possessor (i.e. the antecedent) but also the object in possession. Given such differences, it would be unreasonable for discourse-aware SMT systems to model all anaphoric pronouns in the same way. Methods both for improving the translation of pronouns, and assessing the accuracy of their translations, should consider sub-classifying anaphoric pronouns according to their translation requirements. Reflexive pronouns, excluded from the analysis due to their infrequent use in the ParCor corpus, should be considered when working with genres in which their use is more frequent.

As observed in Section 5.2.3.1, the presence of relativizers in English text may lead to the insertion of a relative pronoun in German. Whilst SMT systems

appear to already perform quite well in terms of inserting relative pronouns when both that-relativizers (explicit) and null-relativizers (implicit) are encountered in English, performance is not prefect. Assuming that it is always appropriate to insert a relative pronoun when a relativizer is encountered, a method for successful translation would require:

- Automatic identification of null-relativizers, using constituent parses to identify relative clauses in English that do not contain an explicit relativizer
- Automatic disambiguation of instances of "that" which may function as either a relativizer or a complementizer in English. When translating "that" into German, a relative pronoun is used if the instance is a relativizer and "dass" is used if the instance is a complementizer
- Automatic identification of the antecedent of the relative pronoun to be inserted
- Automatic identification of whether a preposition is also required in the German translation, and if so, its form ("in/an/auf/mit/bei/zu" etc.). The choice of preposition affects the case of the relative pronoun, and is in turn governed by the verb of the clause

The method could take the form of rules within a rule-based system, an automatic post-editing framework for correcting SMT output, or a decoder feature. The aim would be to encourage, or in the case of post-editing to enforce, the use of an appropriate relative pronoun in the German MT output.

6.8 Conclusion

This chapter described an analysis of automated translation, using two state-of-the-art SMT systems. The analysis revealed that biases in the training data and incorrect selections of the base form for possessive pronouns (i.e. "ihr" vs. "sein" for "its") are both problems which SMT systems must overcome. For relative pronouns selecting the correct preposition is also important as it influences the case of the pronoun.

Possible directions for future work include further analyses of manual and automated translation and applying the knowledge that is gained to build pronoun-aware SMT systems. Initial efforts could focus on syntax-based SMT — lever-

aging information within target-side syntax trees constructed by the decoder, to encourage pronoun-antecedent agreement for intra-sentential anaphoric pronouns (i.e. "it/its" and relative pronouns).

Pronoun-aware SMT systems could also address translation of the ambiguous second-person pronouns "you" and "your". In English, they have both deictic and generic use, while in German, different forms are used ("Sie/du" vs. "man").

6.9 Bridging the Gap: From English-German to English-French

The annotation of the ParCor corpus (Chapter 4) and the corpus analyses (Chapters 5 and 6) were carried out for the English-German language pair. However, the annotation guidelines could also be used in the annotation of other languages. In the annotation of French texts, no changes to the guidelines are envisaged. However, the annotation of some other languages may require some adaptation. For example, if applied to pro-drop languages, such as Czech, the guidelines could be adapted to annotate instances of subject pro-drop. In the case of Czech, this could mean annotating the morphology of the verb (which captures the omitted pronoun). Likewise, the techniques used in the analyses of manual translation and state-of-the-art SMT output could be applied to other language pairs. In order to conduct analyses similar to those presented in this thesis, ParCor-style annotations are required for the source-language texts (and over the target-language text for the analysis of manual translation). The findings of the analyses are expected to differ depending on the language pair that is investigated.

From this point onwards, the focus of the thesis shifts to English-to-French translation. This reflects the choice of language pair for the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015). The techniques used in the automatic post-editing submission to the shared task (Chapter 7) and for the PROTEST test suite for pronoun evaluation (Chapter 8) rely on ParCor-style annotations over the source-language texts and consideration is given to the choice of target language. However, the techniques could be applied to other languages. Where relevant, the adaptation of the techniques to the English-German language pair is described for illustrative purposes.

Chapter 7

Automated Pronoun-focussed Post-editing for SMT

This chapter describes a submission to the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015), based on post-editing. The design of the automatic post-editing system, which handles anaphoric and non-anaphoric pronouns using separate rules, was influenced by the work in the preceding chapters. The post-editing system is a first attempt to explicitly model the translation of the functionally ambiguous pronoun "it" using knowledge obtained via analyses of translation in the ParCor corpus. This decision to divide pronouns into anaphoric vs. non-anaphoric groups follows the recommendation in Chapter 6, that both pronoun function in the source language and translation requirements in the target language should be considered.

This chapter aims to answer the question: Can pronoun translation be improved using simple rule-based post-editing based on this method of dividing pronouns into these two groups? Although the automatic post-editing system was unable to beat the official shared task baseline, important insights were gained from the process of building and analysing the system, and from its failure. These insights are detailed in Section 7.12, which in addition to the system and its analysis, provide one of the main contributions of this chapter.

In addition to describing the design of the system and its results on the shared task, the chapter presents a detailed analysis of system performance in which pronouns are divided into different categories, an analysis of the external tools used in the post-editing system, and the results of a detailed analysis of an oracle experiment using gold-standard ParCor-style annotations over the shared

task *DiscoMT2015.test* dataset. All of these analyses rely upon the provision of ParCor-style annotations over the English source-language texts in the *DiscoMT2015.test* dataset.

Despite its unsuitability for the evaluation of pronoun translation (cf. Section 2.5.1.1), BLEU is the dominant evaluation measure in SMT. BLEU scores are therefore provided throughout this chapter, in addition to pronoun scores computed using manual evaluation methods. BLEU scores also serve to confirm that the overall translation quality does not suffer as a result of post-editing.

Much of the content of this chapter is based on a system description paper published at the 2nd Workshop on Discourse in Machine Translation (DiscoMT) at EMNLP 2015 (Guillou, 2015). Extensions to the paper include the oracle experiment and the in-depth analysis of system performance, the external tools, and the oracle system.

7.1 Overview

The DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015) focusses on the translation of the English subject position pronouns "it" and "they" into French. Both pronouns serve multiple functions in English.

When translated into French, anaphoric pronouns must agree with their antecedent in terms of both number and grammatical gender. Therefore, selecting the correct pronoun in French relies on knowing the number and gender of the antecedent. This presents a problem for current state-of-the-art Statistical Machine Translation (SMT) systems which translate sentences in isolation.

As noted in Section 1.1.1, *inter-sentential* anaphoric pronouns, i.e. those that occur in a different sentence from their antecedent, will be translated with no knowledge of their antecedent. Pronoun-antecedent agreement therefore cannot be guaranteed. Even *intra-sentential* pronouns, i.e. those that occur in the same sentence as their antecedent, may lack sufficient local context to ensure agreement.

As outlined in Section 1.1.2 functional ambiguity poses a problem for SMT. The English pronoun "it" may be used as an anaphoric, pleonastic or event reference pronoun. The pronoun "they" may serve as either an anaphoric or generic pronoun. For each pronoun type, translations into French must meet different requirements.

The first half of this chapter describes an automatic post-editing approach

which applies two pronoun-specific rules to the output of an English-to-French phrase-based SMT system. One rule handles anaphoric pronouns and the other handles non-anaphoric (i.e. event reference and pleonastic) pronouns. The baseline SMT system, post-editing rules, and the external tools on which the rules rely to make judgements as to pronoun function, position, and antecedent (for anaphoric pronouns), are described in detail.

The advantage of a post-editing approach is that the translations of both pronouns and their antecedents (for anaphoric pronouns) are already known. There is therefore no need to keep track of this information within the decoder. Instead, the problem becomes one of identifying incorrectly translated pronouns and replacing them with predicted values based on information extracted from the source-language text. The aim is to leverage knowledge about the target language and through this maximise the number of changes that will improve the pronoun translations, whilst also attempting to minimise those that may have a detrimental effect.

The post-editing rules make use of information automatically obtained from the source-language text. The risk of doing this is that inaccurate information could lead to incorrect translations. As post-editing takes place after translation, the decoder and language model can no longer be relied upon to recover from bad decisions. However, due to the simplicity of the approach and encouraging results from Weiner (2014) for the English-German pair, post-editing appeared to be worth exploring.

As revealed by the results of the shared task (see Section 7.8), the post-editing approach performed poorly. However, so too did all of the other systems, none of which was able to beat the *official shared task baseline*, a simple phrase-based SMT system.

The second half of this chapter outlines an oracle experiment, conducted using ParCor-style manual annotations, to determine how well the post-editing approach performs when *perfect* conditions are assumed. Comparisons of the manual annotations and output of the external tools reveal the extent to which inaccurate dependency parsing, non-anaphoric "it" detection and coreference resolution, may contribute to errors. A detailed error analysis of the oracle experiment is then used to identify the extent to which the baseline SMT system and the post-editing rules contribute to poor performance.

Please note that this chapter contains references to two *baseline* systems.

The first is used to provide the SMT output used as a starting point for post-editing. The second is the *official shared task baseline*, created by the shared task organisers, against which participating systems are scored. For clarity, all references to the latter baseline system will appear in italic text, as "*official shared task baseline*".

7.2 Post-editing

Using the ParCor corpus annotations (Guillou et al., 2014) as a model, automated external tools are applied to the full text of each (sentence-split) source-language document in the dataset to extract the following information: Anaphoric vs. non-anaphoric pronouns, subject vs. non-subject position and the antecedent of each anaphoric pronoun. This information is then leveraged by two post-editing rules; one for anaphoric pronouns and one for non-anaphoric pronouns. These rules are automatically applied to the 1-best output of the baseline SMT system described in Section 7.3. The process for extracting source-language information and application of the post-editing rules is outlined in Figure 7.1 and described in Sections 7.4 and 7.5 respectively.

When compared to other work on pronoun translation, this work is most similar to the post-editing approach taken by Weiner (2014) for English-to-German translation. The method used by Weiner (2014) filters out pleonastic pronouns and concentrates on the translation of anaphoric pronouns. So too does the method of Le Nagard and Koehn (2010). The differentiation between pleonastic and anaphoric pronouns has been addressed by Novák et al. (2013) and Loáiciga and Wehrli (2015). The approach described in this chapter, however, is the first attempt to explicitly handle the translation of both anaphoric and non-anaphoric instances of "it" in a post-editing framework.

7.3 Baseline Machine Translation System

The baseline system used to produce the SMT output is of a similar design to that provided of the *official shared task baseline* system. It is a phrase-based system built using the Moses toolkit (Koehn et al., 2007) and trained/tuned using only the pre-processed (tokenised, lower-cased) parallel data provided for the shared task. Training, tuning and development test data are described in Table 7.1.

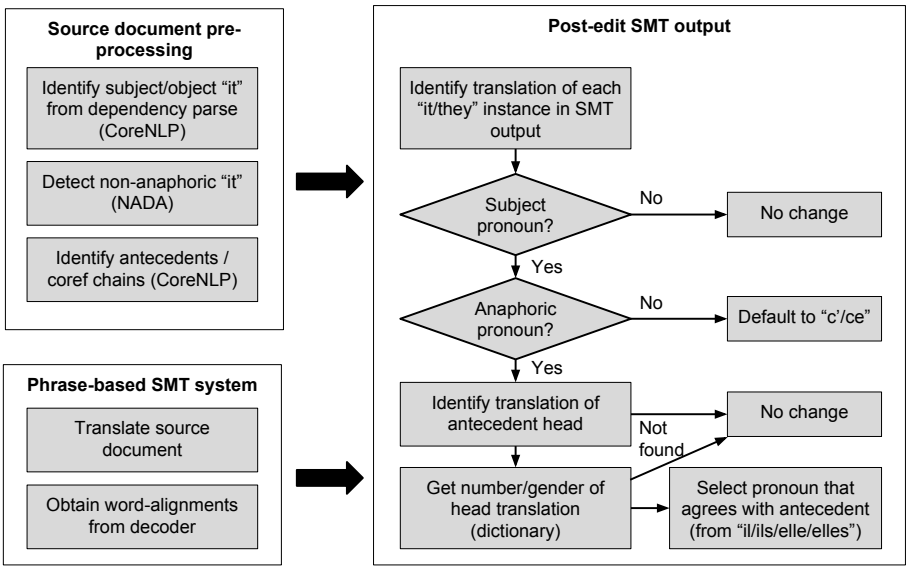


Figure 7.1: The post-editing process

Data	Description	Parallel Sentences	Monolingual Sentences
Training	TED, Europarl, News Commentary	2,372,666	
Tuning	dev2010 + tst2011	1,705	
Development	tst2010	1,664	
Development	tst2012	1,124	
Language model	TED, Europarl, News Commentary, News		33,869,133

Table 7.1: Baseline SMT system training, tuning and development data.

Word alignments are computed using Giza++ with *grow-diag-final-and* symmetrization, and with sentences restricted to 80 tokens or fewer (as Giza++ produces more robust alignments for shorter sentences). The maximum phrase length is set to 7. As memory and disk space are not a concern, sig-test filtering which prunes unlikely phrase pairs from the phrase table, is not used in training the baseline SMT system. Tuning is performed using MERT (Och, 2003) with an N-best list of 200, and using the dev2010+tst2011 data.

The language model is a 5-gram KenLM (Heafield, 2011) model, trained using *lmplz*, with modified Kneser-Ney smoothing and no pruning. The memory optimisations that were made for the *official shared task baseline*[1] are not replicated as they are not required. The language model uses the *probing data structure*; the fastest and default data structure for KenLM, it makes use of a hash table to store the language model n-grams.

By restricting the training data to sentences of 80 or fewer tokens, the baseline SMT system is trained on 27,481 fewer parallel sentences than the *official shared task baseline*. There are no other differences in the data used; for tuning, development-testing or language model construction.

The baseline SMT system scores nearly one BLEU point higher than the *official shared task baseline* for the IWSLT 2010 (34.57 vs. 33.86) and 2012 (41.07 vs. 40.06) test sets. BLEU scores were calculated using the case-insensitive, *multi-bleu* perl script provided in the Moses toolkit.

The decoder is set to output word alignments, which are used later for automatic post-editing.

7.4 Extracting Source-language Information

Guided by the ParCor annotation scheme, the following information is automatically extracted from the source-language text:

- Position: Subject or non-subject ("it" only)
- Function: Anaphoric or non-anaphoric (i.e. pleonastic / event reference, for "it" only)
- Antecedent: For anaphoric pronouns only

[1] The *official shared task baseline* was provided as part of the shared task resources.

The first step is to identify whether the pronoun appears in subject or non-subject position. The pronoun "it" may be used in either position, unlike "they" which is always a subject position pronoun. When translating into French it is necessary to ensure that each instance of "it" is correctly translated, with different French pronouns used depending on the position that the pronoun fills. Instances of "it" are categorised as being either subject- or non-subject position pronouns using the dependency parser provided as part of the Stanford CoreNLP tool[2]. Subject position pronouns are those that participate in an *nsubj* or *nsubjpass* dependency relation.

The next step is to determine the function of each instance of "it". NADA (Bergsma and Yarowsky, 2011) is used as it considers the entire sentence, unlike the pleonastic sieve in the Stanford coreference resolution system (Lee et al., 2011), which uses only fixed expressions to identify pleonastic "it". Instances of "it" with a NADA probability below a specified threshold are treated as non-anaphoric, and those above, as anaphoric. Here, a non-anaphoric pronoun is either an event reference or pleonastic pronoun; a finer distinction cannot be made using currently available tools. The NADA threshold is set to 0.41 (see Section 7.6).

For instances of "it" identified as anaphoric, and all instances of "they", the pronoun's nearest non-pronominal antecedent is extracted using the coreference resolution system (Raghunathan et al., 2010; Lee et al., 2011) provided in the Stanford CoreNLP tool. To avoid falsely identifying coreference chains across document boundaries, the source-language text is split into documents prior to coreference resolution. Full coreference chains are retained in case the nearest antecedent is not translated by the baseline SMT system.

NADA and CoreNLP were run on tokenised, but not lower-cased data, in order to ensure parser accuracy. The tokenisation and sentence segmentation is the same as that used in the pre-processed data distributed for the shared task. The CoreNLP tool was run with the following annotators: *tokenize*, *ssplit*, *pos*, *lemma*, *ner*, *parse* and *dcoref*. The following parameters were set to true: *tokenize.whitespace* and *ssplit.eolonly*.

[2]Stanford CoreNLP version 3.3.1: http://nlp.stanford.edu/software/corenlp.shtml

7.5 Automatic Post-Editing Rules

Automatic post-editing is applied to the 1-best output of the baseline SMT system described in Section 7.3. The process makes use of information extracted from the source-language text (see Section 7.4) and the word alignments output by the decoder.

For each source-language pronoun, one of two post-editing rules is applied, depending on whether the pronoun is identified as anaphoric or non-anaphoric. The rules are outlined in Figure 7.1 and are described in detail in the following sections.

7.5.1 Anaphoric Rule

This rule is applied to all instances of "they" and subject position "it" that are identified as anaphoric, both inter- and intra-sentential. *Cataphoric* pronouns, where the pronoun appears before its antecedent, are very rare in TED Talks data (Guillou et al., 2014) and are ignored for the sake of simplicity. Instances of non-subject position "it" are excluded as the focus of the shared task is on subject position pronouns only. Target-language pronoun forms are predicted using the projected translation of the *head* of the nearest non-pronominal antecedent.

On the source-language side:

1. Identify the nearest non-pronominal antecedent
2. Identify the antecedent head word (provided by CoreNLP for each antecedent)
3. Using word alignments output by the decoder, project source-language pronoun and antecedent head positions to the SMT output

On the target-language side (SMT output):

4. If no antecedent can be found for the pronoun, do not attempt to replace its translation. (It may be non-anaphoric but not detected by NADA)
5. For all other pronouns, use the word alignments to identify the translations of the pronoun and antecedent head. N.B. the pronoun and/or the head of its antecedent may not be translated

6. Extract the number and gender of the antecedent head translation via a dictionary of French nouns extracted from the Lefff (Sagot, 2010) and augmented by entries from dict.cc[3]

7. If the antecedent head word is aligned to multiple words in the translation select the right-most noun (this should be the head in most cases)

8. If the antecedent head translation **is a noun**[4]:

   (a) Predict "elle" for feminine, singular; "il" for masculine, singular

   (b) Predict "elles" for feminine, plural; "ils" for masculine, plural

   (c) If the antecedent is split-reference of the format **N and N**, split it into two nouns. If both are feminine, predict "elles", otherwise predict "ils"

9. If the antecedent head translation **is not a noun** (i.e. not in the dictionary) or is not translated:

   (a) Traverse further back through the coreference chain and repeat from *step 5*

   (b) If the antecedent head is not translated, apply a default value. If the source-language pronoun is translated as a pronoun, but not "il/elle" (for "it") or "ils/elles" (for "they"), predict "il" for "it" and "ils" for "they". If the pronoun is not translated, do nothing as the SMT system may have correctly learned to drop a pronoun

10. If the pronoun in the baseline SMT output and the predicted translation disagree, the post-editing rule replaces the translation in the baseline SMT output with the predicted value

This method allows for the prediction of a plural pronoun for cases where an English singular noun is translated into French using a plural noun (and vice versa). For example, "vacation" is singular in English but may be translated as "vacances" (plural) in French. The method also caters for cases where singular "they" is used in English. For example we may refer to single person as "they"

[3]dict.cc: www.dict.cc

[4]If the word is hyphenated and not in the dictionary, look up the right-most part, which should be the head.

if that person's gender is unknown: "If I find the person who did this, **they** will pay for it". In this example, if "person" is translated as "personne" (feminine) in French, the pronoun "they" (plural) should be translated as "elle" (feminine, singular). Cases where singular "they" is used are relatively rare, when compared to plural uses of "they".

7.5.2 Non-Anaphoric Rule

This rule is applied to instances of subject position "it" that are identified as non-anaphoric, i.e. those with a NADA probability below the specified threshold. It does not apply to instances of "they".

The first step is to identify the translation of the pronoun (using the word alignments). The translation that should appear in the post-edited SMT output is then predicted.

1) Translation is an event reference / pleonastic pronoun:
As NADA does not appear to distinguish event reference and pleonastic pronouns (i.e. both are considered equally non-anaphoric; see Section 7.6) it is not straightforward to predict a correct translation for non-anaphoric "it". The French pronoun "ce" may function as both an event reference and a pleonastic pronoun, but "il" is used only as a pleonastic pronoun. All instances of "it" translated as "ce/c'/il" are left as they are in the SMT output. Changing them may do more harm than good and would be performed in an uninformed manner. The hope is that these pronouns, or at least the pleonastic ones, may be correctly translated using local context.

2) Translation is another pronoun:
If an instance of "it" is translated as a pronoun outwith the set "ce/c'/il", it will be corrected to the default "ce" (or "c' " if the next word in the baseline SMT output starts with a vowel or silent "h"). The French pronouns "ce/c'/cela/ça" may be used as neutral pronouns, referring to *events*/actions/states or general classes of people/things, and "il/ce/c'/cela/ça" may be used as impersonal pronouns, marking the subject position but not referring to an entity in the text, i.e. *pleonastically* (Hawkins et al., 2001). "ce/c'/cela/ça" may all be used as either pleonastic or event reference pronouns. "ce" is selected as the default as it oc-

curs most frequently in the training data, suggesting common usage. There are some cases in which only "il" should be used as the impersonal pronoun, such as expressions of time. These are not easy to detect and are therefore ignored.

3) Translation is not a pronoun:
If an instance of "it" is translated using something other than a pronoun, it is not replaced. This may also indicate that the pronoun has been dropped.

4) No translation:
There is no provision for handling cases where a pleonastic or event reference pronoun may in fact be required but was dropped in the baseline SMT output. Automated tools that can separate pleonastic and event reference instances of "it" are not available (at least, not for English) and inserting a pronoun might not be the correct thing to do in all cases.

If the pronoun in the baseline SMT output and the predicted translation disagree, the post-editing rule replaces the translation in the baseline SMT output with the predicted value.

7.6 Setting the NADA Threshold

NADA returns a probability between 0 and 1, and the decision as to whether an instance of "it" is anaphoric can be made by setting a threshold for this probability. The NADA documentation suggests a general threshold value of 0.5; for probabilities over this value the pronoun is said to be referential (i.e. anaphoric) and for those below this value, that it is non-referential. However, different threshold values may be appropriate for different genres[5].

The TED-specific NADA threshold was set using the manual ParCor (Guillou et al., 2014) annotations over the TED Talks portion of the corpus. NADA was run over the English TED Talks in ParCor and the probabilities it assigned for each instance of "it" were compared with the pronoun type labels (i.e. anaphoric/pleonastic/event reference) in the ParCor annotations.

There are 61 instances of "it" marked as pleonastic in the ParCor annotations. Looking at *all* 133 instances of "it" in the ParCor TED Talks for which

[5]TED Talks are considered out-of-domain. NADA was trained using the Penn Treebank and Google N-Grams corpus.

		ParCor Label	
		Pleonastic	**Not** Pleonastic
NADA	Non-referential	37	75
	Referential	24	400

Table 7.2: NADA scores vs. ParCor labels at a threshold of 0.41

their NADA probabilities fall below 0.5, there are a mixture of pleonastic, event reference, and *anaphoric with no explicit antecedent* pronouns. These could acceptably be treated as non-referential. However, there are also a number of anaphoric pronouns that fall into this range and it would be unacceptable to treat these as non-referential. Setting the threshold is a trade-off between precision and recall. Whatever threshold is set, there will be both false positives and false negatives (see Table 7.2).

Figure 7.2 displays a comparison of the NADA scores vs. the ParCor pronoun type labels at various thresholds of NADA. The comparison is between true positives and false negatives for both pleonastic and event reference pronouns. Although NADA is designed to identify non-referential (i.e. pleonastic) "it", the non-anaphoric post-editing rule which relies on the NADA score is used for both pleonastic and event reference pronouns. The point at which lines of the graph, representing true positives and false negatives, intersect perhaps represent the best thresholds for NADA. For pleonastic pronouns this is between 0.2 and 0.3. For event reference pronouns it is close to 0.7. In an attempt to strike a balance the NADA threshold was selected, by manual inspection, as 0.41, after which there appears to be a noticeable increase in the number of anaphoric pronouns mis-identified as non-referential by NADA.

At a threshold of 0.41, 37 (60.66%) pronouns marked as pleonastic in ParCor are correctly identified and 24 (39.34%) are not. 37 pronouns marked in ParCor as event reference pronouns are correctly identified as non-referential and 35 anaphoric pronouns (of which 4 have no explicit antecedent) are misidentified as non-referential.

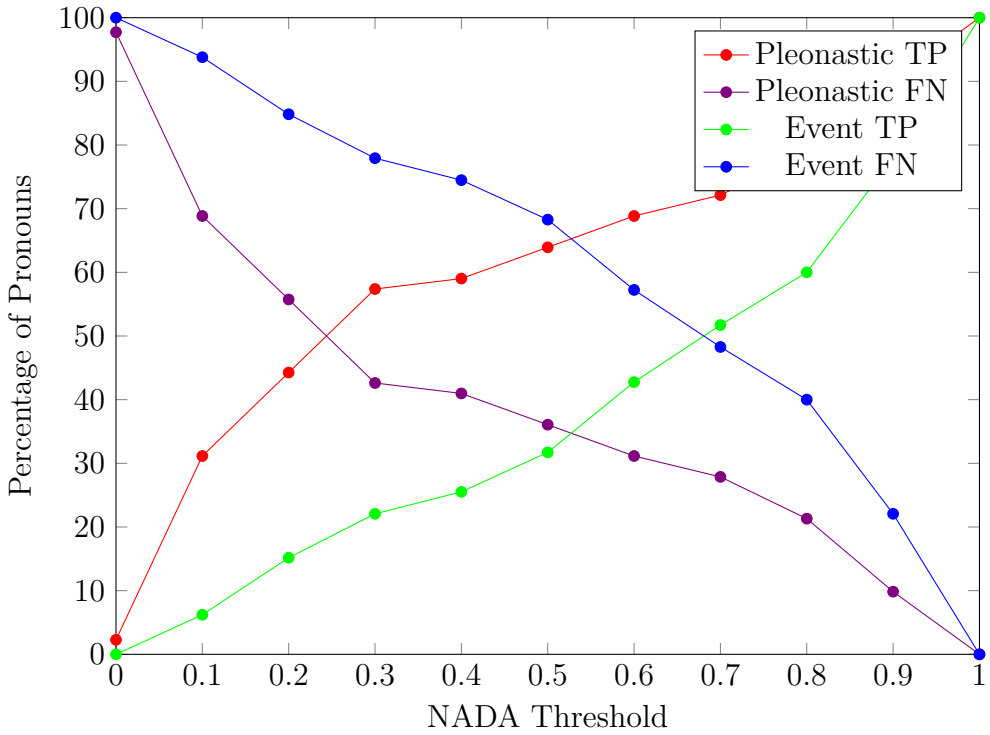


Figure 7.2: NADA scores vs. ParCor labels: True positives and false negatives for pleonastic and event reference pronouns at different NADA thresholds

7.7 Post-editing Changes

The shared task test set (*DiscoMT2015.test*) contains 307 instances of "they" and 809 instances of "it". Automated pre-processing of the source-language texts identifies 581 instances of "it" as subject position pronouns and 228 as non-subject position pronouns (for which no change will be made). Of the 888 instances of "it" and "they" identified as subject position pronouns, the translation of 316 are changed in the baseline SMT output by the post-editing rules. 303 changes are applied to pronouns identified as anaphoric (36 "they" and 267 "it") and 13 to pronouns identified as non-anaphoric. The pronoun changes are summarised in Table 7.3. 10 pronouns were not translated by the baseline SMT system, and as such, were not considered for replacement. It is not possible to determine whether these changes lead to a better, worse, or equivalent translation of the pronoun as this would require manual evaluation over the complete test set. As part of the shared task evaluation, the translation of 210 pronouns from the test set were manually evaluated for each participating system. Statistics on the number

Pronoun type	Form	Before	After	Count
Non-anaphoric	it	ça	ce/c'	7
Non-anaphoric	it	cela	ce/c'	3
Non-anaphoric	it	elle	ce/c'	1
Non-anaphoric	it	le	ce/c'	1
Non-anaphoric	it	on	ce/c'	1
Anaphoric	it	il	ils	3
Anaphoric	it	il	elle	51
Anaphoric	it	il	elles	3
Anaphoric	it	elle	il	17
Anaphoric	it	elle	ils	1
Anaphoric	it	le/l'	il	3
Anaphoric	it	on	il	1
Anaphoric	it	ça	il	10
Anaphoric	it	ça	ils	2
Anaphoric	it	ça	elle	5
Anaphoric	it	cela	il	6
Anaphoric	it	cela	elle	3
Anaphoric	it	cela	elles	1
Anaphoric	it	ce/c'	il	84
Anaphoric	it	ce/c'	ils	5
Anaphoric	it	ce/c'	elle	68
Anaphoric	it	ce/c'	elles	4
Anaphoric	they	ils	elles	32
Anaphoric	they	elles	ils	4
Total				316

Table 7.3: Automated post-editing changes

of better/worse translations according to the manual evaluation are provided in Section 7.9.1.

The most frequent changes are "c'/ce" → "il" (84), "c'/ce" → "elle" (68), "il" → "elle" (51), and "ils" → "elles" (32). The change "c'/ce" → "il/elle" takes place due to the decision to use gendered translations of all instances of "it" identified as anaphoric (even if "c'/ce" might also have been an acceptable translation). Biases in the training data may account for some of the other changes. For example, the change "ils" → "elles" may result from the common alignment of "they" to "ils" which arises due to the rule in French that "ils" is used unless all of the antecedents are feminine (in which case "elles" is used). This may result in more masculine pronouns requiring replacement with a feminine pronoun than vice versa.

The changes "il" → "elle" and "ils" → "elles" are made to conform with the gender of the translation of the head of the anaphoric pronoun's antecedent. The post-editing rules also allow for changes from singular to plural (and vice versa) and from one number and gender to another. For example in translating "it" → "vacation" the anaphoric rule would allow for an instance of "il" (masculine, sg.) in the baseline SMT output to be changed to "elles" → "vacances" (feminine, pl.).

7.8 Results

The official shared task results taken from Hardmeier et al. (2015), are displayed in Table 7.4 and show that none of the participating systems was able to beat the *official shared task baseline*, a simple phrase-based SMT system.

The results include a number of scores from pronoun-specific and general purpose metrics. *Accuracy with OTHER* is the primary score. It measures the number of correctly translated pronouns per the human judgements in the manual evaluation of 210 pronouns via a pronoun selection task (similar to that described in Section 6.3). It is computed over the set of pronoun tokens and includes pronouns marked as "other" by the annotator – if the annotator marks a pronoun as "other" and the MT output contains a pronoun outwith the set of options given to the annotator (ce, ça/cela, il, ils, elles and elles), this is counted as a *good* translation. *Accuracy without OTHER* is a similar measure, but excludes pronouns labelled as "other". *ProsF* is the micro-averaged F-score for all pro-

	Acc with OTHER	**Acc w/o OTHER**	**ProsF**	**AutoP**	**AutoR**	**AutoF**	**BLEU**	**NIST**	**TER**	**METEOR**
Official Shared Task Baseline	0.676	0.630	0.699	0.371	0.361	0.366	37.18	8.04	46.74	60.05
auto-postEDIt	0.543	0.473	0.523	0.329	0.276	0.300	36.91	7.98	46.94	59.70
UU-Hardmeier	0.581	0.525	0.580	0.347	0.333	0.340	32.58	7.66	49.04	57.50
IDIAP	0.657	0.617	0.711	0.346	0.333	0.340	36.42	7.89	48.18	59.26
Its2	0.419	0.339	0.396	0.184	0.187	0.188	20.94	5.96	60.95	47.90
UU-Tiedemann	0.643	0.590	0.675	0.386	0.353	0.369	36.92	8.02	46.93	59.92
A3-108	0.081	0.081	0.188	0.054	0.045	0.049	4.06	2.77	88.49	25.59

Table 7.4: Official Shared Task Results

nouns, based on the results of the manual evaluation. *AutoP*, *AutoR* and *AutoF* are the automatic pronoun precision, recall and F-score scores from the BLEU-inspired pronoun metric defined by Hardmeier and Federico (2010). Scores from the following general-purpose automatic metrics are also provided: *BLEU*, *NIST*, Translation Error Rate (*TER*), and *METEOR*.

An overview of the results and a brief comparison of the participating systems are provided in Hardmeier et al. (2015).

The official results report a BLEU score of 36.91 for the post-edited SMT output (*auto-postEDIt*). This score is lower than the *official shared task baseline* system (37.18), comparable with *System E* (36.92), and higher than the other competing systems. However, the post-editing system outperformed only two of the other five competing systems in terms of the *accuracy* measures, reinforcing the general opinion within the *discourse in SMT* community, that BLEU is a poor measure of pronoun translation performance. The *accuracy with OTHER* measure reveals that the post-edited SMT output contains correct translations for only 114/210 pronoun instances, according to human judgements.

There is a small decrease of 0.36 BLEU between the baseline system used to provide SMT output and the post-edited version for the *DiscoMT2015.test* dataset (38.83 vs. 38.47 respectively, as calculated using case-insensitive multi-bleu[6]).

7.9 Analysis of System Performance

The scores reported in Section 7.8 relate to the overall performance of a system and do nothing to guide future development. In order to improve upon the design of a given system its developers may wish to know which pronouns are translated well and which are translated poorly.

This section describes a detailed analysis of the performance of the automatic post-editing system. The analysis was conducted using the human judgements for the set of 210 pronouns included in the manual evaluation, and ParCor-style annotations over the English texts in the *DiscoMT2015.test* dataset. The annotations were used to facilitate the analysis of the translation of pronouns according to their functional *type*. Anaphoric pronouns were further sub-categorised accord-

[6]The official shared task BLEU scores appear to have been calculated using a different method.

Pronoun type	Count
Anaphoric	1,524
Cataphoric	8
Event reference	360
Extra-textual reference	110
Pleonastic	123
Speaker reference	1,880
Addressee reference	727
Total	**4,732**

Table 7.5: Pronoun distribution by type for the *DiscoMT2015.test* dataset

ing to whether they are in subject or non-subject position (for "it"), and whether they are used as singular or plural pronouns (for "they").

The *DiscoMT2015.test* dataset (Hardmeier et al., 2016) comprises 12 TED Talks; their English transcriptions and French translations. Each English transcription was manually annotated following the ParCor annotation guidelines outlined in Chapter 4. (Annotation of the texts was coordinated by the shared task organisers.) The annotations were produced and released *after completion of the shared task* and as such were not available to participants during the development of their systems.

Corpus statistics for the *DiscoMT2015.test* dataset are displayed in Table 7.5.

7.9.1 Analysis of Post-editing using Human Judgements and ParCor Annotations

An examination of the human judgements for the 210 pronoun tokens that were evaluated manually, revealed that the post-editing process makes many mistakes. 34 instances were worsened by post-editing and only 9 improved. The remaining instances were neither better nor worse following post-editing: 70 instances were correct in both the baseline and post-editing output and 97 were incorrect. Translation accuracy differed for "it" and "they". For "it" 32 instances were judged to be correct vs. 60 incorrect. The opposite was observed for "they", with 47 instances judged to be correct vs. 14 incorrect. (Instances marked as "other" or "bad translation" cannot be commented upon further and are excluded from

the counts). The poor translation of "it" could be due to the method used to identify anaphoric and non-anaphoric instances (no such method was required for "they"), differences in coreference resolution accuracy for "it" and "they", or something else entirely.

An example of a translation that is improved though post-editing is presented below. In the English input the pronoun "it" refers to "car" (Ex. 7.1). In the baseline MT output (Ex. 7.2) the antecedent is translated as "voiture" [fem. sg.] and the pronoun as "il" [masc. sg.] and as such pronoun-antecedent agreement does not hold. In the post-edited output (Ex. 7.3) the translation of the pronoun is changed to "elle" [fem. sg.] to agree with the antecedent. The human judgement for this pronoun confirms that "elle" is the correct translation.

(7.1) it was in the early days of gps , the **car** knew where **it** was , and it would give audio instructions to the driver , when to turn right , when to turn left and so on . [Input]

(7.2) C' était dans les premiers jours de GPS , la **voiture** savait où **il** était , et il donnerait des instructions audio au conducteur , quand de tourner à droite , quand de tourner à gauche et ainsi de suite . [Baseline]

(7.3) Il était dans les premiers jours de GPS , la **voiture** savait où **elle** était , et il donnerait des instructions audio au conducteur , quand de tourner à droite , quand de tourner à gauche et ainsi de suite . [Post-edited]

Post-editing changes can also lead to a degradation in performance. Consider the instance of anaphoric "it" in Ex. 7.4. The coreference resolution system incorrectly identified the antecedent as "substitute" – in the ParCor-style annotations the antecedent is "homosexuality" which appears in an earlier sentence (not shown here). The baseline system (Ex. 7.5) translates "it" as "elle" [fem. sg.], which is the correct translation per the human judgement. However, in Ex. 7.6 the anaphoric post-editing rule changes the pronoun translation to "il" to agree with the translation of "substitute" as "substitut" [masc. sg.].

(7.4) it is a pathetic little second-rate **substitute** for reality – a pitiable flight from life . as such , **it** deserves no compassion , it deserves no treatment as minority martyrdom , and it deserves not to be deemed anything but a pernicious sickness . " [Input]

(7.5) C' est un peu pathétique médiocre **substitut** à la réalité , une fuite pitoyable de la vie . En tant que telle , **elle** ne mérite aucune compassion , il ne mérite pas de traitement comme martyre minoritaire , et mérite de ne pas être considérée comme une maladie pernicieuse ." [Baseline]

(7.6) Elle est un peu pathétique médiocre **substitut** à la réalité , une fuite pitoyable de la vie . En tant que telle , **il** ne mérite aucune compassion , il ne mérite pas de traitement comme martyre minoritaire , et mérite de ne pas être considérée comme une maladie pernicieuse . " [Post-edited]

However, in many cases the post-editing changes have no effect at all. The pronoun translation may remain unchanged by the post-editing rules or the rules may substitute one poor translation for another. For example, the baseline system translated "it" in Ex. 7.7 as "il" (Ex. 7.8) and this was changed to "elle" by the post-editing system (Ex. 7.9). However, the human judgement is that "ce" or "cela" would have provided a better translation. As the post-editing system identified the instance of "it" as anaphoric (with the antecedent automatically identified as "nature"[7]), and the anaphoric rule makes no provision for translation using impersonal pronouns, the system was unable to produce a correct translation.

(7.7) but i was also struck by the burdensome **nature** of such mutual secrecy . depression is so exhausting . it takes up so much of your time and energy , and silence about it , **it** really does make the depression worse . [Input]

(7.8) Mais j' ai également été frappé par la **nature** d' une telle mutuelle lourd secret . La dépression est tellement épuisant . Il reprend beaucoup de votre temps et de l' énergie , et le silence , **il** est réellement la dépression pire . [Baseline]

(7.9) Mais j' ai également été frappé par la **nature** d' une telle mutuelle lourd secret . La dépression est tellement épuisant . Elle reprend beaucoup de votre temps et de l' énergie , et le silence , **elle** est réellement la dépression pire . [Post-edited]

The human judgements alone are not sufficiently informative for a complete analysis. Combined with the information from the ParCor-style annotations, they

[7]In the ParCor-style annotations the antecedent is identified as "silence" and not "nature".

	anaphoric													total
	it							*they*						
	intra		inter		no specific ant		–	intra		inter		no specific ant	–	
	subj	non-subj	subj	non-subj	subj	non-subj	subj	sg	pl	sg	pl	pl	pl	
Tokens	*19*	*17*	*39*	*8*	*2*	*1*	*1*	*2*	*20*	*3*	*31*	*5*	*1*	*149*
Correct	*5*	*1*	*20*	*2*	*1*	*0*	*0*	*2*	*19*	*3*	*21*	*5*	*0*	*79*
Incorrect	*14*	*16*	*19*	*6*	*1*	*1*	*1*	*0*	*1*	*0*	*10*	*0*	*1*	*70*

Table 7.6: Anaphoric pronouns: Evaluation of **baseline** SMT system performance using 149 anaphoric pronoun tokens and ParCor-style annotations

	it				
	event	pleonastic	extra textual	cataphoric	total
Tokens	*36*	*20*	*4*	*1*	*61*
Correct	*15*	*16*	*1*	*0*	*32*
Incorrect	*21*	*4*	*3*	*1*	*29*

Table 7.7: Non-anaphoric pronouns: Evaluation of **baseline** SMT system performance using 61 non-anaphoric pronoun tokens and ParCor-style annotations

become much more useful. The annotations allow for a more granular categorisation of the English pronouns, providing a clearer breakdown of which pronouns the system translates well. The pronouns are sub-categorised using the ParCor-style annotations:

- Functional *Type*: Anaphoric, pleonastic, event reference ("it" only)
- *Position*: Subject vs. non-subject "it"
- *Agreement*: Singular vs. plural "they"
- *Antecedent*: Explicit vs. non-explicit, i.e. whether an antecedent is specified or not (anaphoric pronouns only)

The manual evaluation results for the baseline SMT system performance, according to the above sub-categorisation, are displayed in Tables 7.6 and 7.7 (for 149 anaphoric and 61 non-anaphoric pronouns, respectively). The figures in

	anaphoric													total
	it							*they*						
	intra		inter		no specific ant		–	intra		inter		no specific ant	–	
	subj	non-subj	subj	non-subj	subj	non-subj	subj	sg	pl	sg	pl	pl	pl	
Tokens	*19*	*17*	*39*	*8*	*2*	*1*	*1*	*2*	*20*	*3*	*31*	*5*	*1*	*149*
Correct	*4*	*1*	*12*	*2*	*0*	*0*	*0*	*2*	*17*	*2*	*21*	*5*	*0*	*66*
Incorrect	*15*	*16*	*27*	*6*	*2*	*1*	*1*	*0*	*3*	*1*	*10*	*0*	*1*	*83*
Δ*Baseline*	*-1*	*0*	*-8*	*0*	*-1*	*0*	*0*	*0*	*-2*	*-1*	*0*	0	0	

Table 7.8: Anaphoric pronouns: Evaluation of **post-editing** system performance using 149 anaphoric pronoun tokens and ParCor-style annotations, compared against the performance of the baseline SMT system

	it				
	event	pleonastic	extra textual	cataphoric	total
Tokens	*36*	*20*	*4*	*1*	*61*
Correct	*7*	*12*	*1*	*0*	*20*
Incorrect	*29*	*8*	*3*	*1*	*41*
Δ*Baseline*	*-8*	*-4*	*0*	*0*	

Table 7.9: Non-anaphoric pronouns: Evaluation of **post-editing** system performance using 61 non-anaphoric pronoun tokens and ParCor-style annotations, compared against the performance of the baseline SMT system

these tables are used as the basis for comparison, throughout this chapter, with improvements over the baseline SMT system reported as delta values.

The manual evaluation results for the post-editing system performance, according to the same sub-categorisation, are displayed in Tables 7.8 and 7.9. 86 pronouns were translated correctly, with 124 translated incorrectly. Performance was compared with that of the baseline SMT system, with improvement / degradation of performance provided as delta values. The results indicate that the post-editing system is poor at translating instances of event reference "it" and significantly better at translating instances of anaphoric "they" than anaphoric "it". 19 instances of anaphoric "it" were translated correctly vs. 68 incorrectly, compared with 47 correct vs. 15 incorrect for anaphoric "they". This difference

in translation accuracy of anaphoric "it" and "they" is statistically highly significant, $p < .00000001$[8].

Central to the translation of anaphoric "it" and "they", is the problem of selecting the correct grammatical gender in the target language. However, the pronoun "they", which is typically used to refer to a plural entity, is sometimes also used to refer to a singular entity, for example, when referring to a person of unknown gender. As described in Section 7.5.1, the post-editing rule for anaphoric pronouns accommodates this, allowing for the prediction of a singular pronoun for cases where an English plural noun is translated into French using a singular noun, and vice versa.

That the post-editing system is better at translating "they" than "it" may be due to the possible variation in use of the pronouns in English and in translating each pronoun in French. The pronoun "it" can be used as either a subject or a non-subject position anaphoric pronoun, and as a pleonastic or event reference pronoun. Using external tools to identify these properties may lead to errors, which result in poor translation of "it". In contrast, the pronoun "they" may be used as a subject position anaphoric pronoun, or as a generic (i.e. *anaphoric with no specific antecedent*). No external tool was used to identify instances of generic "they", but these are rather rare, with only 14 instances labelled as such in the *DiscoMT2015.test* dataset. Therefore, there are more decisions to be made for instances of "it", and making wrong decisions could lead to errors.

The poor translation of instances of event reference and anaphoric "it" may be linked to the accuracy of the external tools, as described in Section 7.9.2.

The results in Tables 7.6 to 7.9 were generated using the same judgements used to calculate *Accuracy without OTHER*. That is, the scores from those pronouns marked as "other" in the manual evaluation are considered to be incorrect. The "other" category is a *catch-all* category for pronouns not in the set defined for the manual evaluation and as such, is not clearly defined and does not allow for further examination. According to the implementation of the *Accuracy with OTHER* measure, tokens for which the pronoun is untranslated in the SMT output are rewarded if they are marked as "other" in the manual evaluation. The *Accuracy without OTHER* metric is therefore preferred.

Another criticism is that the set of 210 pronoun tokens contained 26 instances of "it" marked as non-subject position in the ParCor-style annotations. The focus

[8]Calculated using Fisher's exact test.

of the shared task was on subject position pronouns, which brings into question the method used to select the 210 pronoun tokens. The inclusion of these non-subject position pronouns in the set, may have had an adverse effect on the performance of the post-editing system, and other participating systems.

7.9.2 Accuracy of External Tools

The post-editing system makes use of external tools to disambiguate between subject and non-subject instances of "it" (CoreNLP dependency parser), between anaphoric and non-anaphoric instances of "it" (NADA) and to identify the antecedents of anaphoric pronouns (CoreNLP coreference resolution). The accuracy of these tools directly affects the quality of the changes made to the baseline SMT output, and should therefore be assessed. The accuracy of NADA has already been reported with respect to the ParCor annotations (see Section 7.6). In Section 7.9.2.2 the same procedure was applied to ascertain the accuracy on the *DiscoMT2015.test* dataset. An assessment was also made of the automatically identified antecedent spans output by CoreNLP compared to the antecedent spans in the manual annotations, and the position labels for instances of "it" generated using a dependency parser based method compared to the manual annotations. The results of these investigations are reported below.

7.9.2.1 CoreNLP Dependency Parser

The CoreNLP dependency parser is used to detect subject vs. non-subject instances of "it". Those instances that form part of an *nsubj* or *nsubjpass* relation are considered to be subject position pronouns. This automated method was compared with the manual annotations of subject vs. non-subject position. There are 809 instances of "it" in the *DiscoMT2015.test* dataset, of which 308 were manually labelled as non-anaphoric instances, and one was not annotated. The results of the comparison for the remaining 500 (anaphoric) instances of "it" are presented in Table 7.10. The comparison shows that the method used to automatically identify subject vs. non-subject instances of "it" is reasonably accurate.

7.9.2.2 NADA

Section 7.6 discussed setting the threshold for NADA (using the ParCor corpus) for post-editing. In this section, the consequences of setting this threshold

		Automated Identification	
		Subject	Non-subject
Manual Label	Subject	315	17
	Non-subject	16	152

Table 7.10: Comparison of manual annotation and automated identification of subject vs. non-subject position "it"

		DiscoMT2015.test Label	
		Pleonastic	**Not** Pleonastic
NADA	Non-referential	71	88
	Referential	52	597

Table 7.11: NADA scores vs. DiscoMT2015.test labels at a threshold of 0.41

are considered. Using the ParCor-style annotations over the *DiscoMT2015.test* dataset, the pronoun *type* labels for instances of "it" were compared with the NADA scores.

At a threshold of 0.41 (the value used in the post-editing experiment), 71 (57.72%) of the 123 pronouns marked as pleonastic in the*DiscoMT2015.test* dataset manual annotations were correctly identified and 52 (42.28%) were not (see Table 7.11). Of the 88 instances of "it" identified by NADA as being non-referential, 39 were marked in the manual annotations as event reference, 4 as extra-textual reference, 2 as cataphoric, and 43 as anaphoric pronouns (of which 4 have no explicit antecedent). One instance of "it" was not annotated and is therefore excluded from the table. The accuracy of NADA as measured against the ParCor-style annotations is similar to that presented in Section 7.6. For the ParCor corpus 60.66% of pleonastic pronouns are correctly identified using NADA at a threshold of 0.41, compared with 57.72% for the *DiscoMT2015.test* dataset.

Results	Pronoun	Count
Match	it	90
	they	84
Partial match (some heads match)	it	0
	they	9
Mismatch	it	189
	they	113
Antecedent not automatically identified	it	182
	they	84
Total		**751**

Table 7.12: Comparison of manual annotation and automated identification of antecedents

7.9.2.3 CoreNLP Coreference Resolution

The post-editing method relies on knowledge of the antecedent head for anaphoric pronouns. Comparing the automatically extracted antecedent head (CoreNLP) and the head of the antecedent span in the manual annotations, is therefore sufficient. All instances of "it" and "they" linked to an antecedent in the manual annotations were considered for comparison.

Antecedent heads were not manually annotated in the ParCor-style annotations. Instead these were automatically extracted from the manually annotated antecedent spans using the dependency parser in CoreNLP, and manually checked. The results of the comparison are presented in Table 7.12. The heads were compared using string match, and were recorded as *matches* and *mismatches*. In many cases, the Stanford coreference resolution system failed to find an antecedent. This was also recorded.

The results suggest that the Stanford coreference resolution system, included in CoreNLP, is rather poor at identifying the antecedents of anaphoric pronouns, at least when used in a domain other than the one that it was trained on. It could therefore be expected that this component would have a large impact on the performance of the post-editing system.

	anaphoric													total
	it							*they*						
	intra		inter		no specific ant		–	intra		inter		no specific ant	–	
	subj	non-subj	subj	non-subj	subj	non-subj	subj	sg	pl	sg	pl	pl	pl	
Tokens	*19*	*17*	*39*	*8*	*2*	*1*	*1*	*2*	*20*	*3*	*31*	*5*	*1*	*149*
Correct	*5*	*1*	*14*	*2*	*1*	*0*	*0*	*2*	*17*	*2*	*22*	*5*	*0*	*71*
Incorrect	*14*	*16*	*25*	*6*	*1*	*1*	*1*	*0*	*3*	*1*	*9*	*0*	*1*	*78*
Δ*Baseline*	*0*	*0*	*-6*	*0*	*0*	*0*	*0*	*0*	*-2*	*-1*	*1*	*0*	*0*	
Δ*Post-edit*	*1*	*0*	*2*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	

Table 7.13: Anaphoric pronouns: **Oracle** performance using 149 anaphoric pronoun tokens and ParCor-style annotations, compared against the performance of the baseline SMT system

7.10 Oracle Experiment

7.10.1 Oracle System

The shared task manual evaluation was framed as a pronoun selection task in which the pronouns in the MT output are obscured, and a human annotator was asked to select the pronoun that best fits the translation. The pronoun selection task is described in Hardmeier et al. (2015) and Hardmeier (2014), and used in the analysis of state-of-the-art SMT output in Chapter 6. In the case of the post-editing system, changes were applied directly to the pronoun translations produced by a baseline SMT system. This means that the same human judgements may be re-used in an oracle experiment, in which the output of NADA and CoreNLP are replaced with the ParCor-style annotations over the *DiscoMT2015.test* dataset. Antecedent heads were automatically extracted from the antecedent spans using the dependency links output by the Stanford dependency parser (in CoreNLP). The heads were then checked manually with minor adjustments made as necessary. The purpose of the oracle experiment was to ascertain how well the post-editing system would perform if we were to assume *optimal conditions*. The performance of the post-editing system using the ParCor-style annotations is shown in Tables 7.13 and 7.14. Performance was compared with that of the baseline SMT system and the post-editing system,

	it				
	event	pleonastic	extra textual	cataphoric	total
Tokens	*36*	*20*	*4*	*1*	*61*
Correct	*9*	*16*	*1*	*0*	*26*
Incorrect	*27*	*4*	*3*	*1*	*35*
Δ*Baseline*	*-6*	*0*	*0*	*0*	
Δ*Post-edit*	*2*	*4*	*0*	*0*	

Table 7.14: Non-anaphoric pronouns: **Oracle** performance using 61 non-anaphoric pronoun tokens and ParCor-style annotations, compared against the performance of the baseline SMT system

with improvement / degradation of performance provided as delta values.

Using the ParCor-style annotations over the *DiscoMT2015.test* dataset leads to only a marginal improvement over the original post-editing experiment (see Tables 7.8 and 7.9). 97 of the 210 pronouns were translated correctly by the oracle system, compared with 86 by the original post-editing system. These numbers were computed using the *Accuracy without OTHER* metric. The gains came from the following pronoun categories:

- Anaphoric (no specific antecedent), subject "it": 1
- Anaphoric, inter-sentential, subject "it": 2
- Anaphoric, intra-sentential, subject "it": 1
- Event reference "it": 2
- Pleonastic "it": 4
- Anaphoric, inter-sentential (plural) "they": 1

The BLEU score for the oracle system is presented in Table 7.15. Scores for the baseline SMT system and the post-editing process (using automated identification of position, anaphoric pronouns and antecedents) are provided for comparison. (These are the same scores reported in Section 7.8.) All scores were calculated using case-insensitive multi-bleu. A very small increase in BLEU score was observed for the oracle system, as compared with post-editing using the output of automated tools (+0.16).

Experiment	BLEU
Baseline SMT system	38.83
Post-editing (automated tools)	38.47
Post-editing (oracle: manual annotations)	38.63

Table 7.15: BLEU scores for the oracle system, baseline SMT system, and post-editing using the output of automated tools

Given the errors introduced by the external tools used in the post-editing pipeline (see Sections 7.9.2.1, 7.9.2.2 and 7.9.2.3) it is surprising that the oracle shows little improvement. It is therefore necessary to consider other factors, such as baseline SMT quality, coverage of the dictionary of French nouns and the post-editing rules themselves.

7.10.2 Error Analysis

The output of the oracle system in Section 7.10.1 was used as the starting point for further investigation. The assumption here is that by using the manual ParCor-style annotations, issues arising from the use of external tools may be avoided. That is, we may assume perfect labelling of subject vs. non-subject "it", anaphoric vs. non-anaphoric "it" and pronoun antecedent, and focus instead on issues with the rules themselves, and the baseline SMT system.

The investigation was framed as an *error analysis* of the 113 pronouns translated incorrectly by the oracle system (out of the 210 pronouns from the manual evaluation). Error categories were derived using the ParCor style annotations, knowledge of how and when the rules are applied, the translation of each pronoun in the MT output and the human judgements. The error categories and the number of pronoun translations that fall into each category are presented in Table 7.16.

The categories are defined as follows. Pronouns marked as "bad translation" could not be evaluated by the manual evaluator as the quality of the MT output was too poor. Pronouns marked as "other" in the manual evaluation were excluded as the "other" category is too broad, including other pronouns and untranslated pronouns. No further analysis of these categories was possible.

Error category	Count
Poor quality SMT output ("Bad translation", excluded)	9
Human judgement: "Other" (excluded)	14
Pronoun not translated by the SMT baseline	14
Pronoun not translated by the SMT baseline as a pronoun	6
Correct pronoun in translation but part of a hyphenated word	3
Event reference/Pleonastic pronoun left as "il" by rule (human judgement: "ce")	12
Event reference pronoun translated as "ce" (human judgement: incorrect)	10
Anaphoric pronoun translated with wrong number/gender	17
Gendered translation for anaphoric pronoun but human judgement is non-gendered	27
Gendered translation for extra-textual pronoun but human judgement is non-gendered	1
Total	113

Table 7.16: Error Analysis of incorrect translations by oracle system

A source-language pronoun may not be translated (i.e. no word alignment exists), or its translation may not contain a pronoun (e.g. source-language pronoun aligned to a verb in the translation). It may be acceptable, or indeed necessary to omit the pronoun from the translation, but the manual judgements do not tell us this. Future manual evaluation tasks might also include a *pronoun not required in target* option, so as to distinguish between valid omissions and errors. The comparison of MT translations and human judgements might also consider hyphen-splitting. For example, in the case that a pronoun is translated as "est-il" and the human judgement is that the translation should be "il", such tokens could be marked as correct.

Other error categories are specific to the post-editing rules themselves. Those instances of "it" identified as event reference or pleonastic are handled by a single rule for non-anaphoric pronouns. According to this rule, the translations "il/c'/ce" were left unchanged and all other translations were replaced with "ce". The use of this rule resulted in some instances of event reference / pleonastic pronouns left as "il", where "ce" would have been the correct translation, and other event reference pronouns changed to "ce" resulting in an incorrect translation. The anaphoric rule, used for all pronouns identified as anaphoric, replaces the pronoun translation with a gendered pronoun which agrees with the French translation of the English antecedent head. Using this rule, an incorrect pronoun

may be selected. This may happen if the French translation of the antecedent head is ambiguous in the dictionary of French nouns, e.g. if the same word form can take different genders. Note also that the pronoun is only changed if the English antecedent head is translated by the baseline SMT system and if the French translation of the antecedent head is listed in the dictionary of French nouns (from which number and gender are extracted). These factors can also lead to incorrect pronoun translations. It is also possible that the human evaluator identifies a different antecedent for the pronoun to the one in the ParCor-style annotations. Without knowledge of what antecedent the human evaluator had identified in English, no further information can be gleaned from the analysis.

The anaphoric rule selects only gendered French pronouns. In 27 instances, the human evaluator suggested that the translation of an English anaphoric pronoun should be a non-gendered pronoun, e.g. “ce” or “cela”.

7.11 Discussion

7.11.1 Limitations of Post-editing

The post-editing system performed poorly in the DiscoMT 2015 shared task on pronoun translation. As with all other participating systems it failed to beat the *official shared task baseline* — a simple phrase-based SMT system. The external tools used in the post-editing process are imperfect and introduce errors. However, even when manual annotations are used in place of the external tools, system performance is still below that of the baseline SMT system. The error analysis of the oracle system revealed other possible causes of error, including the baseline SMT system, the post-editing rules and possible human error in the manual evaluation task.

The analysis of the oracle system’s performance suggests that the post-editing rules themselves are the cause of many of the poor translations. It is clear that it is not sufficient to use a single “non-anaphoric” rule covering both pleonastic and event reference pronouns. As revealed in Table 7.14, even when the ParCor-style annotations are used to identify pronoun function, the translation of event reference pronouns is particularly poor. (Only 9 out of 36 event reference pronouns are correctly translated vs. 16 out of 20 pleonastic pronouns.) The provision of a separate rule for event reference pronouns is not currently possible as no tool

exists for detecting instances of event reference "it". Whilst NADA appears to detect some event reference pronouns, it is an accidental consequence of its inability to distinguish a pleonastic from an event reference pronoun. Using "ce" as the default translation for both pleonastic and event reference pronouns also causes problems. The use of "ce" was judged as unsuitable for a number of cases of event reference "it" during the manual evaluation.

The anaphoric rule, which considers only the use of gendered pronouns in the French translation, is also insufficient. The error analysis of the oracle system revealed that 27 of the 113 errors were the result of using a gendered pronoun where the human annotator believed that a non-gendered one would be correct (see Table 7.16). In these cases the annotators suggested that the use of the non-gendered French demonstrative pronouns "ce", "ça" or "cela" would be more appropriate when translating an instance of "it" than the gendered French personal pronouns "il" and "elle".

As the post-editing rules affect only pronouns, agreement issues may occur. For example, if the baseline SMT system outputs "**ils** sont partis" ("**they**[masc.] have left") and the post-editing rules replace "ils" with "elles", the verb "partis" should also be replaced: "**elles** sont *parties*" ("**they**[fem.] have left"). Agreement issues could be addressed within a dependency-parser-based post-editing framework such as the Depfix system for Czech (Mareček et al., 2011; Rosa, 2014).

While post-editing rules could potentially be written to insert a pronoun in the SMT output where one is syntactically required in the target language, or to delete a pronoun for syntactic or stylistic reasons, this was not done in the current system.

Despite encouraging results from Weiner (2014), post-editing using rules does not appear to be a good choice for the pronoun translation task. The approach may also be difficult to extend to other languages which are less well provisioned in terms of parsers and coreference resolution systems, or for which baseline SMT quality is poor. Considering the more complex problem of English-to-German translation, which also requires selection of grammatical case (determined by the syntactic role of the pronoun in the sentence), the problems a rule-based method faces are further compounded. Rather, the decision of which pronoun to use in the MT output might be better made using a classifier, as in the post-editing approach of Luong et al. (2015).

7.11.2 Limitations of Evaluation

System performance for the shared task was measured in terms of both automated and manual evaluation metrics, with results from the manual evaluation providing the definitive system ranking. This reliance on manual evaluation could be alleviated, to some extent, by the provision of multiple reference translations. Assuming that the type of each pronoun in the source language and what, if anything, it refers to is known, pronoun translations may be assessed as follows. The translations of non-anaphoric instances of "it" could be compared to the range of translations provided in the multi-reference set. In the case of anaphoric pronouns, the translation of both the pronoun and the antecedent head would need to be considered, as pronoun-antecedent agreement needs to hold for a correct translation. The pronoun-antecedent pairs in the MT output would need to be compared to those in each reference translation in the set, in turn. However, test sets with multiple references are rare and expensive to produce. The next chapter focusses on what can be done to reduce manual evaluation efforts when only a single reference translation is available.

7.12 Insights

A number of insights were gained during the development and analysis of the automatic post-editing system. These may be useful to researchers considering applying post-editing, or other methods, to the problem of pronoun translation.

First and foremost, one could argue that the overall approach was flawed, given the poor performance of the post-editing system at the DiscoMT 2015 shared task. However, there are both positive and negative aspects to the design of the system. Whilst the system performs poorly for some pronouns such as event reference and anaphoric "it", even in the oracle scenario, it performs reasonably well for some other pronouns: Pleonastic "it" and anaphoric "they". In evaluating pronoun translation, it is important to look at where the system does well and where it does poorly. This will be explored in greater detail in Chapter 8.

Directly replacing pronouns in the MT output with substitute values is not sufficient for a complete system. For example, changing only the pronoun may result in conflicts with the surrounding words i.e. agreement with the verb (an example of which is provided in Section 7.11.1).

The results of the oracle system serve to highlight the fact that poor coreference resolution is not the only problem that affects the performance of pronoun translation in SMT. In the case of the automatic post-editing system, the design of the rules and their strict application to the SMT output were identified as the biggest factors in the failure of the system.

A more refined implementation of the system would need to provide more complex rules. At minimum these rules would need to cater for the detection and translation of event reference "it" and to allow for the translation of anaphoric "it" using *either* demonstrative or personal pronouns. Another point to consider is whether to directly apply these rules to MT output, or to incorporate them in the decoder. Given the disfluencies that arise from substituting only the pronouns, it seems wise to incorporate rules into the decoder. On the other hand, in the case of this automatic post-editing system, there is some benefit to be gained from initially working outside the decoder as the same human judgements may be re-used if the rules are refined within the current framework in which changes are made to the SMT output at the pronoun level.

7.13 Conclusion

This chapter presented an automatic post-editing system, submitted to the DiscoMT 2015 shared task on pronoun translation. The post-editing approach makes use of two pronoun-specific rules applied to the output of a baseline English-to-French phrase-based SMT system. One rule handles anaphoric pronouns, the other handles non-anaphoric pronouns.

Before extending this work to develop new rules or applying the technique to other language pairs, it is important to first understand where the post-editing method performs well and where it performs poorly. A detailed analysis of the post-editing changes as compared with the human judgements from the manual evaluation provides a logical first step. Limitations of both the external tools and the post-editing rules were assessed, and an oracle experiment was conducted using manual ParCor-style annotations in place of external tools. An error analysis of the oracle system revealed problems with the rules, and shortcomings of the manual evaluation with respect to what can be analysed in detail.

Chapter 8

Pronoun-focussed Evaluation

This chapter describes the development of the *PROTEST* test suite for pronoun translation. The test suite comprises a set of 250 hand-selected pronoun tokens categorised according to the range of problems that SMT systems face when translating pronouns, and an automatic evaluation script for assessing translation accuracy. The pronoun tokens are selected from the *DiscoMT2105.test* dataset.

There are three main aims to the work described in this chapter. The first is to provide a set of hand-selected pronoun tokens that may be used to assess and compare pronoun translations by MT systems, and to make this available to the research community. The second is to discover the extent to which automatic evaluation can be reliably applied to the assessment of pronoun translation. The third is to apply the automatic evaluation to the output of MT systems as a first step to better understanding how well those system perform when translating different categories of pronoun. The main contributions of this work are the *PROTEST* test suite and its application to the automatic evaluation of the systems submitted to the DiscoMT 2015 shared task on pronoun translation.

The content of this chapter is based on Guillou and Hardmeier (2016). The design of the test suite, including the pronoun categories and the methodology behind the automatic evaluation, was outlined in collaboration, with equal contribution from both authors. While I worked on the manual selection of the pronoun tokens and the implementation of the automatic evaluation script, Christian Hardmeier was responsible for running the automatic evaluation over the DiscoMT 2015 shared task systems, and producing statistics based on the translation accuracy of each system for each pronoun category.

The use of the PROTEST test suite is not limited to the evaluation of SMT

output; it may be used to evaluate translation by any MT system provided word alignments can be produced between the source-language text and its translation. Where relevant the term *MT* is therefore used in preference to *SMT* throughout this chapter.

8.1 Overview

Evaluation poses a particular problem for researchers interested in pronoun generation in MT. Owing to the cost and difficulty of manual evaluation (including manual post-editing based methods as a means to assess MT quality), MT researchers rely on automatic evaluation metrics such as BLEU (Papineni et al., 2002) to guide their development efforts. Most automatic metrics assume that overlap of the MT output with a human-generated reference translation may be used as a proxy for correctness. In the case of anaphoric pronouns, this assumption breaks down. If the pronoun's antecedent is translated in a way that differs from the reference translation, a different pronoun may be required: Using one that matches the reference translation may in fact be incorrect (cf. Section 5.3.2).

This shortcoming of existing automatic evaluation metrics is widely recognised (cf. Section 2.5.1.1). Hardmeier (2015a) therefore suggests using a test suite composed of carefully selected pronoun tokens which can be checked individually using an automatic evaluation script, in addition to an aggregate measure over a complete test set (cf. Section 8.3.2), to evaluate pronoun correctness.

The following sections describe the categorisation of pronouns according to ParCor-style annotations over the *DiscoMT2015.test* dataset, the selection of 250 pronoun tokens, the design of the evaluation and its application to the output of the six systems submitted to the DiscoMT 2015 shared task on pronoun translation.

8.2 Test Set Annotations

The test suite is built on top of an existing corpus: The *DiscoMT2015.test* dataset which is described in Section 7.9. This test set contains English transcriptions of 12 TED conference talks (and their French translations), selected in such a way that the texts include a reasonable number of instances of some less frequent pronouns. Since the dataset provides complete texts, rather than a collection

of isolated sentences or passages, any MT system being tested has access to full document context for each example, which is essential for discourse-aware translation.

The English source texts were annotated manually[1] for *reduced coreference* in the style of the ParCor corpus with coreferential pronouns labelled according to their function in the text (i.e. their *type*), additional type-specific attributes labelled for some pronoun types, and coreferential links created between anaphoric pronouns and their nearest non-pronominal antecedent(s). These annotations form the basis for the categorisation and selection of pronoun tokens, and the evaluation procedure.

Whilst gold-standard test sets such as the OntoNotes corpus (Weischedel et al., 2011) exist for the coreference resolution task, they are not suitable for assessing pronoun translation in MT. In particular, monolingual gold-standard test sets lack reference translations, and there exist neither monolingual nor multilingual test sets that provide the additional pronoun type-specific features used to define the fine-grained categories for the test suite pronouns.

8.3 Test Suite Design

8.3.1 Selection of Pronoun Tokens

The distribution of pronoun types in the *DiscoMT2015.test* is presented in Table 8.1. Anaphoric and cataphoric pronouns have been sub-split into *intra-sentential* (pronoun and antecedent appear in the same sentence) and *inter-sentential* (pronoun and antecedent appear in different sentences). For anaphoric pronouns, two additional sub-types are considered: Those *linked to another pronoun* (no NP antecedent was found) and those with *no specific antecedent*. As pronoun-antecedent agreement must hold in French, the translation accuracy of such pronouns would be difficult to assess.

The aim is to extract pronoun tokens that provide good coverage over the range of different pronoun *types* and surface *forms* (e.g. “it”, “they” etc.) and represent the different problems that MT researchers must consider:

- Anaphoric [it/they]

[1] N.B. This is the same annotated dataset used in the analysis of system performance for the automatic post-editing system in Section 7.9.

Pronoun type	**Count**
Anaphoric	
inter-sentential	761
intra-sentential	644
linked to another pronoun	26
no specific antecedent	93
Cataphoric	
inter-sentential	1
intra-sentential	7
Event reference	360
Extra-textual reference	110
Pleonastic	123
Speaker reference	1,880
Addressee reference	727
Total	**4,732**

Table 8.1: Pronoun distribution by type for the *DiscoMT2015.test* dataset

  - Inter-sentential vs. intra-sentential
  - Subject vs. non-subject [it only]
  - Singular vs. plural "they"
  - Referring to group nouns (e.g. "company" could be referred to as singular/plural)
- Event reference [it]
- Pleonastic [it]
- Addressee Reference [you]
  - Generic vs. deictic
  - Singular vs. plural [deictic only]

The categorisation of pronouns follows the taxonomy introduced for the annotation of the ParCor corpus (see Section 4.4).

At the top level, we wish to distinguish between pronouns with different *functional types* in English and requiring different translations in French. For example, the pronoun "it" requires the use of different French pronouns depending on

whether it functions as an anaphoric, pleonastic or event reference pronoun[2]. At the lower level, we wish to consider other differences exhibited by pronouns of the same *type* and *surface form*. For anaphoric pronouns we wish to distinguish between inter- and intra-sentential pronouns, which given the current framework of sentence-by-sentence translation, pose different challenges to MT systems. (The same division of inter- and intra-sentential pronouns was employed in the analysis of the pronoun "it" by state-of-the-art SMT systems in Section 6.4.2.) We also wish to consider position and number as different French pronouns will be required when translating subject vs. non-subject instances of "it", or plural vs. singular "they". For deictic instances of "you" number affects the French translation: "tu" or "vous" may be used to refer to a single person (depending on formality), but when referring to more than one person "vous" must be used. Generic instances of "you" may be translated as "on" (similar to English "one").

The selection provides a balance both in terms of the number of pronoun tokens for each category[3], and of the expected French translation. For example, the selection contains equal numbers of instances of "it/they" that one might expect to be translated as masculine vs. feminine pronouns (i.e. by looking at the reference translation). The selection process also considers instances of singular pronouns that may be translated as plural in French and vice versa.

Another option would have been to define category sizes in proportion to the number of pronouns for each category in the source-language texts. However, if we wish to build MT systems that are linguistically competent, they should demonstrate an understanding of the linguistic system, rather than mere frequencies. The aim is to be able to assess the accuracy of an MT system in translating both commonly occurring source-language pronouns and rare ones (e.g. singular "they").

One use case for the test suite is to complement automatic evaluation with manual evaluation. This motivates the restriction of the set of pronoun tokens to a number that is manageable for manual evaluation and inspection. A number of pronoun groups are therefore excluded. For example groups for which very few instances exist in *DiscoMT2015.test* or for which translation is perhaps less problematic. The following pronoun groups are excluded from the test suite:

[2]"ce" may function as both an event reference or pleonastic pronoun; "il" may be used as both a pleonastic or anaphoric pronoun.

[3]N.B. For some categories, few instances exist in *DiscoMT2015.test*.

- *Reflexive* pronouns, which are very infrequent in TED Talks
- *Relative* pronouns. Those that are marked for number and gender in French (e.g. "lequel" [masc. sg.], "lesquelles" [fem. pl.], etc.) are infrequent in TED talks, and those that are not marked (e.g. "qui", "que", "dont" and "quoi"), are unambiguous as they are in English
- *First-person* (i.e. speaker reference) pronouns, which are unambiguous and do not change between speakers
- *The third-person pronouns "he/she"* which are unambiguous in both French and English
- *Possessive adjectives* ("your/their" etc.), which in French agree with the noun that follows (and not the antecedent)

The set of pronoun forms included in the test suite is restricted to "it", "they" and "you". One could argue for the inclusion of other pronoun forms within some of the pronoun categories. For example "this/that" which like "it" can be used as anaphoric or event reference pronouns, or "your" which requires a similar deictic/generic disambiguation approach as for "you". However, the translation problems are similar to those posed by "it" and "you". In order to keep the number of pronoun tokens manageable when it comes to manual evaluation, exclusions must also be made in terms of pronoun surface forms. New pronoun token sets may be created in the future when the performance of systems, over the current set, has improved.

Pronoun tokens were automatically pre-selected according to the above categories using the ParCor-style annotations over the source-language text, and word alignments[4] between the source language and reference texts. The word alignments allow for the selection of English pronoun tokens according to their expected (i.e. reference) translation. The final selection of pronoun tokens is confirmed following manual examination. The distribution of pronoun tokens selected for the test suite is presented in Table 8.2.

[4]Word alignments were computed using a combination of Giza++ (with standard settings) and fast_align for sentences exceeding the Giza++ limit of 100 tokens.

Pronoun	Type	Primary sub-type	Secondary sub-type	Count
it	anaphoric	intra-sentential	subject	25
it	anaphoric	intra-sentential	non-subject	15
it	anaphoric	inter-sentential	subject	25
it	anaphoric	inter-sentential	non-subject	5
they	anaphoric	intra-sentential	–	25
they	anaphoric	inter-sentential	–	25
they	anaphoric	singular	–	15
it/they	anaphoric	refer to group noun	–	10
it	event reference	–	–	30
it	pleonastic	–	–	30
you	addressee reference	generic	–	20
you	addressee reference	deictic	singular	15
you	addressee reference	deictic	plural	10
Total				**250**

Table 8.2: *DiscoMT2015.test* pronouns selected for the test suite

8.3.2 Automatic Evaluation

The test suite also includes an automatic script to check the translations of the pronoun tokens in the output of an MT system against those in the reference translation. The results are presented in terms of *matches* and *mismatches*. *Matches* are measured in terms of overlap between the reference token and the MT output string. For anaphoric pronouns, the script verifies that both the translation of the pronoun and (each) antecedent head match those in the reference translation. Those cases where the pronoun translations match but the antecedent head translations do not are considered *mismatches*. For all other pronoun types, only the translation of the pronoun is considered. The evaluation script outputs the count of pronoun tokens correctly translated by the MT system (i.e. *matches*), for each category, as well as an accuracy score for each category and for the test suite as a whole.

The tokenisation of the source text is relevant to evaluation and systems may tokenise the source text in ways other than that in *DiscoMT2015.test*. It is therefore necessary to supply the tokenised source text in addition to the MT output and the word-alignments between the source text and MT output. The sentence-

internal word-position of each pronoun token (and antecedent head where relevant), and its MT translation are identified.

While the accuracy score output by the evaluation script can be used as an aggregate metric, the main advantage of the test suite over existing metrics is the possibility to study the system's performance on individual pronoun tokens. The evaluation script outputs a list of *mismatches* between the MT and reference translations to be checked manually – the pronoun translations (and antecedent heads) may be valid alternative translations of the source, not present in the reference. This need for manual evaluation is the driving factor behind restricting the test suite to only a sub-set of the pronouns in the *DiscoMT2015.test* dataset.

The evaluation method is similar to that of the ACT metric (Hajlaoui and Popescu-Belis, 2013) for discourse connectives. In contrast with ACT which considers only the translations of the discourse connectives themselves, the evaluation of anaphoric pronouns requires that two elements be considered: The pronoun and its antecedent(s). The ACT metric also incorporates a list of valid translations for each discourse connective. This dictionary is automatically composed. For pronouns, the manual verification of translations in MT output forms a similar method for obtaining a similar set of valid alternative translations. Another difference lies in what may be evaluated automatically. The ACT metric automatically scores a discourse connective translation as incorrect if it differs from the reference (or list of alternatives) or if it is missing from the MT output (but present in the reference). Corresponding scenarios for pronoun translation require manual evaluation as more variation, including the omission of pronouns, is possible.

8.3.3 Use Cases

There are two main use cases for which PROTEST was designed. The first is for the *manual evaluation* of those translations that did not match the reference in the automatic evaluation. By combining automatic and manual evaluation, researchers may obtain a complete evaluation of one or more systems. In addition to the number of matches for each pronoun category, the evaluation script outputs a list of mismatches between the MT and reference translations to be checked manually – the pronoun translations (and antecedent heads) may be valid alternative translations of the source, not present in the reference. Consider

the following example:

(8.1) I have a **bicycle** . **It** is red . [Original English]

(8.2) J'ai un **vélo** . **Il** est rouge . [Reference]

(8.3) J'ai une **bicyclette** . **Elle** est rouge . [MT output]

Here the English anaphoric pronoun "it" in Ex. 8.1 refers to "bicycle". The reference translation translates "it" as "il" (masc. sg.) which agrees with the translation of "bicycle" ("vélo" [masc. sg.]). In the MT output, a valid alternative translation is produced, with "elle" referring to "bicyclette" (both fem. sg.). This translation, although correct, does not match the reference and would therefore be referred for manual evaluation.

During development, translations found in the MT output could be added to the set of translations accepted by the evaluation script once they have been manually verified for correctness. This would serve to expand the set of valid translations against which future MT output could then be scored, thereby reducing manual evaluation effort over time. Obviously, doing this will make it impossible to compare the scores output by the evaluation scripts with values reported by other groups[5], but it enables a much more precise evaluation of progress for the developer's internal use.

The second use case is for the *measurement of the incremental progress of a system (or systems)*, where it may be sufficient to simply compare the results of the automatic evaluation, for example where a new system extends a baseline, or provides a small incremental change over an existing system. In such scenarios, it may be sufficient to check whether performance of the new system improves for the desired pronoun categories, or at least does not show a degradation in performance over the baseline system.

When comparing the performance of two systems, it would be useful to identify whether the difference in performance is statistically significant. For this we would need to have first conducted a complete evaluation so that we know for each pronoun in a category whether the translation by each system was correct or not. A correct translation would be assigned a score of 'one' and an incorrect

[5]Unless researchers were to report two sets of results so as to facilitate comparisons: Results using only the *DiscoMT2015.test* reference translations, and using the reference translations *plus* their own expanded set of valid translations.

translation a score of zero. Given that the number of pronouns per category is rather small, between 5 and 25 pronouns, one option might be to use Fischer's exact test which is suitable for use with small sample sizes. To compute the test we would construct a 2x2 contingency table containing the number of correct vs. incorrect translations for each of the two systems, A and B. However, we might not wish to compute the test for each pronoun category, for each run of a new system. We might therefore rather reverse the test so as to ascertain how many more correct translations system B needs to produce than system A so that we could say with 95% confidence that system B is overall better than A. We might also consider t-tests and their reverse. Whichever test is selected, reverse tests should be computed for each pronoun category in PROTEST. Additionally, pronoun categories could also be combined so that statistical significance could be computed for the set of anaphoric "it" instances, the set of anaphoric "they" instances and the set of deictic "you" instances, for example.

8.4 Evaluation Results

This section demonstrates the application of the test suite to the results of the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015) for English-to-French. The shared task focussed on the translation of subject position "it" and "they" which are represented by pronoun categories in PROTEST. The evaluation also provides results for the translation of other PROTEST categories which fall outside the scope of the shared task: Non-subject position "it" and the addressee reference pronoun "you". As noted in Section 7.8, all participating systems were beaten by a simple phrase-based SMT baseline according to the official evaluation. PROTEST is used to gain a better understanding of these results.

The results presented in this section were computed by Christian Hardmeier and the brief discussion that follows arose from an investigation conducted by both authors.

Table 8.3 shows the number of matches in the test suite for all participating systems, including the *official shared task baseline* (labelled as "Baseline" in the table).

The results reveal, subject to confirmation following manual evaluation of the mismatches, that some of the participants do outperform the *official shared*

	anaphoric								event	pleonastic	addressee reference		
	it				*they*				*it*	*it*	*you*		
	intra		inter		intra	inter	sing.	group			generic	deictic	
	subj.	non-subj.	subj.	non-subj.								sing.	plural
Tokens	*25*	*15*	*25*	*5*	*25*	*25*	*15*	*10*	*30*	*30*	*20*	*15*	*10*
Baseline	8	1	**11**	1	12	**12**	8	6	**15**	**18**	**13**	**9**	9
auto-postEDIt	**10**	**6**	6	**2**	**13**	11	8	**7**	6	11	12	8	**10**
UU-Hardmeier	**10**	3	7	**2**	11	8	**11**	5	13	**18**	12	8	**10**
IDIAP	8	3	**11**	1	11	8	6	6	11	15	12	**9**	9
Its2	5	2	**11**	0	5	8	9	4	5	9	9	8	8
UU-Tiedemann	9	0	**11**	**2**	12	**12**	8	6	14	17	**13**	**9**	9
A3-108	0	0	0	0	0	0	0	0	2	3	0	0	0

Table 8.3: Matches per category for the DiscoMT 2015 shared task

task baseline on certain categories such as intra-sentential subject anaphoric "it", whilst most systems perform poorly on event reference and pleonastic pronouns. This breakdown is a good starting point for a more detailed investigation of the problem, including *manual verification* of the mismatches found by the automatic evaluation script.

The counts in Table 8.3 sum the number of pronoun tokens for which the translations by MT systems match those in the reference. However, to get a better idea of how systems compare, we need to look at individual translations. For example, the IDIAP system (Luong et al., 2015) has fewer reference translation matches for intra-sentential anaphoric "they" than the baseline system. However, it produces some pronoun translations that are better than those produced by the baseline. Consider the following example. Here, the IDIAP system translates "corporations" and "they" as "les enterprises" and "elles" (Ex. 8.6) as per the reference (i.e. a translation "match", Ex. 8.5), but the baseline system provides a non-matching (and incorrect) translation of the pronoun: "ils" [masc. pl.] does not agree with "enterprises" [fem. pl.].

(8.4) You are one of those people who believe that **corporations** are an agent of change if **they** are run well . [Source]

(8.5) Vous êtes l' une de ces personnes qui croient que les **entreprises** sont des agents du changement si **elles** sont bien dirigées [Reference]

(8.6) vous êtes de ceux qui croient que **les entreprises** sont un agent de changement , si **elles** sont bien gérées . [IDIAP]

(8.7) Vous êtes de ceux qui croient que **les entreprises** sont un agent de changement s' **ils** sont bien gérés . [Baseline]

Knowing the design of the DiscoMT 2015 systems is also useful when interpreting results. This information can be found in the system description papers, which are available for all systems except A3-108. Taking system design into account the following patterns are observed.

The first is that the auto-postEDIt (Guillou, 2015) and Its2 (Loáiciga and Wehrli, 2015) systems both perform particularly poorly for the event and pleonastic categories. This may be due to design similarities for these systems. Both systems make use of rules; Its2 is a rule-based MT system and auto-postEDIt uses rules to automatically post-edit the output of a baseline phrase-based SMT system. In addition, the focus of both systems is on producing *gendered* pronoun translations. Whereas the auto-postEDIt system uses a simple rule to replace the translations of non-anaphoric pronouns that do not match a predefined set with the token "ce", the Its2 system ignores the problem of translating pleonastic and event reference pronouns altogether. Evidently both of these strategies can be beaten by more sophisticated approaches such as those provided by some of the other systems. This is reflected in the results in Table 8.3.

Another clear pattern is the similarity in performance of the UU-Tiedemann system (Tiedemann, 2015) and the baseline system. Both are phrase-based SMT systems trained using the same data. In contrast to the other systems, the UU-Tiedemann system does not attempt to resolve pronominal anaphora explicitly. Instead, it uses a cross-sentence n-gram model over determiners and pronouns which aims to bias the SMT model towards selecting correct pronouns. In many ways it could be considered the system closest in design to that of the baseline.

The systems generally performed well on the translation of addressee reference "you", as compared with the baseline. However, none of the systems was designed with the aim of handling addressee reference pronouns, given that the focus of the shared task was on translating instances of "it" and "they".

8.5 Conclusion

This chapter presented the PROTEST test suite for evaluating pronoun translation. The test suite is intended to support developers in evaluating the performance of MT systems on the task of pronoun translation. The set of pronoun tokens covers a range of different pronoun *types* and *surface forms*, tailored to the problems that challenge MT.

There are a number of possible areas for future work. A graphical user interface (GUI) could be developed for PROTEST to aid in the manual evaluation of those pronoun translations that do not match the reference, and as a way for researchers to browse pronoun translations and manual evaluation annotations in context. Similar test suites of pronoun tokens could also be created for other datasets and language pairs, and the benefits of using multiple reference translations could be explored.

Chapter 9

Conclusion

The focus of this thesis is the problem of translating pronouns in SMT. Despite recent interest in the problem, little progress has been made to date in terms of building discourse-aware SMT systems that are successful in improving pronoun translation.

This thesis marks two shifts in focus from that of previous work. The first is a shift from the development of systems, to understanding what problems SMT systems face and why system performance has been poor so far. The second is a shift from a narrow focus on the anaphoric pronoun problem (with all anaphoric pronouns typically treated as alike), to the broader focus of pronouns in general. Central to the work in this thesis is the development of the ParCor annotation scheme, in which pronouns are labelled according to their functional *type* (i.e. the function that they perform in text). The provision of the ParCor corpus of English-German parallel texts facilitates both shifts in focus. It provides the means to analyse manual translation, with the aim of identifying systematic differences in pronoun use between English and German. The annotations over the source-language text provide a way to sub-categorise anaphoric pronouns and assess how well state-of-the-art SMT systems perform when translating each subtype.

The analysis of manual translation revealed that the German translations in ParCor include significantly more anaphoric and pleonastic pronouns than the original English texts. Many of the pleonastic pronouns inserted into the German translations correspond to the use of short, fixed phrases including an existential "there" in English. As phrase-based SMT systems are typically very good at translating short, fixed phrases which appear frequently in training data, and

because no other clear patterns were identified in the data, no further investigation into the insertion of pleonastic pronouns was carried out. With anaphoric pronouns, many insertions in the German translations were found to correspond to the use of relativizers in English. The ability of SMT systems to translate both null-relativizers (implicit in text) and that-relativizers (explicit) was assessed in the analysis of the output of two state-of-the-art systems. The study found, perhaps surprisingly, that both systems were able to output the correct relative pronoun in the German SMT output, with a similar accuracy for implicit and explicit relativizers. Also included in the study of SMT output were the anaphoric pronoun "it" (disambiguated from event reference and pleonastic instances using the ParCor annotations) and the possessive pronoun "its" (also anaphoric). As relative pronouns, "standard" anaphoric (i.e. "he/she/it") and possessive pronouns all have different translation requirements in German, their sub-categorisation provides a logical way to analyse their translation.

The main recommendation of this thesis is therefore that when considering the translation of pronouns for a given language pair, they should be categorised according to their function in the source language, **and** by their translation requirements in the target language. For example, the pronoun "it" may be categorised as anaphoric, event reference or pleonastic in English, each requiring different translations in both German and French. This is termed *functional ambiguity* and may occur for different combinations of types, for different languages. In terms of translation requirements it may not be suitable to treat all pronouns of the same type in the same way, as we've seen for the translation of anaphoric pronouns in German. Therefore, future work on translating anaphoric pronouns should consider splitting the anaphoric type into sub-categories: Possessive, relative and reflexive pronouns, and those that do not fall into any of these other categories (e.g. "it"), which may have different translation requirements in a language such as German. This recommendation applies both to the development and analysis of systems.

The second recommendation is that system performance should be analysed in detail. It is important to know where a system does well and where it performs poorly. Such analyses can be aided by categorising pronouns according to their types and relevant sub-types, to help identify where future efforts would be best directed. The PROTEST test suite comprising a set of pronoun tokens and automatic evaluation script is a first attempt at providing a framework for evaluating

English-to-French translation. The set of pronoun tokens covers a range of different categories, representing the range of different problems that MT researchers face for English-to-French translation. For example, different translations are required when translating instances of anaphoric, event reference and pleonastic "it". However, while it is necessary to classify source-language pronouns according to their functional type, this alone is not sufficient. Other information is required. For example, the anaphoric "it" category is further sub-categorised to consider subject vs. object position, for which different sets of French pronouns are required. Further sub-categorisations are also necessary. For example, for addressee reference pronouns which have generic and deictic use, and for anaphoric "they" which may be used as a plural or singular pronoun in English. The PROTEST test suite provides the first step of a detailed analysis of pronoun translation for the systems submitted to the DiscoMT 2015 shared task on pronoun translation, including the post-editing system described in chapter 7.

With respect to SMT system design for pronoun translation, post-editing in which rules are used to *enforce* changes to the output of a baseline system, appears to be a poor approach. Despite encouraging results from Weiner (2014) for English-to-German translation, the English-to-French post-editing system presented in Chapter 7 failed to beat the official DiscoMT 2015 shared task baseline – a simple phrase-based SMT system. The post-editing system was, however, not the only one to perform poorly on the pronoun translation task. As reported in Section 7.8, all of the participating systems failed to beat the baseline. A detailed analysis of the post-editing system components revealed a number of possible sources of error, including the external tools, baseline SMT system and the rules themselves. We might also draw from this that *enforcing* changes to a translation is the wrong approach to take. Instead, it might be better to *encourage* translation changes using techniques that can be more closely integrated in the SMT system, and which allow the decoder to recover from bad decisions. This then shifts the problem back to one of automatic evaluation. Without suitable pronoun-sensitive metrics, tuning systems so that their pronoun-specific modules or features are effective, remains an open problem.

Chapter 10

Future Work

Despite growing interest in pronoun translation in SMT, and discourse in general, much work remains. The results of the DiscoMT 2015 shared task on pronoun translation help to highlight just how much of a problem pronoun translation is for SMT.

Future work on pronoun translation is currently limited by the following factors: Availability of automatic evaluation methods, provision of high-quality automatic external tools as components for discourse-aware SMT systems and a deeper understanding of why previous attempts at pronoun translation have yielded limited improvement. Work on addressing these limitations is crucial to future progress.

The work outlined in this thesis focussed on corpus analyses of English-German pronoun use, the design and analysis of a participating system in the DiscoMT 2015 shared task on pronoun translation for English-French, and the development of the PROTEST test suite for English-French. Similar efforts in terms of corpus annotation, corpus analyses and analysing the performance of discourse-aware SMT systems should be extended to other language pairs and genres, with the ultimate aim of improving SMT system design. However, due to the precise nature of this work, progress in this area is likely to be fairly slow. The PROTEST test suite may also be extended to include hand-selected pronoun token sets for other language pairs, and to use multiple translations instead of a single reference.

Other issues such as pro-drop and other scenarios in which it may be desirable to insert or omit pronouns in translation have not been considered in this thesis, but should be addressed in future work.

The following sections outline the above-mentioned areas for future work in more detail.

10.1 Understanding of Previous Attempts

In order to make progress it is first necessary to understand why previous discourse-aware SMT systems have failed to make improvements in pronoun translation. Work should begin with an extensive analysis of the SMT output of the six participating systems from the DiscoMT 2015 shared task on pronoun translation. The system output of each could be analysed for pronouns sub-categorised by functional type and surface form as outlined in Chapter 7 for the post-editing system. Using the results of the manual evaluation of the same 210 pronouns for each participating system provides plus the annotations over the *DiscoMT2015.test* dataset provides an obvious starting point for such an analysis. The PROTEST test suite provides another possible starting point, but will entail some manual evaluation. To facilitate learning and progress within the community, researchers should be encouraged to perform analyses of their own systems and share the findings.

10.2 External Tools for Discourse-aware SMT

The performance of the external tools used by many researchers in building discourse-aware SMT systems for pronoun translation is far from perfect. Coreference resolution systems and methods for detecting non-anaphoric "it" make incorrect decisions. There has also been little or no attention given to the problems of detecting event reference pronouns (in any language), or to disambiguating between generic vs. deictic "you", or singular vs. plural uses of the English pronoun "they". The availability of tools that can provide accurate automatic annotation of these features could be of great benefit to the development of discourse-aware SMT systems. However, that is not to say that even with the availability of such high-quality tools, pronoun translation will be a solved problem. That is, the poor performance of previously developed SMT systems may not entirely be due to the poor quality of the external tools used. One way to assess this is to conduct oracle experiments in which the systems use gold-standard ParCor-style annotations in place of the output of external tools.

There also remains the question of how best to integrate these external tools, and knowledge gained from detailed analyses of pronoun translation for different language pairs, within the SMT pipeline.

10.3 Extensions to the PROTEST test suite

The pronoun test suite was developed for English-to-French translation and the pronoun tokens are for this language pair. However, the methodology could be applied to any language pair. The automatic evaluation script is language independent and pronoun token sets may be extracted for any language pair. Depending on the language pair, different pronoun categorisations may be appropriate. Based on the *functional ambiguity* of pronouns in the source language, i.e. ambiguity arising from the same surface form pronoun performing many functions, different categorisations may be required to make accurate distinctions between pronoun tokens. A simple extension would be to obtain translations of the texts in the*DiscoMT2015.test* dataset, or a subset of them, in another language. For example, a subset of the TED Talks already have German translations. The same ParCor-style annotations over the English side of the *DiscoMT2015.test* dataset together with alignments between the English and German texts could then be used to extract a set of pronoun tokens for English-German with no further annotation effort.

The ParCor annotation guidelines could also be applied to the manual annotation of new English TED Talks, English texts of a different genre, or texts in languages other than English. These annotations could then be used to identify and categorise pronouns to form pronoun token sets for use with the PROTEST automatic evaluation script. In the case of new language pairs or text genres, pronoun translation frequency distributions may be used to identify those source-language pronoun tokens or categories for which more translation options exist in the target language (as in Section 6.2). We may infer that pronouns with more translation options could be more difficult for an SMT system to translate and therefore warrant attention.

The *DiscoMT2015.test* dataset contains a single reference translation from which the gold-standard translation is extracted for each pronoun token in the test suite. Pronoun translations in the MT output that do not match the reference are referred for manual evaluation, including anaphoric pronouns for which

the pronoun translations match but the antecedent heads do not. However, the provision of multiple reference translations might help to reduce the number of pronoun tokens requiring manual evaluation. Multiple reference translations may provide a number of valid alternative translations for a given pronoun, or in the case of anaphoric pronouns, alternative pronoun-antecedent pairs. In the case of translating anaphoric pronouns into languages with grammatical gender, variation in the reference translation may arise from choosing different translations of a pronoun's antecedent head and selecting a pronoun with the appropriate gender. For other pronoun functions variation may exist from the use of different pronouns which capture the same meaning and may therefore be used interchangeably. For example "you" and "one" may be used interchangeably as generic pronouns, and "this" and "it" as event reference pronouns.

Multiple translations may be gathered through the manual translation of the TED Talks by other translators, or a set of pronoun tokens may be defined over a new corpus which already contains multiple reference translations. Another option is to build up a corpus of manually-verified translations provided by MT systems through manual evaluation, to be used as alternative valid translations (i.e. in addition to those provided in the reference) during automatic evaluation.

10.4 Automatic Evaluation Methods

One of the major factors inhibiting the progress of current research is the lack of automatic evaluation methods for pronoun translation. Manual evaluation is both slow and costly and is not suitable for the rapid develop-and-test environment in which SMT researchers are accustomed to working. BLEU has been rejected on the basis that it is a general purpose metric not capable of reflecting changes in pronoun translation. The BLEU-inspired pronoun evaluation metric introduced by Hardmeier and Federico (2010) is also rejected as it is insufficient in evaluating cases in which the SMT translation is valid, but differs from the reference translation. In order to develop and quickly test multiple discourse-aware SMT systems, the provision of an automatic metric is crucial.

The PROTEST test suite described in Chapter 8 is a step towards fully-automatic evaluation. However, those cases where the SMT output differs from the reference translation still require manual evaluation. Whilst this need can, to some extent, be mitigated by building up a list of manually verified valid alter-

natives for pronoun translations (or pronoun-antecedent pairs in the anaphoric case) from MT output, this still requires a great degree of manual effort. The benefit of the manually verified translations will not be immediate, rather it will be realised gradually, over time. To support manual evaluation, a graphical user interface (GUI) should be added to the test suite. This will allow researchers to browse the translations of the pronoun tokens in context. The GUI could also allow for pronoun-antecedent pairs not present in the reference translation but valid alternatives, to be added to the set of acceptable translations. In its present state, the test suite provides a way to address the shortcomings of BLEU and reduce the cost of manual evaluation. However, the ultimate aim should be to develop a fully-automated method for evaluating pronoun translation.

The value of automatic evaluation metrics is not limited to evaluating SMT output. It also extends to tuning SMT systems. A number of previous attempts at improving pronoun translation have made use of decoder features integrated in the log-linear SMT model. In order to be effective, decoder features need to be assigned sufficient weight during tuning – too little weight and the features will have little or no effect on the translation of pronouns in the SMT output. Decoder features are typically tuned using BLEU, but as previously discussed, BLEU is not suitable for scoring (and therefore tuning) pronoun translation. Without the availability of a suitable tuning metric, researchers may have to resort to manually setting the weights of their features, or focus on pre- and post-processing methods, thereby imposing limits on what may be achieved.

10.5 Corpus Annotation

A simple extension to the ParCor corpus would be to add translations in another language for the existing texts. A number of the EU Bookshop documents have also been translated into French and the TED Talks have been translated into many different languages. The annotation guidelines developed for English and German annotation could be applied to other languages, with extensions added should they be required for the annotation of another language. Knowing more about pronoun use and translation for other language pairs would be useful both in identifying generalisations across languages and the particular problems that exist for specific language pairs. Work on pro-drop languages would be of particular benefit in better understanding those problems that arise when translating

between a pro-drop language and a non-pro-drop language.

The EU Bookshop documents and TED/TEDx Talks[1] provide two distinct text genres; written texts and transcribed planned speech, respectively. In analysing the influence of genre in pronoun use and translation, these two genres provide an important data-point. However, for the same reason that it is important to look at other language pairs, it is also important to consider a wider range of genres to better understand the effects of genre.

The annotation of full coreference chains/sets in a gold-standard corpus could be useful for researchers wishing to study the differences in coreference between two languages from a general perspective. However, the benefit of full coreference with respect to the problem of pronoun translation in SMT, is not known. In the case of using silver-standard annotations (i.e. provided by a coreference resolution system), using only an anaphoric pronoun's nearest non-pronominal antecedent for pronoun-antecedent agreement could result in errors if the wrong antecedent is identified. In this case, obtaining a consensus (perhaps the majority case antecedent head-noun) may help to reduce the errors such that one incorrect antecedent choice is outweighed by a greater number of correct antecedents in the remainder of the coreference set. A similar suggestion has been made by Novák (2011). Another possible direction for investigation would be into the translation of NPs in coreference sets – does lexical consistency exist within NPs in the same coreference set and is this consistency preserved in translation?

It is recognised that additional manual annotation is costly, but this is something that could be coordinated within the community in such a way that the work is divided effectively between different research groups.

10.6 Corpus Analyses

The corpus analysis of manual translation described in Chapter 5 and the analysis of state-of-the art SMT system output in Chapter 6 were conducted for the English-German pair. The analyses focussed on anaphoric and pleonastic pronouns, motivated by differences observed in their distribution in English and German. The analysis of anaphoric pronouns excluded reflexives as these are uncommon in the ParCor texts, but should be considered to provide a complete picture for the anaphoric type. In order to study reflexives, additional texts

[1]TEDx Talks are considered to belong to the same genre as TED Talks.

would need to be added to the ParCor corpus. These texts would need to be taken from a genre in which reflexives appear more frequently. The analysis also excluded other pronoun types due in part to the effort required to conduct such manual analyses. For example, wh-relativizers were excluded from the analysis of state-of-the-art SMT output. Another example of a pronoun that requires further investigation is the addressee reference pronoun "you" which has both deictic and generic use in English and may be translated as "Sie/du" (deictic) or "man" (generic) in German. Correct translation of an instance of "you" into German therefore requires disambiguation in English. Extensions of the English-German analysis to other pronoun types and forms, provides a natural progression of the work already carried out in this thesis.

Corpus analyses for other language pairs may also yield valuable insight into the individual problems that exist for translating different language pairs. For example, different languages may exhibit different biases in terms of pronoun surface forms used for different pronoun types. Different target languages may have different agreement constraints for anaphoric pronouns, or there may be syntactic reasons for omitting/inserting pronouns (e.g. when translating to or from a pro-drop language). These differences may be identified through analyses of manual translation, and their effects on translation through the analysis of SMT output. The analyses conducted for English-German pronoun use and translation (Chapters 5 and 6) hinged on the availability of the ParCor corpus. In order to conduct similar analyses for other language pairs, additional annotation effort would be required.

10.7 SMT System Design

The aim of corpus analyses (and so too analyses of SMT system performance) should be to identify ways in which to improve the design of SMT systems. For example, as revealed by the analysis of state-of-the-art SMT systems in Chapter 6, future efforts for the translation of intra-sentential anaphoric pronouns for English-German could focus on syntax-based SMT — leveraging information within target-side syntax trees constructed by the decoder, to encourage pronoun-antecedent agreement.

The automatic post-editing methods outlined in Chapter 7 and in Luong et al. (2015) replace pronoun tokens with predicted values, which may lead to

local agreement issues. For example, changing only the pronoun may lead to agreement issues with neighbouring verbs or other lexical tokens. As noted in Section 7.11.1 these agreement issues could be addressed within a dependency-parser-based post-editing framework such as the Depfix system (Mareček et al., 2011; Rosa, 2014) — the scope of post-editing changes could be expanded so as to also include tokens surrounding the pronoun. When post-editing is applied to other languages, such as German, the prediction of pronoun tokens may be more complex. As noted in Section 7.11.1, a particular problem for German is the prediction of the case of the pronoun. Possible approaches include employing classifier-based methods for German case prediction, or for jointly predicting the number, gender and case of the pronoun.

10.8 Automatic ParCor-style Annotation

The manually annotated ParCor corpus is useful for investigating differences in pronoun use between a pair of languages and as a gold-standard test set for SMT research. However, it limits research to a particular set of texts. To alleviate this problem, additional texts from the TED Talks and EU Bookshop genres could be manually annotated. Annotation could also be extended to different language pairs (cf. Section 10.6), different genres and to include full coreference chains.

Given the cost of manual annotation, work could also concentrate on the automatic generation of *silver-standard* ParCor annotations over a different range of texts. For English, this could be supported by combining the pre-processing pipeline described in Chapter 4 with the output of additional external tools and hand-crafted rules. The pipeline already makes use of external tools for non-anaphoric "it" detection, dependency parsing (for subject vs. non-subject position pronouns) and NP span detection (for candidate antecedents). To this could be added automatic coreference resolution (originally excluded from the pipeline so as not to bias the annotators), and hand-crafted rules for identifying first-person (speaker reference) pronouns. The disambiguation of instances of "you" as deictic vs. generic could be achieved by training classifiers such as those previously used in disambiguating "you" in meeting transcripts (Gupta et al., 2007). Again, as with external tools used to support the development of discourse-aware SMT systems, there is a limit to both the quality of currently available tools (e.g. coreference resolution) and to what is available (e.g. for event reference detection).

10.9 Functional Phrases

One way to alleviate the reliance on external tools to label pronoun tokens according to their function is to consider other methods for identifying pronoun function. One possible option is to identify *functional phrases* (or *fixed expressions*) that contain pronouns, to be used as indicators of pronoun function. For example, instances of "**They** say..." might indicate the use of a generic, singular "they" as in "**They** say that you should brush your teeth twice a day". This is turn might inform the translation of the pronoun into another language. This is similar to the observation made in Section 5.2.3.2, that instances of "There +*be*" are commonly translated in German as "Es gibt", where "There" is an *existential there* and "Es" is a pleonastic pronoun. Other examples of functional phrases include "Were **it** the case" or "**It** should be pointed out that". Both examples, in which the instance of "it" is pleonastic, may be generalised as "It +*be*" or "*be*+ it" patterns. Other patterns may signify that an instance of "it" is referential. For example, "**It** caused" and "**It** leads to" are patterns which strongly indicate that the instance of "it" is an event reference pronoun (Bin et al., 2010).

Once collated, these sets of *functional phrases* or patterns could be used to identify the function of pronouns in the source-language text, via string or pattern matching. Work on translating the pronouns might then focus on translating the entire phrase (i.e. the pronoun in context), rather than considering the pronoun in isolation. For example, in a post-editing scenario, this would mean replacing entire phrases, rather than just pronoun tokens. This may better fit the current SMT paradigms, which work beyond the word (or token) level.

10.10 Other Issues

The study of subject pro-drop, in which subject position pronouns are omitted if they can be inferred from the text using other means, has been excluded from this thesis. When considering the translation to or from a pro-drop language such as Czech or Spanish, this issue cannot be ignored. Future work should consider the analysis of pro-drop in translation, its evaluation in SMT, and approaches suitable to the omission of subject position pronouns when translating into a pro-drop language, as well as the insertion of pronouns when translating from a pro-drop language to a non-pro-drop language. There may also be other

reasons for which pronouns should be inserted or omitted during translation — syntactic or stylistic requirements, both of which deserve further attention. This also opens up more general questions surrounding the eventual aim of research into pronoun translation, including whether researchers should aim to provide accurate translation, or *natural* (i.e. *human-like*) translation.

Appendix A

APPENDIX A: Annotation Guidelines

This is a copy of the annotation guidelines given to the annotators. A copy is also included with the ParCor 1.0 corpus Guillou et al. (2014) release. The core annotation guidelines were developed at the University of Edinburgh. Credit for the TED-specific annotation rules goes to Christian Hardmeier and Aaron Smith, at the University of Uppsala.

The ParCor corpus is available to download from OPUS, the online parallel corpus website[1], free of charge. Use and redistribution of the corpus texts is governed by the terms of use laid out by the EU Bookshop and TED organisations, as displayed on the EU Bookshop[2] and TED[3] websites.

A.1 Note

These guidelines were presented to the English and German annotators who worked on the annotation of the EU Bookshop and TED Talks texts in the ParCor 1.0 corpus. The document is split into three main sections: General guidelines (applicable to both text genres) and additional guidelines specific to the annotation of the EU Bookshop and TED Talks portions of the corpus respectively.

[1]ParCor: http://opus.lingfil.uu.se/ParCor/
[2]EU Bookshop: http://bookshop.europa.eu/
[3]TED: http://www.ted.com

A.2 Pre-populated Markables

In order to assist the annotation process pre-populated MMAX-2 *markables* are provided as a starting point to the manual annotation. These markables represent:

- Pronouns: These will first appear as bold blue text with a magenta background and the background colour will disappear once you have determined a new *type* for them – e.g. anaphoric, pleonastic or event.
- Noun Phrases (NPs): A set of potential antecedents for pronouns. These will appear as normal blue text (as for any other markable in MMAX-2).

Please note the markables were produced by automated tools and may not be 100% accurate. You should look for errors such as:

- Pronouns that have not been identified (and therefore are not labelled as markables)
- Words that have been mis-labelled as pronouns
- Potential pronoun antecedents (the set of NPs) which may be missing and therefore need to be added (manually), or ones where their span may be too large or small and therefore needs adjusting (manually)

A.3 General Guidelines: What to Include

We wish to construct links between pronouns and their antecedent(s). These linked pronoun-antecedent pairs should exclude events (and references to the events) and pleonastic/dummy pronouns (see Section A.4 on what to exclude).

The following set of guidelines has been condensed from the MUC-7 guidelines[4].

Pronoun *forms* to be annotated, include:

- Personal: First, second and third-person
- Possessive
- Demonstrative

[4]MUC-7: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html

- Relative
- Reflexive (TED Talks only)
- Pronominal adverbs (EU Bookshop only)
- Generic

The possessive forms of pronouns used as determiners are markable. Thus in:

> The company and [**its chairperson**]

there are two potentially markable elements: **its** and the entire NP, **its chairperson**.

First, second, and third-person pronouns are all markable, so in:

> "There is no business reason for [**my**] departure", [**he**] added.

my and **he** should be marked as coreferential.

A.3.1 Anaphoric and Cataphoric Pronouns

We are interested in marking pronouns (e.g. he, she, it, they,...) and their *antecedent* (the thing that the pronoun refers to). For example, in the following example:

> [**Alan Turing**]$_1$ was born at Paddington, London. [**His**]$_1$ father, [**Julius Mathison Turing**]$_2$, was a British member of the Indian Civil Service and [**he**]$_2$ was often abroad...

the pronoun **His** refers to **Alan Turing** - in other words the antecedent of **His** is **Alan Turing**. The pronoun **he** refers to **Julius Mathison Turing**, not to **Alan Turing**.

In some cases a pronoun can refer to more than one entity. Consider the following example in which the pronoun **They**, refers to two people: **John** and **Mary**, captured in the single conjoined NP antecedent **John and Mary**:

> [**John and Mary**] went to the cinema. [**They**] saw a film about penguins.

When a pronoun appears after its antecedent/referent in a text we call this *anaphora* (the relationship is *anaphoric*). The pronouns in the above examples are anaphoric. When a pronoun appears before its referent in a text we call this *cataphora* (the relationship is *cataphoric*). The pronoun **she** in the example below is cataphoric:

If [**she**] is in town, [**Mary**] can join us for dinner.

We are only interested in cataphoric relations in which the pronoun and its referent occur in the same sentence. Also, consider the following rule for deciding if a pronoun is anaphoric/cataphoric: If the pronoun can be marked as anaphoric, mark it as such. If no possible antecedent appears before the pronoun, then consider linking it as cataphoric. (We will use the term *antecedent* to refer to the NP that either a cataphoric or anaphoric pronoun refers to.)

A.3.2 Speaker/Addressee Reference Pronouns

These are defined as:

- *Addressee reference*: Where the pronoun primarily refers to the addressee (person being addressed)
- *Speaker reference*: Where the pronoun primarily refers to the speaker or may not include the addressee

As a guideline:

- First-person pronouns normally refer to the speaker, in the case of the singular (e.g. the English "I"), or to the speaker and others, in the case of the plural (e.g. the English "we")
- Second-person pronouns normally refer to the person or persons being addressed (e.g. the English "you"); in the plural they may also refer to the person or persons being addressed together with third parties
- Third-person pronouns normally refer to third parties other than the speaker or the person being addressed (e.g. the English "he", "she", "it", "they")
- Plural pronouns like "us", "you" and "we" may be more difficult. "we", "us" and "our" will most likely be speaker reference, and instances of "you" and "your" will likely be addressee reference

- Be aware that it may be difficult to distinguish between *speaker reference* and *addressee reference* in some cases

A.3.3 Pleonastic Pronouns

These are pronouns that do not actually refer to an entity. In other words, the pronoun could not be replaced with an NP as with a regular pronoun. Often a *subject* is required by syntax i.e. something is required in that position. In some cases there will not be a *subject* so a "dummy" pronoun is required to fill the gap. For example in the following sentences the pronoun **it** does not refer to anything but is included as something is required by the syntax of the language in the subject position:

- **It** is raining
- **It** is well known that apples taste different from oranges

It is commonly used as a pleonastic pronoun in English. Other pronouns such as **they** and **you** may also be used in cases where they do not refer to a specific entity:

- In this country, if **you** own a house **you** have to pay taxes
- **They** say you should never mix business with pleasure

In the case of *pleonastic* pronouns we wish to make a partial annotation: Marking the pronoun as *pleonastic*, but not linking it to anything (because it does not refer to anything).

A.3.4 Identifying the Antecedent(s)

Once an *anaphoric* or *cataphoric* pronoun has been identified a pronoun, its *antecedent* needs to be determined. There are several cases. The pronoun may refer to:

- An entity (represented by a noun or NP)
- An event (see Section A.4.1)
- Nothing (see Section A.3.3)

- It may be possible to tell that a pronoun is anaphoric, but there is no specific antecedent in the text. For example the pronoun **these** in "Access to 0800 numbers...**these** calls" (see Section A.3.7)
- A word may have been marked as a pronoun in error (i.e. the automated pre-processing pipeline made an incorrect choice)

In order to identify what a pronoun refers to, the pronoun itself should be used as a starting point. Look back earlier in the text (working backwards sentence by sentence) until the nearest non-pronominal antecedent is identified. For example, in:

> The details of [**Miyamoto Musashi**]'s early life are difficult to verify. [**Musashi**] simply states in Gorin no Sho that [**he**] was born in Harima Province

the pronoun **he** should be linked to **Musashi**, the nearest antecedent, and not to **Miyamoto Musashi** which appears earlier in the text.

A.3.5 Special Case: Pronoun has Multiple Antecedents

In cases like:

> [**John**] likes documentaries. [**Mary**] likes films about animals. The last time [**they**] went to the cinema [**they**] compromised and saw a film about penguins.

They refers to both **John** and **Mary**, who are mentioned in separate sentences so there is no NP span that covers both **John** and **Mary**. In cases like these, if all of the antecedents can be identified and it is clear from the texts what the antecedent are, the pronoun should be linked to each of the separate antecedent "parts". It is important to ensure that all "parts" are linked.

It is important to first ensure that no NP exists that covers all parts of the antecedent.

A.3.6 Special Case: They

When the pronoun **they** is used to refer anaphorically to a collective noun (such as **the government**), it should be considered a plural pronoun and marked as such.

A.3.7 Special Case: No Specific Antecedent

For example, in the sentence:

> There's a study called the streaming trials. **They** took 100 people and split them into two groups

There is no antecedent in the text to which the pronoun **They** may be linked. In this case, the pronoun should be marked as "anaphoric but no specific antecedent".

A.3.8 Special Case: "he or she", "him or her", "his or her" and "his or hers"

English lacks un-gendered person pronouns, and the former solution of just using "male pronouns" (e.g. "he") is now considered bad form. Therefore, you may come across instances of "he or she" in a text. For example:

> If your child is thinking about a gap year, [**he or she**] can get good advice from this website.

In such cases, **he or she** should be considered a single unit (or markable), just as if it had been written "s/he" (which is a common alternative). This solution will also make the phrase easier to resolve, since it can only be linked to a non-specific antecedent.

The same applies to instances of "him or her", "his or her" and "his or hers".

A.3.9 Special Case: "s/he"

Treat this as a complete unit (or markable) and as a pronoun.

A.3.10 Special Case: The Pronoun Refers to a Modifier

In some cases, the pronoun may refer to a modifier in an NP. Consider the following example:

> The unionists used to be [**EU supporters**], but now they are questioning how [**it**] has developed...

Here, the pronoun **it** cannot be linked to the complete NP **EU supporters**, but **it** can be linked to **EU** (the modifier). If none of the automatically generated markables are suitable, the span of an existing markable should be adjusted or a new markable created. The resulting markable may or may not be an NP.

However, with compounds like EU-supporters, these exist as a single unit and cannot be split any further (i.e. it is not possible to construct a markable that covers only the **EU** part). In such cases, it is necessary to search for a stand-alone instance of **EU** earlier in the text and link the pronoun **it** to that instance (assuming one can be found).

A.3.11 How Much of a Markable to Annotate

A markable is any pronoun, noun or NP that will be "marked" because it forms part of pronoun-antecedent pair, or a pronoun for which there is no antecedent to be marked. For pronouns, the markable will be a single word. For a pronoun's antecedent(s), the markable will be a noun or an NP. For noun or NP markables, the following rules apply. The markable must:

- Contain the head (main) noun
  - E.g. **task** is the head in **the coreference task**
  - If the head is a name then the entire name (not just a part of it) should be marked. E.g. **Frederick F. Fernwhistle Jr.** in **the Honorable Frederick F. Fernwhistle Jr.**
- Also include all text which may be considered a modifier of the NP
  - E.g. **the Honorable Frederick F. Fernwhistle Jr.**
  - E.g. **Mr. Holland**
  - E.g. **the coreference task** (where **task** is the head) – this provides information about what the task is and separates it from **other coreference tasks**, **the scheduling task**, etc.
  - E.g. **the big black dog** (where **dog** is the head)
  - Determiners such as **the** should be included

N.B. The automatically generated set of markables may contain NPs that have incorrect spans. The spans may therefore require manual adjustment.

A.3.12 Relationships Between Markables

For a pronoun and its antecedent(s), the relationship between the elements is termed as *anaphoric/cataphoric*. For *pleonastic* and *event* pronouns there will not be a link to an antecedent markable.

A.4 General Guidelines: What to Exclude

A.4.1 The Events in Event Reference

The events in event reference — where pronouns are used to refer to an event that has happened or will happen, should not be marked. Event pronouns can refer back to whole sections of text or concepts evoked by the text. For example in:

> Ted [**arrived late**]. [**This**] annoyed Mary.

This refers to the event **arrived late**.
Another example:

> Vulnerable consumers in particular might need [**specific support**] to enable them to finance necessary investments to reduce energy consumption. [**This**] task...

Using deictics that vaguely refer to what the speaker is talking about (as in the above example) is bad writing, but examples like this exist in some of the texts. Here **this** should be treated as an instance of event reference.

In general, events should be easy to identify as they should contain verbs. As with the annotation of *pleonastic* pronouns a partial annotation is required: The pronoun is marked as *event*, but is not linked to the event itself.

Identifying pronouns that refer to events can be difficult, therefore the following simple rule is proposed:

- *English:* Try replacing the pronoun with a period and then start a new sentence *or* test if you can replace an instance of **which** with **this**
- *German:* Try replacing the pronoun with a period and then start a new sentence with **das**

If the resulting "new text" reads OK, then it is likely that the pronoun refers to an event. As an example of how this test would work, consider the following sentence:

Ted arrived late, [**which**] annoyed Mary.

Question: Is "**which**" an event pronoun?

Replace the pronoun **which** with a period and start the new sentence with **This**:

Ted arrived late[**. This**] annoyed Mary.

Result: Mark **which** as an event pronoun as the "test" passed.

If two pronouns refer to the same event, each should be marked as an *event* pronoun (as opposed to marking the second as anaphoric to the first) and the two instances linked together.

A.5 Special Instructions for the Annotation of Written Text: EU Bookshop

The following instructions are specific to the annotation of written text and should be used when annotating the EU Bookshop documents.

A.5.1 Reflexive Pronouns

Reflexive Pronouns should not be marked.

In cases like "the man himself" we do not treat **himself** as a pronoun. Instead it should be considered an NP (the markable span can be amended in MMAX-2) if it has been automatically marked as a pronoun in error.

A.5.2 Indefinite Pronouns

An indefinite pronoun is a pronoun that refers to one or more unspecified beings, objects, or places. For example:

[**Anyone**] can see that she was looking for trouble.

Here, **Anyone** is an indefinite pronoun as it does not refer to a specific person or group of people.

Indefinite pronouns should be marked as *pronoun*, to indicate that they have been "seen" in the text. As they will be marked as instances of the type *pronoun*, they will not be linked to anything, nor will any other features be recorded.

A.5.3 Numbers/Quantifiers Used as Pronouns

When deciding whether to link a pronoun to an *antecedent*, the following rules apply:

- many of **them** ...: **them** should be linked to its antecedent
- **one** of the fast growing economies: **one** should be marked as a pronoun but not linked to anything
- **others** ...: **others** is anaphoric and has an antecedent, but it is not coreferent with its antecedent. It should be marked as a pronoun but not linked to anything
- **both**: This is anaphoric, either to two individuals or two events or situations. If **both** here is a *bare* pronoun, it should be marked and linked. If it has a head (as in "both boys"), then it should be marked as a pronoun but not linked to anything
- **each**: This is anaphoric to a set. If **each** here is a *bare* pronoun, it should be marked and linked. If it has a head (as in "each boy"), then it should be marked as a pronoun but not linked to anything

A.5.4 Pronominal Adverbs

Pronominal Adverbs are a type of adverb occurring in both English and German (although they appear to be used more frequently in the German texts). They are formed by replacing a preposition and a pronoun. We wish to annotate these.

For example:

- For that → therefore
- In that → therein

- By this → hereby
- To this → hereto
- In which → wherein

A.5.5 Pronouns Within Quoted Text

Labelling a first-person or second-person pronoun within quoted text can become difficult. Furthermore, the focus is on translating coreference in *normal* text, not *quoted* text. We therefore simplify the annotation task using the following rules:

- First and second-person pronouns within quoted text should simply be marked as instances of type *pronoun*
- Third-person personal pronouns should be marked as normal
- In some cases, the text may read like an interview (with questions and answers) but with no quotes. In this case, the text is not to be treated as quoted text. Speaker/addressee reference pronouns should be annotated as normal.

A.5.6 Difficult Choices: Deciding Between Anaphoric or Event Categories

In some scenarios is it possible to read the text in more than one way and both readings appear to be equally likely. For example, it may be possible to mark the pronoun as either *event reference* (referring to a phrase with a verb) or *anaphoric* (referring to an NP), i.e. it is ambiguous. As an example, consider:

> In the framework of the North Seas Countries' Offshore Grid Initiative, ENTSO-E is already conducting grid studies for northwestern Europe with a 2030 horizon. [**This**] should feed into ENTSO-E's work for a modular development plan of a pan-European electricity highways system up to 2050.

In this example, the pronoun **This** could refer to:

- North Seas Countries' Offshore Grid Initiative (NP)
- conducting grid studies for northwestern Europe with a 2030 horizon (Verb Phrase)

In scenarios such as these, if multiple labels would be possible, select *anaphoric* and link the pronoun to the NP. This will provide more information when the data is used for training translation systems.

If it is *impossible* to tell what the pronoun refers to or if the text is very poorly written, the pronoun may be marked as *Not sure. Help!*. This will help to identify those scenarios that are very difficult for humans (and therefore even more difficult for machines) to determine.

A.6 Special Instructions for the Annotation of Spoken Text: TED Talks

The following instructions are specific to the annotation of transcribed spoken text and should be used when annotating the TED Talks documents.

A.6.1 Reflexive Pronouns

Reflexive pronouns should be annotated in English and German.

For English:

- Exclude instances of **myself** from the annotation as it is a singular first-person pronoun

For German:

- Include instances of **mich** even though it is a singular first-person pronoun as it can be reflexive *or* personal and it is important to make the distinction
- The pronouns **mich**, **dich**, **uns** and **euch** can all be used as either personal or reflexive pronouns. Mark whether they are personal or reflexive

A.6.2 First-person Pronouns

Singular first-person pronouns (I, me, etc.) do not need to be marked as they can be recovered automatically.

A.6.3 Speaker Reference

For pronouns that fall into the *speaker reference* category (used for all instances of **we**), the audience should be recorded. There is an *audience* attribute (in MMAX-2), which can be set as either:

- *Exclusive we*, meaning the speaker and his/her clique but not the audience
- *Co-present we*, meaning the speaker plus everyone physically present in the room
- *All-inclusive we*, incorporating everything else

A.6.4 Addressee Reference

For pronouns that fall into the *addressee reference* category, the audience should be recorded. There is an *audience* attribute (in MMAX-2), which can be set as either:

- *Deictic*, meaning that the speaker is really referring to the audience or a specific person
- *Generic*, as in phrases such as: In England, if [**you**] own a house [**you**] have to pay taxes

When a speaker uses deictic **you**, talking to the whole audience, it should always be marked as plural, even in cases like "Imagine **you**'re walking alone in the woods", where there is clearly a singular sense to the word.

For generic cases of **you**, it is not necessary to make a singular vs. plural distinction.

N.B. **you** should not be labelled as as pleonastic

A.6.5 Pronouns Within Quoted Text

These pronouns should be annotated strictly from the point of view of the quoted speaker, not of the speaker who quotes the utterance. In particular, this means:

- First-person pronouns are always speaker reference
- Second-person pronouns are always addressee reference

- A coreference relation is never marked between a first-person or second-person pronoun inside quoted speech with a pronoun outside the quoted speech passage (as in ‘ He said, “I do.” ’, where **he** and **I** could arguably be marked as coreferent)

In examples such as:

> I said, “[**Miguel**], what makes your fish taste so good?” [**He**] pointed at the algae.

Do not link the pronoun **He** (outside of the quote) to **Miguel** (inside of the quote). Instead, look for an earlier instance of the entity (i.e. Miguel) in the text that does not appear in quotes and link the pronoun (i.e. **He**) to that instance.

A.6.6 Extra-textual Reference

For cases where the speaker refers to something such as a slide or prop, the pronoun should be marked as *extra-textual*. Two pronouns referring to the same object should both be marked as *extra-textual* and linked together as co-referents.

The *extra-textual* category can also be used within quoted text when a third-person is referred to such as the **he** in:

> People when they see me say “[**he**]’s a bit weird”

N.B. This is rarely required

A.6.7 No Explicit Antecedent

In cases like:

> There’s a study called the streaming trials. [**They**] took 100 people and split them into two groups

where there is no explicit antecedent for **They**, the pronoun should be marked as *anaphoric* and the *no explicit antecedent* sub-category should also be selected.

Do not mark **They** as pleonastic.

A.6.8 Split Antecedent

This should be marked if the pronoun has multiple antecedents. All components of the antecedent should be linked to the pronoun directly, and not to each other.

A.6.9 Simple Antecedent

For all cases except where there is *no specific antecedent* or there is a *split reference*.

A.6.10 Indefinite Pronouns, Pronominal Adverbs and Numbers/Quantifiers Used as Pronouns

Instances of these pronouns should not be marked.

Appendix B

APPENDIX B: Pronoun Forms

The pronoun *form* categories referred to in Chapter 4 are defined in the following sections. For both English and German, the pronouns that belong to each pronoun *form* category. Please note that the pronouns listed cover those annotated in the ParCor corpus.

Pronoun Form Category	**List of Pronouns**
First-person personal	I, we, us, me
Second-person personal	you
Third-person personal	he, she, him, her, it, they, them, "he or she"
Possessive	my, mine, our, ours, your, yours, his, her, hers, its, their, theirs, "his or her", oneś
Relative/Demonstrative	which, who, whom, whose, where, this, that, these, those
Reflexive	himself, herself,itself, themselves, themself, yourself, ourselves, ourself, yourself, yourselves, myself, oneself
Pronominal Adverbs	hereabout, hereabouts, hereafter, hereat, hereby, herein, hereinafter, hereinbefore, hereinto, hereof, hereon, hereto, heretofore, hereunder, hereunto, hereupon, herewith, herewithin, thereabout, thereafter, thereagainst, herearound, thereat, therebeyond, thereby, therefor, therefore, therefrom, therein, thereinafter, thereof, thereon, thereover, therethrough, therethroughout, thereto, theretofore, thereunder, thereunto, thereupon, therewith, therewithal, therewithin, whereabout, whereabouts, whereafter, whereas, whereat, whereby, wherefore, wherefrom, wherein, whereinto, whereof, whereon, whereover, wherethrough, whereto, whereunder, whereupon, wherever, wherewith, wherewithal, wherewithin, wherewithout

Table B.1: Pronoun **form** categories for English

Pronoun Form Category	**List of Pronouns**
First-person personal	ich, mich, mir, wir, uns
Second-person personal	ihr, euch, Sie, Ihnen,du, dich, dir
Third-person personal	er, sie, es, ihn, ihm, ihr, ihnen, er oder sie, man
Possessive	meiner, deiner, seiner, ihrer, unser, euer, eures, eurer, eure, Ihrer, Ihres, mein, meinen, meinem, meines, meine, sein, seinen, seinem, seines, seine, dein, deinen, deinem, deines, deine, unsere, unseren, unserem, unseres, unserer, ihre, ihres, ihrem, ihren
Relative/Demonstrative	dieser, diese, dieses, diesem, diesen, jener, jene, jenes, jenen, der, die, das, dass, dem, den, des, dessen, denen, deren, derer, was, wer, wo, welche, welches, welcher, welchen, dies, denjenigen, diejenigen, diejenige, denjenigen, derjenige, derjenigen, desjenigen, demjenigen, denjenigen, dasjenige
Reflexive	sich, sich selbst, einander, micht, er selbst, du selbst
Pronominal Adverbs	Start with “wo” or “da”
Other (found in ParCor texts)	all das, alle, allem, allen, alles, alles andere, andere, beide, beides, drei, eine, eines, einer, etwas, jedem, jemand, jemanden, nichts, solche, viele, vielen, wir alle, wir als menschen, ’s

Table B.2: Pronoun **form** categories for German

Bibliography

Amr Ahmed and Greg Hanneman. Syntax-Based Statistical Machine Translation: A Review. *Unpublished Manuscript*, 2005.

Nicholas Asher. *Reference to Abstract Objects in Discourse*. SLAP 50, Dordrecht, Kluwer, 1993.

Kathryn Baker, Bonnie Dorr, Michael Bloodgood, Chris Callison-Burch, Nathaniel Filardo, Christine Piatko, Lori Levin, and Scott Miller. Use of Modality and Negation in Semantically-Informed Syntactic MT. *Computational Linguistics*, 38(2):411–438, 2012.

Mona Baker. Corpus Linguistics and Translation Studies: Implications and Applications. In Gill Francis and Elena Tognini-Bonelli, editors, *Text and Technology: in Honour of John Sinclair*, pages 233–252. John Benjamins, Amsterdam, 2003.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

Adrien Barbaresi. German Political Speeches, Corpus and Visualization: 2nd release. Technical report, ENS Lyon, 2012. URL `http://purl.org/corpus/german-speeches`. <halshs-00677928>.

Marco Baroni and Silvia Bernardini. A New Approach to the Study of Translationese: Machine-Learning the Difference Between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274, 2006.

Viktor Becher. *Explicitation and Implicitation in Translation. A Corpus-based Study of English-German and German-English Translations of Business Texts.* PhD thesis, University of Hamburg, 2011.

Shane Bergsma and David Yarowsky. NADA: A Robust System for Non-Referential Pronoun Detection. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, pages 12–23, Faro, Portugal, 2011.

Chen Bin, Su Jian, and Tan Chew Lim. A Twin-Candidate Based Approach for Event Pronoun Resolution using Composite Kernel. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling)*, pages 188–196, Beijing, China, 2010.

Alexandra Birch, Matthias Huck, Nadir Durrani, Nikolay Bogoychev, and Philipp Koehn. Edinburgh SLT and MT System Description for the IWSLT 2014 Evaluation. In *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT 2014)*, pages 49–56, Lake Tahoe, CA, USA, 2014.

Anders Björkelund and Richárd Farkas. Data-driven Multilingual Coreference Resolution using Resolver Stacking. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 49–55, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Ondřej Bojar and Kamil Kos. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 60–66, Uppsala, Sweden, 2010. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland USA, 2014. Association for Computational Linguistics.

Adriane Boyd, Whitney Gegg-Harrison, and Donna Byron. Identifying Non-referential It: A Machine Learning Approach Incorporating Linguistically Motivated Patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, FeatureEng '05, pages 40–47, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A Centering Approach to Pronouns. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*, ACL '87, pages 155–162, Stanford, California, 1987. Association for Computational Linguistics.

Samuel Broscheit, Simone Paolo Ponzetto, Yannick Versley, and Massimo Poesio. Extending BART to Provide a Coreference Resolution System for German. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 164–167, Valletta, Malta, 2010. European Language Resources Association (ELRA).

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. How Comparable Are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. In *Proceedings of BUCC 2011*, pages 78–86, Portland, Oregon, 2011.

Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, 2012.

David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

Nancy Chinchor and Lynette Hirschman. MUC-7 Coreference Task Definition (Version 3.0). In *Proceedings of MUC-7*, 1998. URL `http://acl.ldc.upenn.edu/muc7/co_task.html`.

Noam Chomsky. *Lectures on Government and Binding*. Foris, Dordrecht, 1981.

J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 1960.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 449–454, Genoa, Italy, 2006. European Language Resources Association (ELRA).

Mona T. Diab. An Unsupervised Approach for Bootstrapping Arabic Sense Tagging. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Semitic '04, pages 43–50, Geneva, Switzerland, 2004. Association for Computational Linguistics.

Simon C. Dik. *Functional Grammar*. Amsterdam, North-Holland, 1978.

Gerda Dippmann. *A Practical Review of German Grammar*. Macmillan Publishing Company, 1 edition, 1987.

George Doddington. Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Diego, California, 2002. Morgan Kaufmann Publishers Inc.

Markus Dreyer and Daniel Marcu. HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada, 2012. Association for Computational Linguistics.

Greg Durrett and Dan Klein. Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, 2013. Association for Computational Linguistics.

Andreas Eisele and Yu Chen. Multiun: A multilingual corpus from united nation documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 2868–2872, Valletta, Malta, 2010. European Language Resources Association (ELRA).

Ulrich Engel. *Deutsche Grammatik*. Julius Groos Verlag Heidelberg, 2 edition, 1988.

Antonio Ferrández and Jesús Peral. Translation of Pronominal Anaphora between English and Spanish: Discrepancies and Evaluation. *Journal Of Artificial Intelligence Research*, 18:117–147, 2003.

William A. Gale and Kenneth W. Church. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102, March 1993.

Martin Gellerstam. *In Lars Wollin and Hans Lindquist (eds), Translationese in Swedish Novels Translated From English*. Lund: CWK Gleerup, 1986.

Kevin Gimpel and Noah A. Smith. Rich Source-Side Context for Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 9–17, Columbus, Ohio, 2008. Association for Computational Linguistics.

Alexander Gode and Hugh Blair. *Interlingua: A Grammar of the International Language*. New York: Storm Publishers, 1 edition, 1951.

Zhengxian Gong, Min Zhang, and Guodong Zhou. Cache-Based Document-Level Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 909–919, Edinburgh, United Kingdom, 2011. Association for Computational Linguistics.

Joseph H. Greenberg. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, Massachusetts, 1963.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Providing a Unified Account of Definite Noun Phrases in Discourse. In *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*, ACL '83, pages 44–50, Cambridge, Massachusetts, 1983. Association for Computational Linguistics.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. Centering: A Framework for Modeling the Local Coherence Of Discourse. *Computational Linguistics*, 21:203–225, 1995.

Helmut Gruber and Gisela Redeker. *The Pragmatics of Discourse Coherence: Theories and applications*, volume 254 of *Pragmatics & Beyond New Series*. John Benjamins Publishing Company, 1 edition, 2014.

Liane Guillou. Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 1–10, Avignon, France, 2012. Association for Computational Linguistics.

Liane Guillou. Automatic Post-Editing for the DiscoMT Pronoun Translation Task. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 65–71, Lisbon, Portugal, 2015. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

Liane Guillou and Bonnie Webber. Analysing ParCor and its Translations by State-of-the-art SMT Systems. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 24–32, Lisbon, Portugal, 2015. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3193–3198, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

Surabhi Gupta, Matthew Purver, and Dan Jurafsky. Disambiguating Between Generic and Referential "You" in Dialog. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 105–108, Prague, Czech Republic, 2007. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký,

Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

Najeh Hajlaoui and Andrei Popescu-Belis. Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric. In *14th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 236–247, Samos, Greece, March 2013. University of the Aegean, Springer.

Michael A. K. Halliday. *An introduction to functional grammar*. Hodder Arnold London, 3rd ed. / rev. by Christian M.I.M. Matthiessen. edition, 2004.

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

Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German.* De Gruyter, Berlin, 2012.

Christian Hardmeier. *Discourse in Statistical Machine Translation*. PhD thesis, University of Uppsala, 2014.

Christian Hardmeier. On Statistical Machine Translation and Translation Theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 168–172, Lisbon, Portugal, September 2015a. Association for Computational Linguistics.

Christian Hardmeier. A Document-Level SMT System with Integrated Pronoun Prediction. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 72–77, Lisbon, Portugal, September 2015b. Association for Computational Linguistics.

Christian Hardmeier and Marcello Federico. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT 2010)*, pages 283–289, Paris, France, 2010.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal, 2015. Association for Computational Linguistics.

Christian Hardmeier, Jörg Tiedemann, Preslav Nakov, Sara Stymne, and Yannick Versely. DiscoMT 2015 Shared Task on Pronoun Translation, 2016. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. `http://hdl.handle.net/11372/LRT-1611`.

Roger Hawkins, Richard Towell, and Marie-Noëlle Lamy. *French Grammar and Usage*. Hodder Arnold, 2 edition, 2001.

Kenneth Heafield. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011.

Jerry Hobbs. Resolving Pronominal References. *Lingua 44*, pages 311–338, 1978.

Rodney Huddleston. *Introduction to the Grammar of English*. Cambridge University Press, 1 edition, 1988.

Mitesh M. Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. Projecting Parameters for Multilingual Word Sense Disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 459–467, Singapore, 2009. Association for Computational Linguistics.

Paul Kiparsky. Disjoint Reference and the Typology of Pronouns. In I. Kaufmann and Barbara Stiebels, editors, *More than Words: a Festschrift for Dieter Wunderlich*, pages 179–226. Berlin: Akademie Verlag, 2002.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: The Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT.

Philipp Koehn and Hieu Hoang. Factored Translation Models. In *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, 2007. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Edmonton, Canada, 2003. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Prague, Czech Republic, 2007. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. Translationese and its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1318–1326, Portland, Oregon, 2011. Association for Computational Linguistics.

Kamil Kos and Ondřej Bojar. Evaluation of Machine Translation Metrics for Czech as the Target Language. *Prague Bulletin of Mathematical Linguistics*, 92:135–147, 2009.

Kerstin Kunz and Ekaterina Lapshinova-Koltunski. Cross-linguistic Analysis of Discourse Variation Across Registers. *Nordic Journal of English Studies*, 14 (1):258–288, 2015.

Sadao Kurohashi, Toshiaki Nakazawa, Kauffmann Alexis, and Daisuke Kawahara. Example-based Machine Translation Pursuing Fully Structural NLP. In

Proceedings of the 2nd International Workshop on Spoken Language Translation (IWSLT 2005), pages 207–212, Pittsburgh, PA, USA, 2005.

Shalom Lappin and Herbert J. Leass. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20:535–561, 1994.

Ekaterina Lapshinova-Koltunski and Kerstin Kunz. Annotating Cohesion for Multilingual Analysis. In *Proceedings of the 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 57–64, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

Sara Laviosa. *Corpus-Based Translation Studies. Theory, Findings, Applications.* Rodopi, Amsterdam and New York, 2002.

Ronan Le Nagard and Philipp Koehn. Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden, 2010. Association for Computational Linguistics.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Portland, Oregon, 2011. Association for Computational Linguistics.

Geoffrey Leech, Marianne Hundt, Christian Mair, and Nicholas Smith. *Change in Contemporary English: A Grammatical Study.* Cambridge University Press, Cambridge, 2009.

Ding Liu and Daniel Gildea. Semantic Role Features for Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 716–724, Beijing, China, 2010. Association for Computational Linguistics.

Sharid Loáiciga and Eric Wehrli. Rule-Based Pronominal Anaphora Treatment for Machine Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 86–93, Lisbon, Portugal, 2015. Association for Computational Linguistics.

Annie Louis and Bonnie Webber. Structured and Unstructured Cache Models for SMT Domain Adaptation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 155–163, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Ngoc Quang Luong, Lesly Miculicich Werlen, and Andrei Popescu-Belis. Pronoun Translation and Prediction with or without Coreference Links. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 94–100, Lisbon, Portugal, 2015. Association for Computational Linguistics.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June 1993.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. Two-Step Translation with Grammatical Post-Processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 426–432, Edinburgh, Scotland, 2011. Association for Computational Linguistics.

Thomas Meyer and Lucie Poláková. Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013*, pages 43–50, Sofia, Bulgaria, 2013.

Thomas Meyer and Andrei Popescu-Belis. Using Sense-Labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the EACL2012 Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138, Avignon, France, 2012.

Thomas Meyer, Andrei Popescu-Belis, Jeevanthi Liyanapathirana, and Bruno Cartoni. A Corpus-Based Contrastive Analysis for Defining Minimal Semantics of Inter-Sentential Dependencies for Machine Translation. In *Proceedings of the GSCL2011 Workshop on "Contrastive Analysis - Translation Studies - Machine Translation: What can we learn from each other?"*, Hamburg, Germany, 2011.

Ian Michael. *English Grammatical Categories: And the Tradition to 1800*. Cambridge University Press,, Cambridge, 1970. Accessed digitally printed version 2010.

Ruslan Mitkov. Introduction: Special Issue on Anaphora Resolution in Machine Translation and Multilingual NLP. *Machine Translation*, 14:159–161, 1999a.

Ruslan Mitkov. Anaphora Resolution: The State of the Art. Technical report, School of Languages and European Studies, University of Wolverhampton, 1999b.

Ruslan Mitkov, Sung-Kwon Choi, and Randall Sharp. Anaphora Resolution in Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 5–7, Leuven, Belgium, 1995.

Christoph Müller and Michael Strube. Multi-Level Annotation of Linguistic Data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany, 2006.

Makoto Nagao. A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, Lyon, France, 1984. Elsevier North-Holland, Inc.

Hiromi Nakaiwa and Satoru Ikehara. Intrasentential Resolution of Japanese Zero Pronouns in a Machine Translation System Using Semantic and Pragmatic Constraints. In *Semantic Constraints Viewed from Ellipsis and Inter-Event Relations (in Japanese), IEICE-WGNLC*, pages 96–105, 1995.

Toshiaki Nakazawa, Kun Yu, Daisuke Kawahara, and Sadao Kurohashi. Example-based Machine Translation based on Deeper NLP. In *Proceedings of the 3rd International Workshop on Spoken Language Translation (IWSLT 2006)*, pages 64–70, Kyoto, Japan, 2006.

Karin Naumann and V. Möller. Manual for the Annotation of in-document Referential Relations. Technical report, Universität Tübingen Seminar für Sprachwissenschaft, 2007. URL `http://www.sfs.uni-tuebingen.de/resources/tuebadz-coreference-manual-2007.pdf`.

Vincent Ng. Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Com-*

putational Linguistics, ACL '10, pages 1396–1411, Uppsala, Sweden, 2010. Association for Computational Linguistics.

Michal Novák. Utilization of Anaphora in Machine Translation. In *Proceedings of Contributed Papers, Week of Doctoral Students 2011*, pages 155–160, 2011.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. Translation of "it" in a deep syntax framework. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 51–59, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Sapporo, Japan, 2003. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania, 2002. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440, Sydney, Australia, 2006. Association for Computational Linguistics.

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. Discourse-level Annotation over Europarl for Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2716–2720, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of CoNLL 2011*, pages 1–27, Portland, Oregon, 2011. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 492–501, Cambridge, Massachusetts, 2010. Association for Computational Linguistics.

Rudolf Rosa. Depfix, a Tool for Automatic Rule-based Post-editing of SMT. *The Prague Bulletin of Mathematical Linguistics*, 102:47–56, 2014.

Nicholas Ruiz and Marcello Federico. Complexity of Spoken Versus Written Language for Machine Translation. In *Proceedings of the 17th annual conference of the European Association for Machine Translation, EAMT 2014*, pages 173–180, Dubrovnik, Croatia, 2014.

Lorenza Russo, Sharid Loáiciga, and Asheesh Gulati. Improving Machine Translation of Null Subjects in Italian and Spanish. In *Proceedings of the Thirteenth Conference on European Chapter of the Association for Computational Linguistics - Student Research Workshop*, pages 81–89, Avignon, France, 2012a.

Lorenza Russo, Sharid Loáiciga, and Asheesh Gulati. Italian and Spanish Null Subjects. A Case Study Evaluation in an MT Perspective. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1779–1784, Istanbul, Turkey, 2012b. European Language Resources Association (ELRA).

Benoît Sagot. The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2744–2751, Valletta, Malta, 2010. European Language Resources Association (ELRA).

Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands, 1986.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, 2006.

Michael Strube. *Corpus-Based and Machine Learning Approaches to Anaphora Resolution*. Anaphors in Text: Cognitive, Formal and Applied Approaches to Anaphoric Reference. John Benjamins Pub Co., 2007.

Michael Strube and Udo Hahn. Functional Centering: Grounding Referential Coherence in Information Structure. *Computational Linguistics*, 25(3):309–344, September 1999.

Joel R. Tetreault. A Corpus-based Evaluation of Centering and Pronoun Resolution. *Computational Linguistics*, 27(4):507–520, December 2001.

Jörg Tiedemann. Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 8–15, Uppsala, Sweden, 2010a. Association for Computational Linguistics.

Jörg Tiedemann. To Cache or Not To Cache? Experiments with Adaptive Models in Statistical Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 189–194, Uppsala, Sweden, July 2010b. Association for Computational Linguistics.

Jörg Tiedemann. Baseline Models for Pronoun Prediction and Pronoun-Aware Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 108–114, Lisbon, Portugal, 2015. Association for Computational Linguistics.

Dániel Varga, Németh László, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel Corpora for Medium Density Languages. In *Proceedings of RANLP 2005*, pages 590–596, 2005.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the ACL-08: HLT Demo Session*, pages 9–12, Columbus, Ohio, 2008. Association for Computational Linguistics.

Yannick Versley, Kathrin Beck, Erhard Hinrichs, and Heike Telljohann. A Syntax-first Approach to High-quality Morphological Analysis and Lemma Disam-

biguation for the TüBa-D/Z Treebank. In *Proceedings of the 9th Conference on Treebanks and Linguistic Theories (TLT9)*, Tartu, Estland, 2010.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 167–170, Columbus, Ohio, 2008. Association for Computational Linguistics.

Katie Wales. *Personal pronouns in present-day English*. Cambridge University Press, 1996.

Bonnie L. Webber. *A Formal Approach to Discourse Anaphora*. PhD thesis, Harvard University, 1978.

Bonnie Lynn Webber. Tense As Discourse Anaphor. *Computational Linguistics*, 14(2):61–73, June 1988.

Jochen Weiner. Pronominal Anaphora in Machine Translation. Master's thesis, Karlsruhe Institute of Technology, 2014.

Ralph Weischedel and Ada Brunstein. BBN Pronoun Coreference and Entity Type Corpus, 2005. URL `http://catalog.ldc.upenn.edu/LDC2005T33`.

Ralph Weischedel, Martha Palmer, Marcus Mitchell, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. OntoNotes Version 4.0, 2011. URL `http://catalog.ldc.upenn.edu/LDC2011T03`.

Dekai Wu and Pascale Fung. Semantic Roles for SMT: A Hybrid Two-pass Model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 13–16, Boulder, Colorado, 2009. Association for Computational Linguistics.

Tong Xiao, Jingbo Zhu, and Shujie Yao. Document-Level Consistency Verification in Machine Translation. In *Proceedings of MT summit XIII*, pages 131–138, Xiamen, China, 2011.