ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT

Liane Guillou*, Christian Hardmeier†, Aaron Smith†, Jörg Tiedemann† and Bonnie Webber*
University of Edinburgh*, University of Uppsala†
L.K.Guillou@sms.ed.ac.uk, christian.hardmeier@lingfil.uu.se, aaron.smith.4159@student.uu.se,
jorg.tiedemann@lingfil.uu.se, bonnie@inf.ed.ac.uk

Abstract

We present ParCor, a parallel corpus of texts in which pronoun coreference – reduced coreference in which pronouns are used as referring expressions – has been annotated. The corpus is intended to be used both as a resource from which to learn systematic differences in pronoun use between languages and ultimately for developing and testing informed Statistical Machine Translation systems aimed at addressing the problem of pronoun coreference in translation. At present, the corpus consists of a collection of parallel English-German documents from two different text genres: TED Talks (transcribed planned speech), and EU Bookshop publications (written text). All documents in the corpus have been manually annotated with respect to the type and location of each pronoun and, where relevant, its antecedent. We provide details of the texts that we selected, the guidelines and tools used to support annotation and some corpus statistics. The texts in the corpus have already been translated into many languages, and we plan to expand the corpus into these other languages, as well as other genres, in the future.

Keywords: Corpora, Discourse, SMT

1. Introduction

Reduced coreference – coreference in which reduced expressions such as pronouns, verb morphology, or nothing (zero pronouns) are used in place of full referring expressions – is common in all but the most formal of texts. Reduced coreference – *using* reduced referring forms in a text – is the complement of anaphora resolution – *recovering the referent* of such expressions. Languages differ in how and when they use reduced coreference. This continues to challenge Statistical Machine Translation (SMT) systems (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Novák, 2011; Guillou, 2012). We believe that annotated parallel corpora can shed some light on the use and frequency of reduced coreference, providing valuable insights into where and when pronoun translation is necessary, where the target language requires or permits the use of constructions other than pronouns, and where it is acceptable or necessary to drop pronouns.

This paper describes such annotation in a parallel English-German corpus of documents from two textual genres – TED Talks and EU Bookshop publications. It includes details of the texts that we selected, the guidelines and tools used to support annotation and some corpus statistics. The corpus may be used as a gold standard for testing approaches to pronoun translation, as well as a resource for understanding systematic differences in pronoun use. As parallel data also contains valuable information for anaphora resolution, the corpus may also be useful in the development of anaphora resolution systems (Mitkov and Barbu, 2003; Postolache et al., 2006; de Souza and Orăsan, 2011; Hardmeier et al., 2013). To our knowledge, this is the first attempt at building a parallel corpus of pronoun coreference annotation for the purpose of improving SMT.

2. Previous Work

2.1. Analyses of Pronoun Coreference in SMT

Analyses of pronoun coreference in SMT illustrate problems that arise when a pronoun is translated without knowledge of its antecedent. Novák (2011) highlights how an English-Czech TectoMT system (Žabokrtský et al., 2008) always translates the English pronoun "it" into a third-person neuter pronoun in Czech, resulting in 62 out of 81 tokens (76.5%) being translated incorrectly. Le Nagard and Koehn (2010) made similar observations about their English-French phrase-based SMT system: The English pronouns "it" and "they" were too often translated into masculine forms in French. They reported pronoun translation accuracy at 69%. Because the 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-English translation. They claim that the extent of the errors is different depending on the type of pronoun being translated, an effect caused by the interplay between morphological syncretism in the two languages and the cross-lingual alignment of the morphological paradigms.

2.2. Coreference-Annotated Parallel Corpora

We are aware of only two parallel corpora in which pronoun coreference is annotated in some way. Popescu-Belis et al. (2012) used *translation spotting* to annotate pronouns (Cartoni et al., 2011): That is, each pronoun in the source training data was manually annotated with its translation from the target, thereby avoiding the need for anaphora resolution in the translation pipeline. They used this method to annotate ∼400 tokens of the pronoun "it" in English-French Europarl data, which were then used to train classifiers to predict the French translation of new instances of "it". This resulted in a small but significant increase in BLEU score.

The second parallel corpus annotated with pronoun coreference is the Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0) (Hajič et al., 2012), which is a close translation into Czech of the Penn Treebank corpus (Marcus et al., 1993). The PCEDT corpus offers rich linguistic annotation over several treebank layers. It was originally designed as a multi-purpose linguistic resource, just like the Penn

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

Table 1: Documents taken from the EU Bookshop online archive

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

Table 2: Documents taken from the TED Talks in the IWSLT2013 2010 test set

Treebank. It has, however, been used in two recent SMT experiments (Guillou, 2012; Meyer and Poláková, 2013). Guillou (2012) takes advantage of the parallel annotation of a subset of pronoun coreference in the BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005). Meyer and Poláková (2013) exploit the parallel annotation of discourse connectives in the Penn Discourse TreeBank (Prasad et al., 2008).

3. Data

The corpus described here was designed specifically for use in SMT. It thus comprises a collection of documents taken from large multilingual parallel SMT corpora. The remainder of the data (i. e., that which has not been annotated) in these larger corpora may then serve as in-domain training data for the purpose of creating SMT systems or as additional texts (and languages) to be annotated.

Because reliance on reduced coreference varies across genres, we felt it important that the corpus include texts from at least two different genres. We selected a collection of English texts (and their German translations) from the EU Bookshop online archive (EUB) (Table 1) and (transcriptions of) TED Talks from the IWSLT2013 test set (TED) (Table 2)[1].

The EU Bookshop contains texts written for an educated but non-expert public. The TED Talks are orally-delivered public lectures. Neither the texts nor the talks require any specific expertise, thereby simplifying the annotation task. Nevertheless, as will be evident in the next section, the two genres differ markedly in their use of pronoun coreference.

4. Statistics

We provide counts for both pronoun *type* (Table 3) and *form* (Table 4). Because there are differences in some of the pronoun types annotated for the TED Talks and EU Bookshop corpora, some type categories are marked as not applicable in the tables. The differences in annotation guidelines are described in Section 7.

Differences in Table 3 suggest differences in the use of pronouns in English and German, with more anaphoric pronouns and many more pleonastic pronouns marked in the German part of each corpus. However, a true comparison will require further analysis of the data to determine the extent to which these are systematic differences between English and German versus the effect of translation from English. All documents were written or spoken in English, with the possible exception of one EU Bookshop document whose source language is unknown. The counts displayed in the table 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 "removal" of another from a different sentence will cancel out).

4.1. TED Talks

Talks at TED conferences ("TED Talks") address topics of general interest and are delivered to a live public audience. They are also recorded for online viewing by other members of the public, all around the world. They generally aim to be persuasive and to change viewers' behaviour or beliefs. The genre of the TED Talks is transcribed planned speech. The texts we annotated were taken from the test sets prepared for the IWSLT 2013 machine translation shared task distributed

[1] Sentence alignments computed with LFAligner: http://sourceforge.net/projects/aligner/

Pronoun Type	TED Talks		EU Bookshop	
	English	German	English	German
Anaphoric	887	1,226	2,767	3,036
Anaphoric (pronominal adverb)	N/A	N/A	70	84
Cataphoric	5	16	67	19
Event	266	331	239	255
Event (pronominal adverb)	N/A	N/A	0	78
Extra-textual reference	52	26	N/A	N/A
Pleonastic (non-referential)	61	223	191	391
Addressee reference	497	395	112	75
Speaker reference	656	789	536	567
Generic	N/A	N/A	9	58
Pronoun (other)	N/A	N/A	135	126
Pronoun (unsure)	N/A	N/A	14	0
Total	2,424	3,006	4,140	4,689

Table 3: Pronoun **type** counts for English and German texts in the TED Talks and EU Bookshop portions of the corpus

Pronoun Form	TED Talks		EU Bookshop	
	English	German	English	German
First-person personal	562	712	420	449
Second-person personal	454	395	79	75
Third-person personal	950	881	1,491	1,458
Possessive	218	163	1,001	905
Relative/Demonstrative	210	771	984	1,547
Reflexive	30	84	N/A	N/A
Generic	N/A	N/A	9	58
Pronominal Adverbs	N/A	N/A	70	164
Other	N/A	N/A	86	33
Total	2,424	3,006	4,140	4,689

Table 4: Pronoun **form** counts for English and German texts in the TED Talks and EU Bookshop portions of the corpus

with the WIT^3 corpus (Cettolo et al., 2012). Besides English and German, WIT^3 contains translations of the same texts into 11 other languages.

With respect to pronouns, TED speakers frequently use first and second-person pronouns (singular and plural) – first-person to refer to themselves and their colleagues or to themselves and the audience, second-person to refer to the audience, the larger set of viewers, or people in general. They also use the pronoun "they" without a specific textual antecedent, in phrases such as "This is what they think", as well as using deictic and third-person pronouns to refer to things in the speaker's spatio-temporal context, such as props and slides.[2]

Only one translation of a given TED Talk is included per language. This is important, as the presence of multiple translations in a given language of the same source text could lead to variation in pronoun use between the translations as a side effect of the translation process (which could, of course, be interesting as well). Translations are provided by (named) volunteers.

4.2. EU Bookshop

The EU Bookshop provides a range of documents on topics connected with the EU's activities and policies. While the documents are intended for a wide audience (hence, for non-experts), they were produced by European institutions and thus have a fairly formal style. Translations have been provided by professional translation companies and are available in a number of European languages.

The documents in our corpus (Table 1) were originally written in English[3] and then translated into German and other languages. These documents were selected because they are available as E-books, making it easy to extract the raw text using the Calibre E-book management tool[4]. So far, only the English and German texts have been annotated. As a collection, they contain a good balance of personal and demonstrative/relative pronouns. Generic pronouns are rather rare. Pronoun distribution within documents varies, particularly with respect to the use of speaker and addressee reference which is used throughout in some cases, or limited to specific sections of the document in others. Anaphoric reference is largely restricted to entities explicitly mentioned in the text.

Additional EU Bookshop data for training SMT systems is provided in the EUBookshop corpus available via OPUS[5]. The documents that we annotated post-date the collection of data for the EUBookshop corpus on OPUS.

5. Annotation Process and Guidelines

For each language and each genre, two annotators worked in parallel until agreement on the annotation features described

[2] Each online TED Talk has both a video and a transcript.

[3] With the exception of document MI3112464 for which the source language could not be confirmed.

[4] `http://calibre-ebook.com/`

[5] `http://opus.lingfil.uu.se/EUbookshop.php`

in Section 8 was sufficiently high that a single annotator would suffice for the remaining texts. Our annotators were all native speakers of the language they were annotating.

The MMAX-2 annotation tool (Müller and Strube, 2006) was used throughout, with automated pre-processing pipelines used to generate MMAX-2 markables for pronouns and candidate noun phrases (NPs), as the starting point for manual annotation. These pipelines are described in Section 6 and the inter-annotator agreement scores are presented in Section 8.

Annotation guidelines were adapted from the pronoun annotation guidelines in the MUC-7 Coreference Task Definition (Chinchor and Hirschman, 1998). The annotation guidelines for the EU Bookshop and TED Talks annotation tasks were largely similar, as the aim is to provide comparable annotation. However, there are a number of genre and language-specific differences. These are summarised in Section 7. Our annotated corpus together with the annotation guidelines used by our annotators during the manual annotation phase is publicly available[6].

6. Automatic Preprocessing

To reduce manual effort and improve inter-annotator agreement, we aimed to start with pre-defined pronoun and NP markables. These are generated by separate pipelines for English and German.

The English pipeline starts by defining markables using the Berkeley Parser (Petrov et al., 2006) to identify NPs and pronouns. It then uses NADA (Bergsma and Yarowsky, 2011) to identify instances of pleonastic "it" (with no equivalent system for German) and the Stanford Dependency Parser (de Marneffe et al., 2006) to identify whether instances of the pronoun "it" are in subject or non-subject position. In addition, other markables are recognised using predefined lists, including pronominal adverbs (e. g., "thereafter", "herein") and idiomatic pronoun expressions ("he or she", "his or her(s)", "him or her" and "s/he") which should be treated as a single pronoun. As pronouns used as speaker and addressee reference are unambiguous in English, we also use predefined lists to automatically set the *type* of these pronouns.

The German pipeline is described in the papers by Broscheit et al. (2010) and Versley et al. (2010). It first parses the texts using the Berkeley Parser and then extracts nominal (both minimal and maximal noun projections) and pronominal mentions from the parse trees. The morphological tagging described by Broscheit et al. (2010) provides number and gender information as well as the mention type (definite/indefinite NP, name, personal/relative/reflexive pronoun). We extract the information contained in the top level "markable" xml file and from it, construct our own format xml file to match the annotation scheme in Section 7.

Annotators are presented with the output of the relevant pipeline, as a starting point for their manual annotation.

7. Annotation Guidelines

7.1. General Principles

As noted in Section 1, the treatment of reduced coreference in SMT requires a better understanding of this phenomenon, how frequently the different forms occur and how they appear in corresponding translations. Our focus is, therefore, set on pronouns and their coreferential properties. The annotation process aims to mark all pronouns in each text (personal, possessive, demonstrative, relative, adverbial and generic) as being one of eight types: Anaphoric/cataphoric reference, event reference, extra-textual reference, pleonastic, addressee reference, speaker reference, generic reference, or other function (see Section 7.9).

Annotation of the two corpora differed in minor ways. For example, reflexive pronouns were annotated in the TED Talks corpus but not in the EU Bookshop corpus, since they are rare in the latter.

Here we explain in more detail the various pronoun categories and the differences between the two corpora.

7.2. Anaphoric and Cataphoric Pronouns

Anaphoric pronouns refer to explicitly mentioned entities, where the entity precedes the pronoun in the text. In the example "David Cameron is the Prime Minister of the United Kingdom; **he** is 47 years old", the pronoun "he" is labelled as *anaphoric*, and linked to the NP "David Cameron" – the nearest non-pronominal antecedent.

In the annotation of TED Talks, anaphoric pronouns are by default also sub-classified with the label *simple antecedent*, implying that they refer to a single NP. This can be changed to *split reference*, when the pronoun replaces two or more NPs (in which case it should be linked to each of the antecedents), or *no explicit antecedent*. This final option was introduced to deal with cases observed occasionally in the TED Talks corpus such as "In this study **they** took 100 people and split them into two groups", where the pronoun "they" has no explicit antecedent. In the EU Bookshop annotation, the latter is called *anaphoric but no specific antecedent*. Antecedents are not explicitly sub-classified into "simple" or "split" in the EU Bookshop corpus – instead, this information can be gleaned from the number of NP (antecedent) markables to which a pronoun is linked.

The NPs marked by the automated pipelines comprise the set of candidate antecedents to which a pronoun may be linked (See Section 6). The annotators were instructed to select antecedents from this set wherever possible. If no suitable NP exists, the next closest match may be expanded such that it spans the necessary text; failing that, a new markable may be created. When amending an existing NP markable or adding a new one, the following rules (taken from the MUC-7 guidelines) apply:

- The markable must contain the head noun.
- If the head is a name, the entire name should be marked. For example, given "Frederick F. Fernwhistle Jr.", it is insufficient to simply mark "Frederick".
- The 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).
- Determiners should be included for definite NPs.

The guidelines for marking one or more antecedent spans are taken from the MUC-7 guidelines and handle the cases of conjoined NPs. They also mirror those used in the Tüba-D/Z

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

corpus (Naumann and Möller, 2007). For example, in "*John and Mary* like watching films. The last time **they** went to the cinema...", the conjoined NP "John and Mary" is marked as the single antecedent span of the plural pronoun "they". However, in the following text "*John* likes documentaries. *Mary* likes films about animals. The last time **they** went to the cinema...", there is intervening text between the entities "John" and "Mary" that is not part of an NP. In this case, each entity is marked as a separate NP, and both NPs are linked to the pronoun "they". The addition of marking whether the antecedent is *simple antecedent* or *split reference* in the TED Talks is simply a clarification of the number of antecedents to which a pronoun is linked. This is not a part of the MUC-7 guidelines but it does not constitute a change to the guidelines.

We also record some morphosyntactic features that are difficult to recover automatically. *Agreement* is needed for those pronouns that may be ambiguous so we record whether they are singular or plural (e. g., "they" can be singular or plural in English). *Position* (subject or object), in English, and *case*, in German, are used to identify the syntactic role of the pronoun in a sentence. In the TED Talks, *audience* denotes whether the audience is included when speaker and addressee reference pronouns are used. This is described in Sections 7.6 and 7.7.

We added some additional rules to cover specific cases. In particular:

- When the pronoun "they" (and its equivalents in German) is used to refer anaphorically/cataphorically to a collective noun (such as "the government"), it should be considered a plural pronoun.
- Due to the lack of un-gendered person pronouns in English, cases such as "he or she" and "s/he" exist. These should be treated as a single pronoun.
- A pronoun may refer to a modifier in an NP. In these cases the pronoun should be linked to the modifier if no other suitable antecedent can be found.
- Reflexive pronouns such as *himself/itself* are labelled in the normal way (with the exception of first-person). In cases like "Here comes the the man himself", the token "himself" is not considered a pronoun, and if identified as so by the automated pre-processing pipeline, should be corrected by the annotator.

Cataphoric pronouns, where the entity that a pronoun refers to occurs after the pronoun in the text, are much less common than anaphoric pronouns. Cataphoric pronouns have their own category but are otherwise treated in exactly the same way as anaphoric pronouns in both corpora.

In the EU Bookshop corpus, we also marked pronominal adverbs. These may take the anaphoric or event function.

7.3. Event Pronouns

We use the *event* category for pronouns that refer to propositions, facts, states, situations, opinions, etc. In the most basic case the *event* category is used in examples like "John arrived late; **this** annoyed Mary", where the pronoun "this" refers to the action of John arriving late, and not an explicit NP. In the TED Talks corpus, *event* is also widely used when a pronoun refers back to whole section of text, or a concept evoked by the text. For example, if the speaker says "**This** got me thinking", where "this" refers to a story she has just told, it would be labelled as *event*. In the EU Bookshop corpus, events usually refer to concrete or hypothetical events such as "...spot prices could decrease and remain low... **This**...". Event pronouns are not linked to any section of text in either corpus. In the TED Talks, two or more event pronouns that refer to the same event are linked together.

Many monolingual coreference-annotated corpora ignore event reference, as do many coreference resolution systems, perhaps because events pose unique challenges and tend to be relatively rare when compared to the number of pronouns and NPs in the data (Pradhan et al., 2011). From the perspective of SMT, event pronouns should be identified so that they can be handled differently in translation. In English to German translation, the event pronouns "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 German, event pronouns refer to verbs, verb phrases, clauses, sentences, etc. and agreement with the main verb is not required. At least in translating between English and German (either direction), event pronouns pose less of a challenge as these can be mapped directly from one language to the other. Nevertheless, we wish to exclude event pronouns from the set of possible distractors when considering the translation of anaphoric instances of "it", "this" and "that".

7.4. 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*. This category is used for deictic pronouns only. It is useful in the TED Talks corpus, where the speaker often refers to items physically present in the room, such as her slides. For example, the speaker might say "The house looked like **this**" whilst pointing at a photo that the listener can see. This category may also be used within quoted text when referring to a third-person, e.g. the "He" in "People when they see me say, '**He**'s a nice guy'".

7.5. Pleonastic

The *pleonastic* category is used in both corpora for pronouns that are syntactically necessary but have no semantic content. Pleonastic pronouns are found in both English and German. For example "It" in "**It** is raining", and "Es" in the equivalent German phrase: "**Es** regnet".

Pleonastic pronouns are typically not marked in monolingual corpora annotated with coreference information. There is no provision for the handling of pleonastic pronouns in the MUC-7 guidelines and they are not marked in OntoNotes (Weischedel et al., 2011) corpora or the BBN Pronoun Coreference and Entity Type corpus (Weischedel and Brunstein, 2005). However, they are explicitly marked in the Tüba-D/Z corpus as per the coreference annotation guidelines (Naumann and Möller, 2007).

The removal of instances of pleonastic "it" has been used by a number of coreference resolution systems including the sieve-based Stanford Coreference Resolution System

(Lee et al., 2011). In a similar way to the removal of such instances in coreference resolution, we also wish to identify pleonastic pronouns for the purpose of SMT. As with event pronouns, pleonastics belong to the set of distractors when considering the translation of anaphoric instances of "it".

7.6. Addressee Reference

In the EU Bookshop corpus the *addressee reference* category is used for pronouns that refer to the person being addressed; usually the second-person pronouns "you" and "your" (and their German equivalents).

In the TED Talks corpus, second-person pronouns are always labelled as *addressee reference*. The job of the annotator is to sub-classify and decide whether the audience is *deictic*, meaning that the speaker is referring to the audience or a specific person, or *generic*, as in phrases such as "In England, if **you** own a house **you** have to pay taxes".

When a speaker uses deictic "you", addressing a whole audience, it is always marked as plural, even in cases like "Imagine **you**'re walking alone in the woods".

7.7. Speaker Reference

In the EU Bookshop corpus the *speaker reference* category is used for pronouns that refer to the speaker; usually first-person pronouns. The reference may or may not also include the addressee. Plural pronouns "we", "us" and "our" (and their German equivalents) are labelled as *speaker reference*. In these texts singular first-person pronouns are rare, but are marked when they do occur.

In the TED Talks corpus singular first-person pronouns are identified automatically and labelled as *speaker reference*. Plural first-person pronouns, meanwhile, are annotated manually. They are always labelled as *speaker reference*, and then sub-classified as *exclusive*, meaning the speaker and her clique but not the audience, *co-present*, meaning the speaker plus everyone physically present in the room, or *all-inclusive*, incorporating everything else.

7.8. Generic

The *generic* category was used in the EU Bookshop corpus for pronouns that refer to an unspecified person, such as "you" and "one" in English and "man" in German. In the TED Talks corpus this category was not used; generic pronouns were always labelled as *addressee reference* or *speaker reference*.

Generic pronouns are not marked in OntoNotes, or the BBN Pronoun Coreference and Entity Type corpus and there is no provision for generic pronouns in the MUC-7 guidelines. In the Tüba-D/Z corpus (Naumann and Möller, 2007), the generic German pronoun "man" is labelled, but as "indefinite". We opted to use more specific labels. Again, we wish to mark generic pronouns so as to handle them differently to referential instances of the same pronoun form in SMT.

7.9. Pronoun

The *pronoun* category was used for words that are clearly pronouns but do not belong to any of the above categories. For example, indefinite pronouns (e.g. "anyone") and some numbers/quantifiers used as pronouns but are not themselves bare pronouns (e.g. "others", "each", "both"). Such pronouns are simply labelled as *pronoun* and no additional features are recorded. This category was only used in the EU Bookshop corpus.

Indefinite pronouns are marked in the Tüba-D/Z corpus (Naumann and Möller, 2007), but the annotation guidelines also specify that two or more indefinite pronouns may be linked together.

7.10. Difficult Cases

The English EU Bookshop section of the corpus contains a small number of pronouns whose type is unclear: Both event or anaphoric readings are possible and would make sense. In these cases, we mark the pronoun as anaphoric. If it is impossible to determine, the pronoun is labelled as "unsure" (see the "Pronoun (unsure)" entry in Table 3). These problems were not identified in the German translations or in the annotation of the TED Talks.

7.11. Pronouns in Quoted Text

The annotation of pronouns in direct quotes is more complex. Because direct quotes occur only infrequently in our texts, we developed simple guidelines for the annotation of first and second-person pronouns in quoted text. Third-person personal pronouns are marked according to the relevant guidelines above.

In the TED Talks, pronouns in quoted text are annotated strictly from the point of view of the quoted speaker, not of the speaker who quotes the utterance. First-person pronouns are always labelled as *speaker reference* and second-person pronouns as *addressee reference*. Coreference relations between a first-person or second-person pronoun inside quoted speech and a pronoun outside the quoted speech passage are not marked (as in "He said, 'I do.'", where you could arguably mark "He" and "I" as coreferent.)

In the EU Bookshop documents all first and second-person pronouns within quoted text are simply labelled as *pronoun* to indicate that they have been seen by the annotator. The surface form of such pronouns indicates whether they are speaker/addressee reference pronouns. In some cases, specific to the EU Bookshop documents, 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 and speaker/addressee reference pronouns are annotated as normal.

7.12. Exclusions

The annotation of the corpus is limited to pronoun coreference. As such, full coreference chains/sets are not provided. 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. This is commonly accommodated in annotation guidelines and annotated corpora such as the MUC-7 guidelines and OntoNotes. However, it is not necessarily useful for SMT where head finding techniques will be required for all antecedents of coreferential pronouns (not just appositives) to ensure agreement holds between the pronoun and head noun. Implicit pronouns are not annotated – that is, we follow the MUC-7 guidelines in assuming that English has no zero pronouns and extend this assumption to

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 5: IAA Scores for English and German EU Bookshop documents

German. In practice, this means that the empty string is not considered to be a markable.

8. Inter-Annotator Agreement

Inter-Annotator Agreement (IAA) was measured using Cohen's Kappa (Cohen, 1960) for a number of annotation features. We separately measured IAA for pronoun *type*, *agreement* (to disambiguate pronouns such as "they" which can be singular or plural), *position* (English only) and *case* (German only), and *audience* (TED Talks only). 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.

The annotation of the EU Bookshop German texts preceded that of the English texts, so IAA was calculated for two German documents to assure the quality of the annotation guidelines. For English, we were already using a stable annotation scheme and IAA was therefore only computed for a single document (Table 5). IAA scores for the TED Talks corpus (Table 6) 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 changes. Secondly, our German annotator was already familiar with the annotation guidelines used in the EU Bookshop annotation and even provided us with assistance in ensuring that the options for the additional features were captured using German equivalents in the MMAX-2 templates. Computing IAA for English TED Talks therefore serves to ensure that the changes to the annotation guidelines do not adversely affect the quality of the annotations.

9. Conclusions and Future Work

This is the first report on a growing corpus of parallel texts in which pronouns have been manually annotated for location, type and, where appropriate, antecedents. Establishing correspondences between reduced coreferring forms in parallel texts should allow us to improve the realisation of coreferring forms in SMT, improving both fluency and accuracy.

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 6: IAA Scores for English TED Talks

Future work will continue on two main tracks. Firstly, we plan to use the corpus to build SMT systems with a specific focus on improving the translation of pronoun coreference. We are keen to encourage participation from other SMT researchers and plan to introduce a shared task on coreference translation in the near future.

Secondly, we will continue working on corpus development, expanding the existing corpus to include additional documents from the existing genres as well as new languages and text genres. We would welcome involvement from researchers at other institutions who are interested in participating in these efforts. Manual annotation, however, is both time consuming and expensive to produce. Future work on annotation will therefore focus on expanding the capabilities of the automated pipelines in order to provide additional information in the partial annotation that is presented to human annotators as a starting point for manual annotation. Work on developing methods to resolve *addressee reference* pronouns and provide the *audience* information described in Section 7.6, has already begun at Edinburgh.

10. Acknowledgements

We would like to thank our annotators, Susanne Tauber, Petra Strom, Samuel Gibbon and David Lawrence for their many hours of painstaking work and Yannick Versley for making his German pre-processing pipeline available to us. We would also like to thank the three anonymous reviewers for their feedback.

The work carried out at Edinburgh University was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE).

The work at Uppsala University was supported by the Swedish Research Council (Vetenskapsrådet) through the project on Discourse-Oriented Machine Translation (2012-916).

11. References

Bergsma, S. and Yarowsky, D. (2011). NADA: A Robust System for Non-Referential Pronoun Detection. In *Proceedings of DAARC 2011*, pages 12–23.

Broscheit, S., Ponzetto, S. P., Versley, Y., and Poesio, M. (2010). Extending BART to Provide a Coreference Resolution System for German. In *Proceedings of LREC 2010.*

Cartoni, B., Zufferey, S., Meyer, T., and Popescu-Belis, A. (2011). How Comparable Are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. In *Proceedings of BUCC 2011*, pages 78–86.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of EAMT 2012*, pages 261–268.

Chinchor, N. and Hirschman, L. (1998). MUC-7 Coreference Task Definition (Version 3.0). In *Proceedings of MUC-7*.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1).

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC 2006*, pages 449–454.

de Souza, J. and Orăsan, C. (2011). Can Projected Chains in Parallel Corpora Help Coreference Resolution? In Hendrickx, I., Lalitha Devi, S., Branco, A., and Mitkov, R., editors, *Anaphora Processing and Applications*, volume 7099 of *Lecture Notes in Computer Science*, pages 59–69. Springer.

(EUB). EU Bookshop Online Archive. https://bookshop.europa.eu. Accessed: 2014-03-07.

Guillou, L. (2012). Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop*, EACL 2012, pages 1–10.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of LREC 2012*.

Halliday, M. A. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Hardmeier, C. and Federico, M. (2010). Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*.

Hardmeier, C., Tiedemann, J., and Nivre, J. (2013). Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction. In *Proceedings of EMNLP*.

Le Nagard, R. and Koehn, P. (2010). Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's Multipass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of CoNLL-2011*, pages 28–34.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Meyer, T. and Poláková, L. (2013). Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the 1st Workshop on Discourse in Machine Translation*, ACL 2013, pages 43–50.

Mitkov, R. and Barbu, C. (2003). Using Bilingual Corpora to Improve Pronoun Resolution. *Languages in Contrast*, 4(2):201–211.

Müller, C. and Strube, M. (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In Braun, S., Kohn, K., and Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang.

Naumann, K. and Möller, V. (2007). Manual for the Annotation of in-document Referential Relations. Technical report, Universität Tübingen Seminar für Sprachwissenschaft.

Novák, M. (2011). Utilization of Anaphora in Machine Translation. In *Proceedings of Contributed Papers, Week of Doctoral Students 2011*, pages 155–160.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of ACL 2006*, pages 433–440.

Popescu-Belis, A., Meyer, T., Liyanapathirana, J., Cartoni, B., and Zufferey, S. (2012). Discourse-level Annotation over EuroParl for Machine Translation. In *Proceedings of LREC 2012*, pages 2716–2720.

Postolache, O., Cristea, D., and Orăsan, C. (2006). Transferring Coreference Chains through Word Alignment. In *Proceedings of LREC 2006*, pages 889–892.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of CoNLL 2011*, pages 1–27.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008*.

(TED). IWSLT2013/WIT TED Talks. https://wit3.fbk.eu. Accessed: 2014-03-07.

Versley, Y., Beck, K., Hinrichs, E., and Telljohann, H. (2010). A Syntax-first Approach to High-quality Morphological Analysis and Lemma Disambiguation for the TüBa-D/Z Treebank. In *Proceedings of TLT9*.

Weischedel, R. and Brunstein, A. (2005). BBN Pronoun Coreference and Entity Type Corpus.

Weischedel, R., Palmer, M., Mitchell, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., and Houston, A. (2011). OntoNotes version 4.0.

Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, ACL 2008, pages 167–170.