Assessing the applicability of annotation projection methods for coreference relations

Yulia Grishina

Doctoral Thesis

submitted to the Faculty of Human Sciences,
Department of Linguistics at the University of Potsdam

in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy (Dr. phil.)
in Computational Linguistics

Supervisor: Prof. Dr. Manfred Stede

Place and date of defense:
Potsdam, February 12, 2019

Submitted February 26, 2018
Defended February 12, 2019

Reviewers:
Prof. Dr. Manfred Stede
Prof. Dr. Heike Zinsmeister

Published online at the
Institutional Repository of the University of Potsdam:
https://doi.org/10.25932/publishup-42537
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-425378

Abstract

The main goal of this thesis is to explore the feasibility of using cross-lingual annotation projection as a method of alleviating the task of manual coreference annotation. To reach our goal, we build the first trilingual parallel coreference corpus that encompasses multiple genres. For the annotation of the corpus, we develop common coreference annotation guidelines that are applicable to three languages (English, German, Russian) and include a novel domain-independent typology of bridging relations as well as state-of-the-art near-identity categories.

Thereafter, we design and perform several annotation projection experiments. In the first experiment, we implement a direct projection method with only one source language. Our results indicate that, already in a knowledge-lean scenario, our projection approach is superior to the most closely related work of Postolache et al. (2006). Since the quality of the resulting annotations is to a high degree dependent on the word alignment, we demonstrate how using limited syntactic information helps to further improve mention extraction on the target side. As a next step, in our second experiment, we show how exploiting two source languages helps to improve the quality of target annotations for both language pairs by concatenating annotations projected from two source languages. Finally, we assess the projection quality in a fully automatic scenario (using automatically produced source annotations), and propose a pilot experiment on manual projection of bridging pairs.

For each of the experiments, we carry out an in-depth error analysis, and we conclude that noisy word alignments, translation divergences and morphological and syntactic differences between languages are responsible for projection errors. We systematically compare and evaluate our projection methods, and we investigate the errors both qualitatively and quantitatively in order to identify problematic cases. Finally, we discuss the applicability of our method to coreference annotations and propose several avenues of future research.

Zusammenfassung

Ziel dieser Dissertation ist, die Durchführbarkeit von crosslingualer Annotationsprojektion als Methode zur Erleichterung der manuellen Koreferenzannotation zu erproben. Um dieses Ziel zu erreichen, wird das erste dreisprachige parallele Koreferenzkorpus gebaut, das mehrere Textsorten umfasst. Für die Korpusannotation werden gemeinsame Annotationsrichtlinien entwickelt, die auf drei Sprachen anwendbar sind (Englisch, Deutsch, Russisch) und eine neue domänenunabhängige Typologie von indirekten Wiederaufnahmen und sogenannten Near-Identity-Kategorien enthalten.

Danach werden mehrere Projektionsexperimente entworfen und durchgeführt. Im ersten Experiment wird eine direkte Projektionsmethode mit nur einer Ausgangssprache implementiert. Die Ergebnisse zeigen, dass bereits in einem wissensarmen Szenario der vorgeschlagene Projektionsansatz die Resultate der verwandten Arbeit von Postolache et al. (2006) übertrifft. Da die Qualität der resultierenden Annotationen in hohem Maße von der Wortalignierung abhängig ist, zeigen wir, wie die Verwendung begrenzter syntaktischer Informationen weiterhilft, die Extraktion von referierenden Ausdrücken auf der Zielseite zu verbessern. Im nächsten Schritt, dem zweiten Experiment, demonstrieren wir, wie die Nutzung von zwei Ausgangssprachen zur weiteren Verbesserung der Qualität der Zielannotationen für beide Sprachpaare beiträgt, indem die Annotationen aus zwei Quellsprachen kombiniert werden. Schließlich wird die Projektionsqualität noch in einem vollautomatischen Szenario ausgewertet (mit automatisch erstellten Quellannotationen), und ein Pilotversuch zur manuellen Projektion von Paaren indirekter Wiederaufnahmen vorgestellt.

Für jedes Experiment wird eine detaillierte Fehleranalyse durchgeführt. Daraus schließen wir, dass fehlerhafte Wortalignierungen, Übersetzungsdivergenzen und morphologische sowie syntaktische Unterschiede zwischen den Sprachen für die Projektionsfehler verantwortlich sind. Hierzu werden die Projektionsmethoden systematisch verglichen und ausgewertet, und die Fehler sowohl qualitativ als auch quantitativ untersucht, um problematische Fälle zu identifizieren. Zum Schluss wird die Anwendbarkeit unserer Methode für Koreferenzannotationen diskutiert, und es werden Ansatzpunkte für weiterführende Forschung vorgeschlagen.

Acknowledgements

First and foremost, I am immensely grateful to my supervisor Manfred Stede, who introduced me to the research community and gave me the opportunity to learn from him. I am sincerely thankful to him for sharing his insights and expertise, and for his constructive criticism on a number of issues related to this work. Apart from that, he taught me structure, and how to think several steps ahead.

For her support, enthusiasm and honesty, I am indebted to Tatjana Scheffler, who gave me a helping hand on various occasions. She was always open and reachable, and encouraged me to stay focused and motivated. I am also thankful to Wladimir Sidorenko for being eager to help at all times as well as share his knowledge on technical and non-technical matters. Moreover, I am grateful to my office mates at the Linguistics Department – Andreas Peldszus, Arne Neumann, Fatemeh Torabi Asr, Peter Bourgonje and Maria Skeppstedt – for our fruitful discussions and exchange of opinions. Also, I am thankful to the student assistants – Erik Haegert and Mathias Bisping – for their help with the corpus annotation.

It was a privilege to have been able to visit the NLP group in Copenhagen. I am grateful to Anders Søgaard for hosting my stay and sharing his vision, and I also thank my colleagues at the University of Copenhagen – among others, Dirk Hovy, Joachim Bingel and Maria Barrett, – who navigated me through their research and made me feel very welcome.

In addition, I am grateful to the people I met at various NLP conferences, who provided important feedback to my work. In particular, I would like to express my gratitude to Bonnie Webber for her invaluable insights and challenging questions. Also, I am thankful to the anonymous reviewers for their comments that greatly improved earlier parts of this work.

I would like to acknowledge the Collaborative Research Center (SFB) 632 and the Friedrich-Wingert Foundation for substantially supporting this thesis. Similarly, I am thankful to Potsdam Graduate School and FNK for enabling my travel and conference attendance, and I am also grateful to TextLink (COST Action IS1312) for enabling my research stay. I thank Annett Esslinger for assisting with the paperwork afterwards.

Finally, I am genuinely grateful to my parents and to my husband Tobias. Without their love, understanding, humor and support this thesis would simply never have happened.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem description

Coreference is a linguistic phenomenon that occurs when two or more expressions in a text point to the same entity in the real world. Coreference resolution is the task of identifying all mentions of the same entity in a natural language discourse. For example, the entity *Mr. Baccini* (see example (1)) may be introduced in different ways: by its name (*Frank Baccini*), by a definite description (*her client Frank Baccini, the owner of a warehouse of electrical goods*) or simply by a pronoun (*he*).

(1) Daisy picked up the telephone and tried to get through to [her client]$_1$ again. [Her client Frank Baccini, the owner of a warehouse of electrical goods]$_1$, had not paid for her two days' work. Daisy had managed to discover where [he]$_1$ was now living and was anxiously expecting her cheque.[1]

Linguistic theory shows that the preference towards a certain linguistic form for an entity is determined by the degree of accessibility and salience: The least salient mentions need to be introduced by a complete description while the most salient ones can be expressed by a pronoun. In other words, accessibility and salience are driven by the information status of discourse referents, as described in, among others, (Ariel, 1985, 2001). 'Given' discourse referents that were already introduced in discourse can easily be inferred by the reader and therefore can be expressed by using low accessibility markers (such as pronouns), while 'new' discourse referents are not that easily accessible and therefore require additional information to be provided to the

[1]The example is taken from the coreference corpus developed as part of this work.

reader. To exemplify, in (1), when the discourse referent *Frank Baccini* is being introduced, he is described to the reader using two definite descriptions. Thereafter, he can be unambiguously referred to by a personal pronoun, since it is the only male discourse referent in the context.

Coreference is a complex linguistic phenomenon that occurs not only between referring expressions with identical discourse referents. For instance, in (2), the reader is able to infer that *the door* refers to the previously mentioned *Mr. Baccini's warehouse*, although this is not explicitly stated in the sentence.

(2) Daisy was furious and decided to go to [Mr. Baccini's warehouse]$_{B1}$ to see if he was there. When she arrived, she knocked on [the door]$_{B1}$.[2]

Such cases, the inference of which is based solely on the common background shared by the speaker and the listener, are also comprised by the term coreference. They are called indirect coreference, or *bridging*, as initially introduced by Clark (1975), and encompass a wide range of subrelations, such as part-whole, set-subset, etc.

Another complication arises when two referring expressions are *almost* identical, such as *Rome* and *ancient Rome*, but still differ in one dimension (e.g., time). These cases of *near-identity* (as stated in Recasens et al. (2010a)) have recently attracted attention in the field of coreference resolution, as they are hard to identify in the text, but still pose additional difficulties for automatic resolution: For instance, it is not clear whether nominal phrases representing different values of the same concept should be treated as coreferential or not, as in (3).

(3) [The overnight temperature]$_{NI1}$ fell to five degrees below zero, and by midday [it]$_{NI1}$ rose by 5 degrees Celsius.

Coreference relations are necessary for establishing coherence in discourse (Halliday and Hasan, 1976). In particular, referring expressions operate as discourse relational devices that form a layer of discourse structure, and therefore the successful resolution of referring expressions can be exploited for the derivation of a text's discourse structure. However, the variety of linguistic forms used to express the same entity poses challenges for the field of Computational Linguistics, since their resolution requires various linguistic resources as well as additional knowledge. Coreference resolution had first attracted major attention from the community in the 90s,

[2]The example was taken from the coreference corpus developed as part of this work and partly modified to better demonstrate the described issue.

when it became one of the main topics of the 7th Message Understanding Conference (MUC). So far, various approaches to coreference resolution have been developed, evolving from using rule-based methods (beginning with (e.g. Hobbs, 1978, Lappin and Leass, 1994), and more contemporary ones such as (e.g. Haghighi and Klein, 2009, Lee et al., 2011)) into exploiting neural networks (e.g. Wiseman et al., 2016, Clark and Manning, 2016, Lee et al., 2017). To date, most of the studies on coreference resolution focus on English and a few other languages. However, developing a coreference resolution system for a new language from scratch or adapting an existing one to a new language is challenging due to its technical complexity and the variability of coreference phenomena in different languages. Furthermore, it depends on high-quality language technologies (such as mention extraction, syntactic parsing, named entity recognition) as well as gold standard data, which are not available for a wide range of languages.

Coreference resolution plays an important role in Computational Linguistics, since it is usually a necessary step in higher-level natural language processing (NLP) applications. For example, the quality of anaphora and coreference resolution is the key factor in automatic text summarization: Without establishing coreferential links, it is impossible to summarize information with respect to the entities presented in the text. Furthermore, in information extraction (IE), it is important to recognize whether different pieces of information characterize the same entity. Therefore incorporating a coreference resolution module into information extraction systems is essential for its performance, as shown in e.g., (McCarthy and Lehnert, 1995). This can be illustrated by the following example:

(4) [Familymart Co. of [Seibu Saison Group]$_1$]$_2$ will open [a convenience store]$_3$ in [Taipei]$_4$ Friday in [a joint venture]$_5$ with [Taiwan's largest car dealer]$_6$, [the company]$_2$ said Wednesday.[3]

In this example, the underlined phrases contain relevant information that should be extracted by an IE system. As one can see from the example, these phrases also represent referring expressions that participate in coreference relations and therefore can be referred to by various linguistic devices further in the text. Therefore, in order to extract all the information about these entities, it is important to use coreference resolution as a preprocessing step to collect all the target expressions.

Correct pronoun resolution is important for Machine Translation (MT) as well: Mentions of the same entity can have different gender in different languages, therefore

[3]The example is taken from (McCarthy and Lehnert, 1995).

it is important to identify which entity a pronoun refers to in order to translate it correctly. For example, in English, *the sun* is neuter and could, therefore, be referred to with the personal pronoun *it*, while its German equivalent, *die Sonne*, is feminine and could only be substituted by a pronoun *sie* ('she'). Obviously, simply translating the pronouns without looking at their antecedents would result in erroneous translation, therefore taking into account coreference is important to ensure the translation quality. Interestingly, this problem attracted considerable attention in recent years – for instance, several DiscoMT shared tasks (Hardmeier et al., 2015, Loáiciga et al., 2017) explicitly focused on the translation of anaphoric pronouns across languages.

Coreference resolution also supports automatic text classification. In particular, coreference links help to identify: (a) mentions of the same concepts and entities expressed by different linguistic forms and (b) mentions of different concepts and entities expressed by similar linguistic forms. Typically, incorporating coreference information alleviates these decisions by changing the weight of the terms that occur in coreference chains, which results in a more precise performance of the classifiers (Li and Zhou, 2010, Mitkov et al., 2012).

Additionally, coreference resolution is an essential preprocessing step for Sentiment Analysis. If a specific target is investigated with respect to its sentiment, it is important to apply coreference resolution in order to retrieve all the mentions of the same target in a text, as shown in e.g., (Nicolov et al., 2008). For instance, in (5), by simply looking at the first sentence, it is impossible to infer sentiment for the target *Zune$_i$ 80*; it is the second sentence that contains linguistic markers that express positive sentiment. However, to make this determination, one requires coreference resolution in order to link *it* to its antecedent *the Zune$_i$ 80*, as well as bridging resolution to recognize that *the UI* ('user interface') and *the sound* also refer to the target.

(5) I can't stop playing with [my new Zune$_i$ 80]$_1$. [It]$_1$ is lovely to look at and hold, the UI is great, and the sound is super.[4]

However, common technologies for automatic coreference resolution require either a language-specific rule set or large collections of manually annotated data, which are typically limited to newswire texts in major languages (usually English). This makes it difficult to develop coreference resolvers for a large number of the so-called *low-resourced* languages and also to apply existing technologies to other languages.

[4]The example is taken from (Nicolov et al., 2008).

1.2 Research questions

As stated in the previous section, most of the coreference systems can only work on English data and are not ready to be adapted to other languages. Furthermore, large-scale gold standard datasets are required in order to train and test a new system; however, such datasets are difficult to produce, and therefore they are not available for a wide range of languages.

In this thesis, we propose that these challenges can be alleviated by using cross-lingual annotation projection which allows for automatically transferring existing methods or resources across languages in an aligned parallel corpus. Typically, annotation projection techniques rely upon sentence and word alignment information and are performed using a well-studied language in order to project annotations into a low-resourced language.

Therefore, the main goal of this thesis is to explore the difference in coreference phenomena in several languages with the application to annotation projection. Specifically, we will examine whether annotation projection can be applied to the task of coreference resolution in order to automatically generate new annotated datasets for different languages. For our study, we deliberately choose two language pairs: two relatively similar languages (English-German, both Germanic languages) and two less similar languages (English-Russian, a Germanic and a Slavic language).

In this thesis, we will focus on the following research questions:

1. How comparable are coreference relations – identity, near-identity and bridging – across different languages in a parallel corpus?

Nowadays, the availability of parallel corpora supports, to a high degree, the development of cross-lingual NLP technologies in various fields of Computational Linguistics. Most of the research has recently evolved around Machine Translation, such as pronoun translation. In this thesis, we concentrate on a less investigated application of parallel corpora: Our aim is to contrastively investigate coreference chains across languages in order to enable the creation of multilingual coreference corpora, which, in our opinion, can support the development of cross-lingual coreference resolution systems.

Thus, we are interested in investigating the commonalities and differences in coreference chains across English, German and Russian. In particular, we will examine linguistic devices that can serve as referring expressions in the three languages, and we will compare different types of coreference relations (identity, near-identity and bridging) across languages. To reach our goal, we will focus on developing common

annotation guidelines applicable to the three languages, and we will then build a first parallel coreference corpus, which will be subsequently used in our experiments.

2. How feasible is annotation projection as a method to alleviate manual annotation?

While monolingual coreference resolution systems are being constantly improved, multilingual coreference resolution has received much less attention in the NLP community. One of the reasons for that is the lack of multilingual resources annotated according to common annotation standards. This task is particularly challenging for coreference due to a large variety of coreference phenomena: Even for a single language, coreference guidelines are typically not standardized and may contain controversial decisions. However, applying annotation projection can support the creation of common coreference corpora for multiple language pairs.

Therefore, in this thesis, we concentrate on exploiting various annotation projection approaches to alleviate the task of manual annotation. From a historical perspective, annotation projection was first applied in the pioneering work of Yarowsky et al. (2001), who used the method to induce part-of-speech (POS) taggers, chunkers and morphological analyzers in several languages using parallel corpora. Subsequently, there has been influential work on annotation projection for different NLP tasks which performed quite well cross-lingually, e.g., for syntactic parsing (Rasooli and Collins, 2015) or semantic role labeling (Akbik et al., 2015). In our study, we will apply annotation projection to transfer coreference chains across the three languages. In particular, we will experiment with using different language pairs and different amounts of linguistic information to enable a more effective multilingual resource transfer. Also, we will conduct a manual experiment on projecting bridging pairs.

3. Can exploiting two different languages enhance the performance of annotation projection?

Finally, we will exploit the possibility of using several languages to improve the projection method. As opposed to annotation projection from only one source language, multi-source annotation projection makes use of several sources in order to obtain more reliable target annotations, as shown in e.g., (Rasooli and Collins, 2015, Agić et al., 2015). Typically, such strategies as majority voting are then used to select the most accurate incoming annotation.

With regards to coreference, the main idea of this part of the study is that multi-source annotation projection for coreference resolution would grant a bigger pool of potential mentions to choose from, which can be beneficial for overcoming language divergences. Therefore, the main goals of this part of the study are: (a) to explore

different strategies of multi-source projection of coreference chains, and (b) to evaluate the projection errors and assess the prospects of this approach for multilingual coreference resolution.

1.3 Outline

This section presents an overview of the structure of this work as well as the conventions regarding the provenance and the mark-up of linguistic examples presented in this thesis.

Structure of the thesis

The thesis is organized as follows:

Chapter 2 gives a linguistic introduction to the phenomenon of anaphora and coreference.

Chapter 3, provides an overview of the related work, including (a) annotation of coreference relations in multiple languages and (b) overview of annotation projection technologies.

Chapter 4 introduces the first parallel coreference corpus, which was created for the cross-lingual study of coreference phenomena in the three languages and for running experiments on annotation projection. In this chapter, we describe the process of building the corpus, the development of trilingual annotation guidelines used to annotate the corpus, and report on the results of the inter-annotator agreement. Furthermore, we study the correlation between identity coreference and bridging in our corpus and present the results.

In **Chapter 5**, we focus on the corpus alignment that is required for the subsequent annotation projection experiments. In particular, we will discuss sentence and word alignment, the quality of which is important for the success of the projection algorithm.

In **Chapters 6-8**, we report on our three experiments with projecting coreference chains in different settings. All the experiments are based on the direct projection method, but differ in the types of alignments and projection strategies used. Furthermore, each of the experiments is performed in two settings: a knowledge-lean setting (relying only on automatic word alignments) and a more linguistically informed one (using the output of a syntactic parser to identify mention borders and therefore improve the quality of the mention identification).

Thus, in Chapter 6, we describe a study of projecting coreference chains using only one source and a 'classical' projection direction (from English to other languages) as well intersective word alignments to maximize the projection quality. We compare how well our projection method works for two relatively similar languages (English-German) and less similar languages (English-Russian), and we also study the differences incurred by the text genre.

Chapter 7 presents a novel approach of multi-source annotation projection using all the alignments and additional projection directions, such as German-Russian and Russian-German. The novelty of this experiment is that we implement several multi-source projection strategies based on the concatenation and intersection of the projected mentions and compare them to each other.

In Chapter 8, we adopt a fully automatic pipeline and use automatic source annotations produced by two state-of-the-art coreference systems. We combine the output of our projection method for two source languages (English and German) to obtain target annotations for a third language (Russian), and we compare these results to the projection of manual coreference annotations.

Chapter 9 reports on a pilot experiment of manually transferring bridging pairs from German into English and Russian. In this chapter, we carefully describe our procedure and present the results of this experiment coupled with an analysis of the projection errors.

Chapter 10 summarizes the findings and the contributions of this thesis, and provides insights into the possible avenues for future work.

Conventions

In the following, we use several conventions regarding the provenance of linguistic examples provided in this work:

- Linguistic examples that come from our corpus are marked accordingly (in a footnote). If an example has been slightly modified to better suit the purposes of this work, it is correspondingly acknowledged.

- Linguistic examples coming from other data sources are given with a footnote reference to that source.

- If no data source for an example is specified, then it has been artificially composed for the purposes of this thesis.

Furthermore, we use several mark-up conventions (unless explicitly stated differently) to demonstrate annotated coreference, bridging and near-identity relations. In particular, we will be referring to the annotated referring expressions as *markables*, using the following conventions to highlight them in the text:

- We use square brackets and indices to mark coreference chains and their IDs (e.g., *[the door]*$_1$).
- We use square brackets and the letter B with an index to mark bridging pairs (e.g., *[the door]*$_{B1}$).
- We use square brackets and the letters NI with an index to mark near-identity pairs (e.g., *[the door]*$_{NI1}$).

Previously published material

Since some parts of this thesis have already been published, in the beginning of each chapter, we will specify whether it contains previously published material.

Chapter 2

Linguistic background

In this chapter, we introduce coreference phenomena from the linguistic perspective (Section 2.1) and present an overview of direct (Section 2.2) and indirect (Section 2.3) coreference relations. Furthermore, we discuss other complex cases of coreference, which can pose additional difficulties for coreference resolution (Section 2.4).

2.1 Anaphoricity and coreference

Reference resolution is the task of identifying linguistic expressions that refer to a certain entity in the real world – or *referring expressions* – and establishing the relations between them and discourse entities – or *discourse referents* – they refer to. As Stede (2011) notices, these entities do not necessarily have to exist in the real world, but need to be present in the mental models of writer and reader, because we can have common conceptions of entities that do not have to actually be there at the current point in time. When two referring expressions point to the same discourse referent, they are called *coreferent*, and a set of all expressions referring to the same entity forms a *coreference chain* for that entity. For the illustration, let us consider the following example:

(6) [Daisy Hamilton]$_1$ was a private detective. [She]$_1$ was thirty years old and ∅ has been a detective for the past two years. Every morning [Daisy]$_1$ went to [[her]$_1$ office]$_2$ to wait for phone calls or open [the door]$_3$ to clients needing [her]$_1$ services. One day somebody knocked on [the door of [the office]$_2$]$_3$..[1]

[1]The example is taken from the coreference corpus developed as part of this work.

The example above contains three coreference chains: The first chain points to the entity 'Daisy Hamilton', the second one – to 'the office', and the third one represents 'the door'. Already in the first chain, we see a wide range of linguistic devices used to express coreference relations: a proper name (*Daisy Hamilton*), different types of pronouns (personal – *she*, possessive – *her*) and also an ellipsis (or zero anaphora, see Section 2.2 for details), where a personal pronoun is omitted in the second sentence. In addition, in this example, we encounter two cases of referring expressions that are embedded in one another: *her office* and *her*, and *the door of the office* and *the office*.

Text coherence is maintained only when all the references and their discourse entities can be easily identified by the reader. From this viewpoint, we can speak about anaphora resolution and coreference resolution[2]. On one hand, *an anaphor* is a referring expression that cannot be interpreted without considering the discourse context; in particular, one needs to find another referring expression (an *antecedent*) that makes the intended discourse referent clear (Stede, 2011). Anaphora resolution is then the task of identifying an antecedent for each anaphor in a natural language discourse. To exemplify, looking at the text fragment presented in (6), one can see that the referent of the personal pronoun *she* can only be identified by resolving its antecedent – *Daisy*. On the other hand, coreference resolution is usually defined as the task of grouping all the mentions of discourse referents in a text into classes, coreference chains, corresponding to those referents (Stede, 2011).

Coreference and anaphora are two independent phenomena that do not necessarily have to occur simultaneously. For example, in (7), *the door* is definite and anaphoric but not coreferential, although the two entities are related. As already briefly mentioned in Section 1.1, these are cases of indirect coreference (or bridging, see Section 2.3 for details). Conversely, in (8), two mentions of *London* are coreferential, but they are not anaphoric since the interpretation of the second mention does not depend on the first mention.

(7) The lady entered [the office]$_{B1}$ and closed [the door]$_{B1}$ behind her.

(8) Almost six months ago, G-20 leaders met for a historic summit in [London]$_1$. ... Many of the problems that spurred the summit in [London]$_1$ remain real.

Anaphoric and coreference relations can be expressed by various linguistic devices. For example, for English, some of the most frequently used types of referring expressions also listed in the work of Ariel (1988) are the following:

[2]In this thesis, we will be addressing coreference relations in general, assuming that they include anaphoric relations.

a. Definite noun phrases (NPs): 'the lady'

b. Proper names: 'Daisy Hamilton'

c. Demonstratives: 'there / this dog'

d. Pronouns: 'he / his'

In the following section, we will present a more detailed overview of coreference phenomena, classify them in different dimensions as well as discuss other more complex issues that are also classified as or are related to the coreference phenomena.

2.2 Overview of direct coreference phenomena

In general, coreference phenomena can be categorized in respect to the type of the referring expression, the place of the referring expression in the text, the distance between referring expressions in a coreference chain, etc. In this section, we are interested in describing the typology of direct coreference phenomena from different perspectives, focusing on those types of anaphora, which will subsequently play an important role in the annotation projection experiments. Moreover, we will cover other types of anaphora, which are not annotated in the scope of this work, but that are important to keep in mind when comparing coreference relations across languages and different genres of texts.

First, we present a typology of coreference phenomena according to the **type of the anaphor**, which can be represented by one of the following units:

a. Pronominal anaphora: personal, possessive, reflexive, demonstrative and relative pronouns, when used anaphorically.

(9) [Sue]$_1$ forgot [her]$_1$ umbrella on the bus.

(10) [I]$_1$ made [myself]$_1$ a ginger tea.

(11) All accredited journalists will have access to [the restaurant]$_1$, [which]$_1$ is located on the eighth floor.

The non-anaphoric cases include the pleonastic use of the third person pronoun *it* (such as *It rains.*), which should be distinguished from the anaphoric use.

b. Lexical noun phrase anaphora: definite noun phrases (also called *definite descriptions*), proper names.

(12) I am afraid [Lorna]$_1$ has been kidnapped. ... I don't think my husband is interested in whether [Lorna]$_1$ has been kidnapped or not.[3]

c. Noun anaphora or 'one-anaphora': an indefinite pronoun *one*

(13) I don't think I'll buy [that car]$_1$, I need a cheaper [one]$_1$.

d. Verb anaphora: verbs or verb phrases.

(14) Peter [ordered a wine]$_1$; so [did]$_1$ Sarah.

e. Adverb anaphora: locative (such as *there*) and temporal (such as *then*) adverbs.

(15) She said she had met me at [Harvard]$_1$, but I have never been [there]$_1$.

f. Zero anaphora: anaphors – nouns, verbs and verb phrases – that are not overtly represented by a word or phrase, but can still be recovered in the context.

(16) [I]$_1$ got off at Copenhagen Central and ∅ headed towards the conference venue.

(17) Sam bought an apple and Mary ∅ a cake.

In these examples, the pronoun *I* as well as the verb *bought* are highly salient and therefore can be omitted without any change in the meaning of the sentence.

In our study, we are specifically interested in the cases of nominal anaphora ((a) - (c)) as well as the cases of adverbial anaphora (e); cases (d) and (f) are beyond the scope of this study. Still, it should be noted that other types of expressions can also signal anaphoric relations, and that the realization of certain types of coreference relations can differ across languages.

Furthermore, coreference phenomena differ according to the **type of the antecedent**, as shown in (Mitkov, 2002). Similar to the previous classification, the following discourse units can be antecedents:

(a) Noun phrases, such as pronouns, nouns, definite descriptions.

(b) Verbs, in the case of verb anaphora.

[3]The example is taken from the coreference corpus developed as part of this work.

(c) Bigger discourse units, such as clauses, sentences, sequences of sentences. It should be noted that this type of coreference relation is also called *abstract anaphora*, since the referents are typically abstract facts or events (Kolhatkar et al., 2018). For example:

(18) [They will probably win the match]$_1$. [That]$_1$ will please my mother.[4]

(d) Coordinated antecedents: two or more noun phrases that serve as a single antecedent to a plural noun phrase or a pronoun. For example:

(19) The cliff rose high above [Paul]$_1$ and [Clara]$_2$ on their right hand. [They]$_{1+2}$ stood against the tree in the watery silence.

Since our focus lies on nominal coreference, as possible antecedents, we only consider noun phrases ((a), (d)) and do not take into account verbs or any bigger discourse units.

Another important aspect in the classification of coreference phenomena is the **location of the anaphor and antecedent** (Mitkov, 2002). In particular, the antecedent can be located in the same sentence as the anaphor (intrasentential anaphora) or in a different sentence (intersentential anaphora). The most typical examples of intrasentential anaphors are reflexive and possessive pronouns, which are usually located in the same sentence or even in the same clause as their antecedents (Mitkov, 2002). In contrast, personal pronouns and noun phrases are frequently used intersententially.

Furthermore, coreference phenomena differ in the order the anaphor and antecedent appear in the text. While, typically, the antecedent precedes its anaphor, there are also cases when a reference is made to an entity that is mentioned further in the text. Such cases of anaphora are called **cataphora** and can be illustrated by the following example:

(20) After closing [her]$_1$ bag, [the woman]$_1$ put it on the desk.

Another important distinction is between **identity-of-reference** and **identity-of-sense** coreference: While the former type of identity encompasses cases when the anaphor and the antecedent have the same referent in the real-world (example (21)) and is more frequent, the latter deals with the cases when the anaphor and antecedent do not denote the same referent, but a referent with a similar description, which is also called 'one-anaphora' (example (22), see also example (13)).

[4]Examples (18)-(19) are taken from the work of Mitkov (2002).

(21) [Mary]$_1$ was not feeling well, that is why [she]$_1$ stayed home.

(22) [Two new optional protocols]$_1$ have entered into force in the past five years and [another one]$_1$ is under consideration.[5]

In automatic coreference resolution, this case has to be distinguished from at least two other usages of *one*, such as the 'generic person' (*As one can see.*) or numerical (*one apple*), as pointed out by Stede (2011). Furthermore, as we can see from the example above, the anaphoric pronoun does not necessarily have to share the same morphological features with its antecedent (singular vs. plural number).

Finally, another type of anaphora is the so-called *bound anaphora*. Bound anaphora occurs when the interpretation of an anaphor depends on syntactic constraints, and the anaphors are therefore 'bound' to their antecedents, as compared to other cases of coreference that are interpreted based on discourse pragmatics. Reinhart (1983) provides the following examples:

- Obligatory coreference:

  (23) [Zelda]$_1$ bores [herself]$_1$.[6]

- Obligatory non-coreference:

  (24) [Zelda]$_1$ bores [her]$_2$.

  (25) [She]$_1$ adores [Zelda's]$_2$ teachers.

- Optional coreference:

  (26) [Zelda]$_1$ adores [her]$_{1/2}$ teachers.

  (27) Those who know [her]$_{1/2}$ adore [Zelda]$_1$.

Thus, in the first case, our interpretation of the pronoun *herself* is determined solely by syntactic constraints (and there is only one interpretation possible), while in the third case the interpretation of the pronoun *her* is ambiguous and depends on the previous discourse.

[5]The example is taken from the United Nations (UN) Parallel Corpus (Ziemski et al., 2016).

[6]Examples (23)-(27) are taken from the work of Reinhart (1983).

2.3 Overview of indirect coreference phenomena

In this section, we will focus on indirect anaphora, i.e., when a definite NP refers to some aspect of a previously mentioned entity and is therefore linked to it by a relation other than identity. This phenomenon is also called *bridging* and was first introduced in the work of Clark (1975): A referring expression is definite because it builds a bridge to some previously mentioned entity, without being identical to it. In the following, we will be using the term *(bridging) anaphor* for the former and *(bridging) antecedent* for the latter, and we will be referring to such anaphor-antecedent pairs as *bridging pairs*.

The concept of bridging defined by Clark (1975) became the basis for most of the works that focus on studying bridging relations. In his study, Clark examines the inferences that the listener can make based on the speaker's message (which he calls *bridging from previous knowledge*), making a distinction between direct reference and indirect reference. Specifically, direct reference is what we usually understand by identity coreference, when two NPs share the same referent in the real world.[7] Clark (1975) names three classes[8] of indirect reference, illustrating them with the following examples:

1. **Indirect reference by association**: The antecedent is closely associated with its anaphor and is a part of it (*room - the ceiling*).

2. **Indirect reference by characterization:** The anaphor characterizes a role that something plays in an event or circumstance mentioned before (*murder - the murderer*).

3. **Reasons, causes, consequences, and concurrences**: The antecedent is often an event and not an object, and then it gives reasons for, causes of, consequences to, or concurrences of previously mentioned events or states (*John fell. What he wanted to do was to scare Mary.*[9]).

For each class, Clark makes a distinction between necessary and optional parts and roles which can be explained in the following:

[7]It is worth pointing out that reference to one or more members of a set to the whole set is also seen by Clark as direct reference.

[8]It is worth noticing that only the first two classes deal with nominal coreference.

[9]The example is taken from the work of Clark (1975).

(28) (a) During [the terrorist attack in Mumbai]$_{B1}$ [the attackers]$_{B1}$ did not hide their faces.[10]

(b) Clare walked into [the office]$_{B1}$ and saw a bunch of flowers on [the windowsill]$_{B1}$.

The difference between the two examples is that in (28)(a) *the attackers* is an absolutely necessary role of the mentioned event, while from (28)(b) we can infer that *the office* has one *windowsill* (which is not necessarily true for all the offices). Necessary and optional components of entities or events vary in their predictability by the listener from absolutely necessary to quite unnecessary (Clark lists three levels of 'necessity' on this continuum).

As we can see from the examples above, bridging anaphors are typically represented by definite descriptions that are unique in the natural language discourse. However, some early works propose to also interpret indefinite descriptions as being related to some previously introduced entities. For instance, Asher and Lascarides (1998) provide the following example:

(29) Jack was going to commit [suicide]$_{B1}$. He got [a rope]$_{B1}$.[11]

They observe that the proposition in the second sentence is attached to the proposition in the first sentence; in other words, the reader is able to build a bridge between *a rope* and *suicide*, although the former is represented by an indefinite noun phrase. However, the authors themselves admit that they mostly concentrate on cases with definite descriptions; moreover, this idea did not attract much attention in the follow-up work since it would pose additional difficulties for the annotators (see 3.1.2 for the details on the bridging annotation).

2.4 Other complex cases

In the previous sections, we presented an overview of direct and indirect coreference phenomena from different standpoints. However, there are several further complications that may arise when interpreting coreference relations and that may pose additional difficulties for the resolution. In the following, we list the most important cases that frequently occur in a natural language discourse.

Firstly, finding the correct antecedent for an anaphor is not always straightforward: Coreference relations can be ambiguous. In particular, there are cases in which

[10]Example (a) is taken from the corpus developed as part of this work

[11]The example is taken from the work of Asher and Lascarides (1998).

more than one antecedent is possible, which pose additional difficulties for both the annotation and the resolution of coreference. For instance:

(30) [Ann]$_1$ convinced [Mary]$_2$ [she]$_?$ was a good actress.

In this example, both *Ann* and *Mary* could serve as potential antecedents for the anaphoric pronouns *she*; the correct resolution of this case highly depends on the context and is not possible in the scope of a single sentence.

Second, plural pronouns can have multiple antecedents that are located at different positions in the text (the so-called *split antecedents*). For example:

(31) [John]$_1$ first met [Mary]$_2$ at school. [They]$_{1+2}$ have been good friends since that time.

Furthermore, anaphoric expressions can refer not only to specific referents in the real world but also have generic referents:

(32) I adore [dogs].

(33) Thomas Edison is usually credited with the invention of [the light bulb].

Another difficult case is deictic pronouns, mostly of first and second person, that can point to a referent outside of the scope of the text. Such pronouns are typical for spoken language and, in particular, for dialogue speech. For instance:

(34) Let [me] explain this theory to [you].

In this example, the referents for the anaphoric pronouns *you* and *me* cannot be found in the text, but only in the context of the utterance.

Overall, in this subsection, we presented an overview of various coreference phenomena, classifying them in several dimensions. In the rest of the work, we will mostly focus on nominal anaphora; however, the overview and the understanding of other existing types of coreference will support our further analysis and cross-lingual comparison of coreference in the corpus annotation and the projection experiments.

Chapter 3

Related work

This chapter provides an overview of the related work, including (a) annotation of coreference relations in multiple languages (Section 3.1) and (b) an overview of the annotation projection technologies (Section 3.2). We will first introduce the most influential annotation efforts in respect to direct and indirect coreference relations from both monolingual and multilingual perspectives and then turn to the recent advances in annotation projection applied to various NLP tasks. As for the details about annotation schemes relevant for this work, we will discuss these in the appropriate parts of the thesis.

Previously published material

Some parts of this review of the related work on bridging (in particular, Section 3.1.2) have been published as (Grishina, 2016). The literature review on annotation projection (Section 3.2) has also partly been published as (Grishina and Stede, 2015) and (Grishina and Stede, 2017).

3.1 Coreference annotation

In this section, we will focus on the annotation of coreference, considering different types of coreference phenomena and coreference relations as well as the standard metrics for the evaluation of the quality of coreference annotations. As for coreference relations, we will first examine direct coreference, and, thereafter, we will focus on indirect coreference relations (bridging). Finally, we will introduce near-identity coreference, which is an artefact of annotation that takes place if the two referring expressions are partially the same in that they share most of the important characteristics, but differ in one crucial dimension (such as time, for instance, *Rome – the ancient city of Rome*).

3.1.1 Direct coreference

In this subsection, we will focus on the annotation efforts covering direct coreference relations. In particular, we will investigate both monolingual and multilingual datasets, and we will give an overview of the most widely applied annotation schemes.

Monolingual datasets.

From the historical perspective, coreference became one of the central topics of a series of Message Understanding Conferences (MUC), as briefly mentioned in 1.1. Specifically, the most influential and publicly available early guidelines on coreference are the MUC-7 guidelines (Chinchor and Hirschman, 1997), developed in the framework of the 7th Message Understanding Conference. These guidelines served as a basis for many subsequent annotation efforts as well as for creating the datasets for system evaluation. Therefore, the main goal was to enable relatively quick and cheap annotation of texts with high inter-annotator agreement, which influenced several decisions regarding the annotation task.

Technically, MUC guidelines restricted the annotation of coreference to the identity relation and nominal coreference: nouns, noun phrases and pronouns (in particular, personal, possessive, and demonstrative pronouns), and also including dates, currents, and percentage figures. Verbal expressions were generally not considered as markables (e.g., *riding a bike*); the only exception is verbal phrases with modifiers (such as *reading of a novel, slow reading*). Regarding the spans of referring expressions, the authors proposed to annotate both minimal spans (syntactic heads of NPs) and full spans (NPs with all their modifiers) for all the markables. Interestingly, appositions can also participate in coreference relations and are considered coreferential with the corresponding noun phrases.

However, as pointed out by Stede (2011), MUC guidelines included several controversial cases, such as establishing coreference relations between currencies, percentages, etc. For instance, in (35), *the stock price* corefers with *$ 4.02* and then with *$ 3.85*, although both values are not identical.

(35) [The stock price]$_1$ fell from [$4.02]$_1$ to [$3.85]$_1$.[1]

Recently, there have also been a couple of solid coreference annotation efforts for mostly European languages that will be briefly discussed in this section. We will primarily focus on the corpora development for the languages relevant to this work, and also briefly review several other well-known projects.

To our knowledge, the largest current annotation effort for English is the OntoNotes corpus (Hovy et al., 2006) which comprises several levels of morphological, syntactic and semantic annotation, including coreference. The most recent release is the OntoNotes corpus Version 5.0[2], which consists of 3490 documents annotated with coreference for English and a smaller number of documents for Chinese and Arabic, and it is being widely used for system development and evaluation.

The OntoNotes annotation scheme represents annotation instructions for direct coreference relations, excluding appositive relations that are to be marked separately. Annotation of direct coreference in OntoNotes is mostly limited to the following linguistic expressions:

1) Noun phrases (pronominal, nominal, named entity mentions, including temporal expressions);

2) Proper noun premodifiers[3], such as *[FBI] spokesman - [FBI]*;

3) Verbs that co-refer with a noun phrase, such as *[grew] - [the growth].*

Furthermore, coreference links can only be established between the types of referring expressions described above that are mentions of specific referents. These types do not include generic, underspecified or abstract entities. Such mentions can only be linked to pronouns or other definite mentions if they serve as their antecedents (see example (36)) but not to other generic/underspecified/abstract entities (see example (37)). For instance:

[1]The example is taken from (Chinchor and Hirschman, 1997)

[2]https://catalog.ldc.upenn.edu/LDC2013T19 [accessed on 03.11.2017]

[3]Unless they are adjectival (such as **[American] economy*) or refer to a nationality (such as **[U.S.] economy*).

(36) [Meetings]$_1$ are most productive when [they]$_1$ are held in the morning. [Those meetings]$_1$, however, generally have the worst attendance.[4]

(37) Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for *[cataract surgery]. The lens' foldability enables it to be inserted in smaller incisions than are now possible for *[cataract surgery].

Importantly, spans for all the noun phrases are pre-selected for the annotation, while for verbs and premodifiers they have to be selected by the annotators. In case of nested NPs that share the same head, the NP with the largest span should be selected as a markable. Contrastively, for verbs, only the head of the verb phrase is included into the markable span.

As already mentioned earlier, appositive relations are marked separately. In particular, the annotators should establish appositive links between the referring expression itself (or *head*) as well as one or more of its *attributes*. According to the guidelines, an appositive construction contains a noun phrase that modifies an 'immediately-adjacent' noun phrase, for example:

(38) ... [[the PhacoFlex intraocular lens]$_x$ <HEAD>, [the first foldable silicone lens available for cataract surgery]$_x$<ATTRIBUTE>]

Thereafter, the entire appositive construction can be linked to its antecedent via identity relation.

The inter-annotator agreement scores for the OntoNotes corpus measured as MUC score[5] (Vilain et al., 1995) are quite reliable: The highest score is reported for broadcast conversations (86.7 MUC), while the lowest is for the magazine genre (78.4 MUC) (Pradhan et al., 2011). The scores reported after the adjudication of the annotations are considerably high: The highest score reaches 93.7 MUC (again, for broadcast conversations), and the lowest score is improved to 88.8 MUC. However, we are not aware of any evaluations on the complete dataset[6].

[4]Examples (36)-(38) are taken from the OntoNotes guidelines (BBN-Technologies, 2006).

[5]MUC score considers the minimum number of links between mentions to be inserted or deleted when mapping the system response to the gold standard set and ranges between 0 and 100 (see 3.1.4 for details on its computation).

[6]There is an average agreement score for direct coreference (91.8%) and for appositive constructions (94.2%) as reported earlier by Hovy et al. (2006). However, it is not clear from (Hovy et al., 2006) which metric was used to compute the scores.

A similar effort of annotating a large-scale corpus with multiple layers of linguistic annotation including coreference for German is TüBa-D/Z treebank of newspaper articles, which comprises around 104 000 sentences[7] and contains a coreference annotation layer (Naumann and Möller, 2006). In this corpus, several types of coreference relations were annotated. In particular, three types of relations were introduced to annotate direct coreference:

- *coreferential*, thus marking the relation between noun phrases;
- *anaphoric*, thus marking the relation between noun phrases and pronouns;
- *cataphoric*, thus marking cataphoric relations;
- *bound*, thus marking the relation between anaphoric expressions and quantified noun phrases (such as *niemand, jeder* etc.).

Furthermore, two more relation types – *part-of, instance* – were used to annotate bridging and will be discussed in the corresponding section (3.1.2). Additionally, the *expletive* relation was used to annotate expletive pronoun *es* (e.g., **[Es] wird getanzt.*) and distinguish it from the anaphoric ones.

As for potential markables in TüBa-D/Z, these encompass definite NPs, personal, relative, reflexive and reciprocal pronouns as well as possessive adjectives. These markables were already pre-selected in the corpus so that the annotators' task was only to establish corresponding links between anaphor-antecedent pairs (except for the bound relation, where only one instance was annotated) and choose the appropriate relation.

Another annotated corpus for German is Potsdam Commentary Corpus (PCC) (Stede and Neumann, 2014) that includes 175 short articles from a local newspaper. Among other syntactic and discourse annotation levels, it also contains coreference annotations. The annotation scheme used for this corpus covers nominal coreference and identity relation (Stede, 2016). In particular, it introduces the distinction between *primary* and *secondary* markables: Primary markables are definite referring expressions, that point to some specific discourse referent (such as *my dog*), while secondary markables are all other referring expressions, typically indefinite (such as *a dog*) (Chiarcos et al., 2016). Also, there is a category of *no markables*, which

[7] https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/corpora/tueba-dz.html [accessed on 26.02.2019]

includes non-referring nominal phrases, such as nominal phrases used in idiomatic constructions (for instance, **[auf die Nerven] gehen*[8]). Another special category is the so-called *group markables*, that can include several primary markables to build an antecedent for a plural nominal or pronominal phrase (for instance *[Deutschland]*$_1$ *[gegen Argentinien]*$_2$ *– [beide]*$_{1+2}$ *siegen mit Hilfe des Schiedsrichters*).

Furthermore, Chiarcos et al. (2016) introduced several principles for the annotation task that, among others, support the decisions of selecting the spans of the markables, choosing the antecedents or building the coreference chains. For instance, *the maximality principle* postulates that the annotators have to select the longest possible span of a markable. Additionally, the guidelines provide instructions on how to deal with discontinuous markables (*Ich hatte [einen Tee] getrunken, [den Anna mir geschenkt hatte].*) or in case of recursive embeddings, when one markable is included into the span of the other markable (*[nach Washington, an [dessen]*$_1$ *Topf sie hängen]*$_1$)

Another approach to the annotation of referring expressions is adopted from the information structural perspective. This approach was developed by Riester and Baumann (2017), who created an annotation scheme for German. In particular, their guidelines analyze information status at two levels: a referential (*r-level*) and a lexical level (*l-level*), since the givenness of a constituent has to be defined differently for referring and non-referring expressions. Thus, referring expressions are considered at the *r-level*, which includes the following information status categories (Riester and Baumann, 2017):

- *r-given-sit, r-environment:* Referents contained in text-external context (communicative situation), such as first and second person pronouns, expressions referring to a unique entity in the visual context, etc.;

- *r-given, r-given-displaced*: Referent mentioned in previous discourse context, such as pronominal reference, repetition of the same referent, abstract anaphora etc.;

- *r-cataphor, r-bridging*: Discourse-new entities that depend on other expressions in the discourse context[9];

[8]Here and in the next paragraph, the examples are taken from (Stede, 2016).

[9]The part of the guidelines that deals with annotating bridging will be discussed in 3.1.2.

- *r-bridging-contained, r-unused-unknown, r-unused-known:* Globally unique entities that are discourse-new and independent of the discourse context, such as *the swimming pool of the new townhall, the Pope*;

- *r-new:* Non-unique, discourse-new entities, such as *a friend*;

- *r-expletive, r-idiom:* Non-referring expressions, such as *it* in *It is snowing* or *the drawing board* in *go back to the drawing board (='to start all over').*

From this list, *r-given, r-given-sit* and *r-bridging* can have antecedents, but only *r-given* and *r-given-sit* are allowed to form coreference chains of more than two mentions. Furthermore, Riester and Baumann (2017) introduced two additional features – *generic* (example (39)) and *predicative* (example (40)), to mark the corresponding constructions:

(39) [A cat]$_{r-new+generic}$ is a mammal.

(40) I consider her [a genius]$_{r-new+predicative}$.[10]

As for the Romance languages, some of the notable efforts are the works on annotating coreference in Italian – the GNOME (Poesio, 2000) and Venex (Poesio et al., 2004a) corpora, annotated according to the MATE annotation guidelines (Poesio et al., 1999), (Poesio, 2004). The core scheme only contains identity coreference, while in the extended scheme several associative relations are considered, such as *subset* or *part*. Furthermore, it provides instructions for the annotation of references to the visual situation (deixis) which allows for annotating coreference in both spoken and written texts. This approach was continued in the development of the ARRAU corpus (Poesio and Artstein, 2008), which additionally contained annotation of ambiguous antecedents, and the LiveMemories corpus of social Web texts (Rodrıguez et al., 2010), which, in particular, focuses on incorporated clitics and zero pronouns in Italian.

For Spanish and Catalan, the largest annotation effort known to the author is the AnCora-CO corpus of nominal coreference (Recasens and Martí, 2010), the annotation of which drew upon the MATE annotation scheme. Moreover, the AnCora-CO guidelines distinguish between identity, deixis and predication links.

Coreference in Slavic languages was extensively annotated for Czech, Polish and Russian in the Prague Dependency Treebank (PDT) (Bejček et al., 2012), the Polish Coreference Corpus (Ogrodniczuk et al., 2013) and the Russian Coreference Corpus

[10]The examples are taken from (Riester and Baumann, 2017).

Markable type	OntoNotes (en)	TüBa-D/Z (de)	PCC (de)	PDT (cz)
Relative pronouns	✓	✓	×	✓
Adverbial pronouns	×	×	✓	✓
Adjectives	×	✓	×	✓
Verbs	✓	×	×	✓
Zero pronouns	×	×	×	✓

Table 3.1: Differences in annotation schemes: difficult cases

(Toldova et al., 2015), respectively. While the Polish Coreference Corpus includes only the annotations of nominal coreference as well as experimental annotations of near-identity (Ogrodniczuk et al., 2013), the Prague Dependency Treebank adopts a broad understanding of coreference: Not only pronouns and noun phrases can be marked as markables, but also coreference of adjectives, local and temporal adverbs, verbal and abstract nouns and anaphoric zeros (Zikánová et al., 2015). However, appositions, predications as well as verbal complements are not considered as markables. Since Czech is a language without articles, it is quite difficult to distinguish between definite and indefinite anaphoric expressions, therefore both specific and generic nominal groups can be annotated for coreference. Furthermore, in addition to identity coreference, bridging relations are also annotated.

The Russian Coreference Corpus consists of 88 texts of different genres. The annotation principles only focused on nominal coreference and were partially based on the work of Krasavina and Chiarcos (2007) in terms of distinguishing between the primary and the secondary markables (Toldova et al., 2015). In addition, markables of both minimal and maximal length were annotated.

In sum, as one can see from above, annotation schemes used for different coreference corpora vary considerably in terms of (a) the types of referring expressions to be annotated as markables and (b) types of relations to be annotated. Table 3.1 summarizes some of the differences across several annotation schemes for English, German and Czech[11], presenting the decisions on the types of linguistic expressions to be considered as markables. Specifically, tick and cross symbols show whether the corresponding type of referring expressions is present or absent respectively in

[11]To our knowledge, the Russian coreference guidelines were not publicly available at the point of writing this work.

the selected annotation schemes: For instance, relative pronouns are annotated as separate markables in OntoNotes, TüBa-D/Z and PDT, but not in PCC (where they are included into the markable span); adjectives are annotated only in TüBa-D/Z and PDT, and verbs can only be found in OntoNotes. Even considering only four schemes from the above, with two of them being for the same language, we observe discrepancies on all of the cases. Obviously, there is no complete agreement between any of them and for the languages in question, which inevitably poses difficulties for building coreference systems for multiple languages.

Multilingual datasets.

Multilingual coreference datasets that to a great extent shaped the development and training of multilingual systems were those released for the evaluation tasks. There have been two multilingual coreference tasks in recent years, which provided manually annotated corpora, – SemEval 2010[12] (Recasens et al., 2010b) and CoNLL 2012[13] (Pradhan et al., 2012). The main goal of these tasks was to assess the quality of coreference resolution for different languages; however, the participating systems did not have to work on all the languages, but at least one of them. These tasks used some of the datasets described above as well as some newly annotated texts. In the following, we consider the datasets and the annotation schemes applied to their development.

- **SemEval 2010 Task 1: Coreference Resolution in Multiple Languages.** This task aimed at the evaluation of coreference resolution systems for six languages and released corresponding datasets for each of them. However, this data came from different sources (some of them already discussed above) and was comparable only to a certain extent. The English part of the dataset was taken from the OntoNotes corpus 2.0 (Pradhan et al., 2007) which comprises newswire and broadcast news texts annotated according to the OntoNotes annotation scheme. The German part of the dataset came from the Tüba-D/Z treebank (Hinrichs et al., 2004). The Italian collection was extracted from the LiveMemories corpus (Rodrıguez et al., 2010) which contained texts coming from Wikipedia, blogs, newswire and dialogues. Catalan and Spanish texts were from AnCora-CO corpus (Recasens and Martí, 2010), and the Dutch part was acquired from the KNACK corpus (Hoste and De Pauw, 2006), both built on newswire texts.

[12]http://stel.ub.edu/semeval2010-coref/ [accessed on 03.11.2017]

[13]http://conll.cemantix.org/2012/introduction.html [accessed on 03.11.2017]

T	Scheme	Training			Development			Test		
		docs	sents	tokens	docs	sents	tokens	docs	sents	tokens
I	*ca* AnCora	829	8 709	253 513	142	1 445	42 072	167	1 698	49 260
	nl KNACK	145	2 544	46 894	23	496	9 165	72	2 410	48 007
	en OntoNotes	229	3 648	79 060	39	741	17 044	85	1 141	24 206
	de Tüba-D/Z	900	19 233	331 614	199	4 129	73 145	136	2 736	50 287
	it LiveMemories	80	2 951	81 400	17	551	16 904	46	1 494	41 586
	es AnCora	875	9 022	284 179	140	1 419	44 460	168	1 705	51 040
II	*ar* OntoNotes	359	7 422	242 702	44	950	28 327	44	1 003	28 371
	en OntoNotes	2 802	75 187	1 299 312	343	9 603	163 104	348	9 479	169 579
	zh OntoNotes	1 810	36 487	756 063	252	6 083	110 034	218	4 472	92 308

Table 3.2: Datasets and annotation schemes used for the Shared Tasks: SemEval 2010 (I) and CoNLL 2012 (II)

- **CoNLL 2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes.** As the name of the task suggests, the multilingual data for the task was taken from the OntoNotes corpus (Version 5.0) (Hovy et al., 2006). It contained datasets for three languages – English, Chinese and Arabic (a small portion as compared to the other two). This time, the annotation was performed according to similar annotation standards, however, the languages used were very different which posed certain difficulties for the participating systems.

To our knowledge, the only parallel (bilingual) noun phrase coreference annotation effort is ParCor (Guillou et al., 2014), an English-German corpus with annotated anaphoric relations – which, however, is limited to the annotation of pronouns and their antecedents, without taking into account coreference chains. The ParCor corpus consisted of texts of two genres – prepared speech (TED Talks) and publications of the EU institutions (EU Bookshop) – and contained automatically identified markables prior to the annotation. For the annotation, specific guidelines were developed, which considered two types of referring expressions: pronouns (personal, possessive, demonstrative, relative, reflexive, pronominal adverbs, generic) and noun phrases, excluding pleonastic pronouns and event reference. Although these guidelines only focused on anaphoric links, they contained several interesting decisions that we describe below. For instance, due to the peculiarities of the genres, these guidelines contain instructions on how to annotate speaker and addressee reference pronouns (typically second

and third person pronouns) as well as provide some specific decisions for a certain text genre (e.g., reflexive pronouns in the EU Bookshop texts should not be marked as opposed to TED Talks). Furthermore, they allow for annotating anaphoric pronouns with no specific antecedents (and marking them correspondingly), such as in the following example:

(41) There is a study called the streaming trials. [They] took 100 people and split them into two groups.[14]

Also, according to ParCor, if a plural pronoun refers to a collective singular noun, it should also be annotated and linked to it (e.g., *[the government] – [they]*). Moreover, nominal premodifiers are considered as potential markables that can serve as antecedents for anaphoric pronouns (*[EU] supporters – [it]*).

In sum, it can be easily noticed that multilingual tasks do not share any common annotation standards. The data comes from different sources which makes it difficult to compare the annotations, unify datasets, and also develop systems applicable to these different datasets. For both linguistic comparison and system development, it would be better to have standardized annotations in different languages in terms of both types of referring expressions and types of coreference relations.

3.1.2 Bridging

In general, the recent approaches to the annotation of bridging derive from two different annotation frameworks. Firstly, bridging can be annotated as a part of the information structure of texts, along with other information status categories. Secondly, bridging can be seen as a separate category of textual coreference, besides identity and near-identity coreference. In the following, we consider both approaches.

Bridging at the information structural level.

In the theory of information structure, several types of information status categories are usually distinguished: For instance, *new* usually describes information not familiar to the reader and that appears in a text for the first time, while *given* is typically information already known through previous discourse. Moreover, there is a category of *accessible/mediated/inferable* information, which is information that was not previously mentioned in a discourse, but is still somehow familiar to the reader, who is able to draw certain implicatures based on his/her common knowledge or the knowledge obtained earlier. Usually, bridging falls into the latter category and can be

[14]The example is taken from (Guillou et al., 2014).

annotated together with other types of mediated information. Importantly, for such annotation schemes, the inter-annotator agreement results are reported on the entire scheme and are somewhat lower for the single categories.

Therefore, bridging can be annotated as an individual subcategory among other categories of information status, such as in the work of Nissim et al. (2004), who annotated bridging relations as a part of the *mediated* information structural category. Inside this category, Nissim et al. (2004) distinguished between several types of relations, including typical bridging relations such as subsets and physical parts. The validation of the scheme showed that the authors were able to achieve an average κ of 0.845 for the annotation of the top-level categories and a κ of 0.788 for the annotation of the subcategories, with *mediated* being on the second place, with a κ of 0.8 (Calhoun et al., 2005).

Subsequently, this approach to annotating bridging together with information status was enhanced and applied by Ritz et al. (2008), Markert et al. (2012) and some others. However, all these approaches treat the bridging category as a whole, not making any distinctions between individual subcategories. To our knowledge, the highest agreement on bridging anaphor recognition in particular (κ = 0.6-0.7) was reported by Markert et al. (2012), who, similarly to Nissim et al. (2004), annotated bridging relations inside the *mediated* category of information status. However, it should be noted that their interpretation of bridging is to some extent different from the others. Specifically, they do not restrict the annotation scope to definite noun phrases, allowing indefinite NPs to participate in bridging relations as well (similarly to the theoretical work of Asher and Lascarides (1998), see Chapter 2.3 for details). For instance:

(42) Still [employees]$_{B1}$ do occasionally try to smuggle out a gem or two. [One man]$_{B1}$ wrapped several diamonds in the knot of his tie.[15]

One of the most influential annotation schemes which provides guidelines on the annotation of referential information status including bridging is the scheme of Riester and Baumann (2017). As already mentioned earlier (see 3.1.1), referring expressions are considered at the *r-level*, which among others includes such information status categories as *r-given, r-new, r-bridging* etc. From this perspective, the main difference between the bridging category and other information status categories is, according to the authors, that bridging entities do not have a coreferential antecedent, but can be understood from the discourse and depend on other expressions in the previous

[15]The example is taken from (Hou et al., 2013)

context. In other words, bridging anaphors are unique in the context and necessarily have an 'anchor' they are related to.

In a similar way to the previous work, Riester and Baumann (2017) explicitly state that they only consider entities that are unique in discourse (i.e., nominal phrases with a definite article in German) to be bridging anaphors (example (43)), since allowing indefinite expressions to participate in bridging relations introduces a considerable degree of uncertainty to the annotation scheme: One has to examine each indefinite expression in order to classify it as a bridging anaphor or not (therefore, there is no bridging relation marked between *a bird* and *a feather* in example (44)).

(43) [The referee]$_{B1}$ lost control over the [football match]$_{B1}$.[16]

(44) *A bird* is sitting in the tree. It has just lost *a feather*.

Furthermore, Riester and Baumann (2017) introduce the so-called *bridging-contained* category that applies to bridging anaphors (e.g., *the walls of the old dormitory*) that are anchored to an embedded phrase (e.g., *the old dormitory*):

(45) [The walls of [the old dormitory]$_{B1}$]$_{B1}$ were very thick.

Importantly, these cases need to be distinguished from other entities that are discourse-new and independent of the context (in particular, the categories of *unused-known* and *unused-unknown*) as the authors themselves acknowledge. To exemplify, the following example will not be considered as bridging, although it is very similar to (45):

(46) [The garden of [the old dormitory]] was very crowded in summer.

The main difference between (45) and (46) is that in the first example, the relation between the noun phrases in question is prototypical (i.e., every building has walls) as opposed to the second one. To differentiate the first case from the second, Riester and Baumann (2017) propose to apply a permutation test dislocating the embedded phrase of a complex definite description to the left and testing whether the potential anaphor is still interpretable. In this manner, we can ensure that example (45) contains a bridging anaphor (e.g., *The old dormitory is comfortable and the walls are very thick.*). Conversely, example (46) loses its interpretability (**The old dormitory is comfortable and the garden is very crowded in the summer*).

Bridging at the coreference level.

[16]The examples (43)-(44) are taken from (Riester and Baumann, 2017).

One of the earliest approaches to the classification of bridging references based on corpus studies was introduced in the work of Vieira and Teufel (1997), who looked at the type of knowledge required to resolve bridging. In particular, they listed the following classes of bridging references:

1. **Synonymy, hyponymy, meronymy:** bridging by lexico-semantic relations, e.g., *the house - the wall.*

2. **Names:** bridging from an anaphoric definite description to a proper name, e.g., *Microsoft - the company.*

3. **Events:** bridging from an anaphoric definite description to a referent introduced by a verb phrase, e.g., *riding a bike - the tour.*

4. **Compound nouns:** bridging from a definite description to one of its modifiers, e.g., *the car accident - the car;*

5. **Discourse topic:** bridging to an implicit discourse topic that was not explicitly introduced, e.g., *the half time* with the topic being *a football match.*

6. **Inference:** bridging by a cause, reason, consequence or set-membership relation, such as *the royal family - the queen.*

As one can see from the list below, Vieira and Teufel (1997) included several anaphoric cases into their classification (such as *Synonymy* or *Names*). Furthermore, many of their classes overlap with those introduced by Clark (1975) (see 2.3 for details), in particular, the *Inference* class is similar to the last class listed in Clark's hierarchy.

In the annotation of bridging at the coreference level, recent related literature distinguishes between the following most common types of bridging relations: part-whole, set-membership and generalized possession (Poesio et al., 2004b, Poesio and Artstein, 2008, Hinrichs et al., 2005). However, these relations seem to be underspecified in the sense that part-whole is a very general relation; in contrast, we are interested in a more fine-grained classification of relations that could emerge from part-whole in order to build a more nuanced typology of bridging relations as well as study the differences across languages.

Another corpus that contains bridging annotations is the Prague Dependency Treebank. Since the definiteness in Czech is hard to determine due to the lack of articles, Zikánová et al. (2015) proposed to annotate several specific categories of bridging relations:

- *part-whole, whole-part:* meronymical relation between a part and a whole, such as *hand - finger*;
- *sub-set, set-sub:* the relation between a set and its subsets or elements, such as *drinks - lemonade*;
- *p-funct, funct-p:* the relation between an entity and a singular function on this entity, such as *trainer - team*;
- *contrast:* the relation between coherence-relevant discourse opposites, such as *the fortunes of the Czech Republic - the fortunes of Slovakia*;
- *anaph:* non-coreferential explicit anaphoric relation, such as *the disintegration of the Warsaw pact - that time*;
- *rest:* further underspecified bridging relations, such as location-resident (*Berlin-Berliner*) or event-argument (*research-researcher*).

Since the Prague Dependency Treebank also contains a tectogrammatical layer, bridging relations that are already captured by the tectogrammatical structure are not annotated. For that reason, bridging relations of genitive constructions within a single clause are not annotated, such as *resident* and *village* in *the resident of the village*. However, they have to be annotated if there is no direct syntactic dependency between the nodes (e.g., if they appear in different clauses).

A more complex and detailed annotation scheme for bridging in French was introduced by Gardent et al. (2003), who distinguished between five classes of bridging relations: set-membership, thematic (links an event to an individual via a thematic relation defined by the thematic grid of the event, e.g., *murder - the murderer*), definitional (relation is given by the dictionary definition of either the target or the anchor, e.g., *convalescence - the operation*), co-participants, and non-lexical (relation could be established due to discourse structure or world knowledge). It should be noted that, in this work, only definite descriptions were considered as bridging markables. Furthermore, they introduced several constraints on the basic ontological types that could participate in the relations described (e.g., for set-membership, the anaphor should be an individual and the antecedent should be a set of individuals), in order to make their scheme easily processable automatically.

Another taxonomy of bridging relations used for developing a rule-based system to resolve bridging was introduced by Hou et al. (2014), who used 8 rules to resolve

different types of bridging references. These rules were mainly based on related literature and their document set, which comprises 10 documents from the ISNotes Corpus[17], containing the Wall Street Journal portion of the OntoNotes corpus (Hovy et al., 2006). In particular, Hou et al. (2014) focused on the following types of bridging relations: building - part (*room - the roof*), relative - person (*the husband - she*), geopolitical entity - job title (*Japan – officials*), role - organization (*professor - organization*), percentage NP (*22% of the firms - 17%*), set - member (*reds and yellows - some of them*). Furthermore, they looked at the so-called *argument-taking NPs*, which support the recovery of certain kinds of bridging relations in the corpus: argument taking NP I (different instances of the same predicate in a document likely maintain the same argument fillers; *Marina residents - some residents*) and argument taking NP II (an argument-taking NP in the subject position is a good indicator for bridging anaphora, *Poland's first conference - the participants*).

As for the inter-annotator agreement at the coreference level, bridging was also shown to be a very complex category that poses difficulties for the annotators. The annotation of bridging relations typically includes the following subtasks: (a) recognizing bridging anaphors and selecting their antecedents, and (b) assigning appropriate bridging types. In general, inter-annotator agreement for (a) tends to be lower than for standard identity coreference, as shown in the works of Poesio and Vieira (1998), Poesio (2004), Nedoluzhko et al. (2009). In particular, Poesio (2004) reports that 22% of bridging references were marked in the same way by both annotators and 73.17% of relations are marked by only one or the other annotator. In the study of Nedoluzhko et al. (2009), the reported agreement varies from 42.0 to 59.0 F1 in different annotation rounds.

As for the types of relations, not much was reported lately. To our knowledge, only Nedoluzhko et al. (2009) reported on the scores for four basic relation types, with κ varying from 0.79 to 1 (given that the annotators already agreed on the recognition of bridging pairs). As the most frequent sources of errors, the authors list ambiguous context and therefore different interpretation of relations by the annotators, as well as the distinction between bridging and textual coherence (which was also annotated in the same texts). However, we are not aware of any other agreement studies for more complex relation sets.

In sum, corpus creation approaches to bridging classification are quite coarse-grained, while applied work (bridging resolution) tends to be very domain-specific.

[17]https://www.h-its.org/en/research/nlp/isnotes-corpus/ [accessed on 03.11.2017]

Both paths are rather problematic if we want to create reliable multi-genre annotated resources with a fine-grained classification of bridging relations.

3.1.3 Near-identity

The concept of near-identity is relatively new and was first introduced and subsequently discussed in the works of Recasens et al. (2010a) and Recasens et al. (2012). In these works, near-identity is defined as a middle-ground between identity and bridging, and it emerged out of the inter-annotator disagreements while annotating identity coreference. Near-identity holds between two NPs whose referents are almost identical, but differ in one crucial dimension. This can be, for example, the change of an object through the time or the reference to different roles of the same person.

Recasens et al. (2010a) distinguish between the following types of near-identity relations, providing the following examples (we also mention the subtypes for each of the types):

1. **Name metonymy:** The same entity is referred to via different facets. The subtypes are role (*Gassman, the actor - Gassman, the son*), location (*Iraq, the country - Iraq, the population*), organization (*McDonalds, the organization - the original McDonalds (a branch)*), information realization (*Gone With The Wind, movie - Gone With The Wind, book*) and representation (*Queen Elizabeth, the person - Queen Elizabeth, a portrait*).

2. **Meronymy:** A meronym is used to refer to the whole or a set. The subtypes include part-whole (*president Clinton - the US government*), entity-attribute (*alcoholic drinks - alcohol*) and set-set (*Jews - the crowd*). If meronymy and metonymy co-occur, metonymy is preferred.

3. **Class:** The two NPs share the type (*is-a* relationship), but they stand in a different position in the categorical hierarchy so that one can be viewed as more general (*attackers,* like *the two who killed 13 people*) or more specific (*his character - the characters*) to the other.

4. **Spatio-temporal function:** The discourse entity is split based on different values for its spatial or temporal characteristics: it is the 'same' entity or event but realized in another location or time. The subtypes include place (*New Year's Eve in New York - New Year's Eve in Potsdam*), time (*Postville - the*

old Postville), numerical (*temperature of 10°C - temperature of 1°C*) and role function (*the president of the US - the president of France*).

While Recasens et al. (2010a) reported the results of their study only for preselected NP pairs, in a follow-up paper, Recasens et al. (2012) showed that explicit near-identity annotation is a very difficult task for the annotators, who were able to identify only a small amount of the near-identity links in the corpus (6% and 2% for English and Catalan respectively). In particular, Recasens et al. (2012) claim that the human cognitive system tends to perceive reality in terms of simple dichotomies and ignores inconsistencies; therefore, the annotators overlook a certain number of near-identity cases that are actually present in the text. Thus, for the subsequent annotation rounds, Recasens et al. (2012) selected another approach, based on evaluating the annotators' disagreements on annotating identity coreference. In particular, if an anaphor-antecedent pair was annotated as coreferent and non-coreferent by different annotators (in their study they had 5), then this signalized a possible near-identity relation. Using this approach and after excluding the annotations on which all the 5 annotators agreed as well as cases annotated by only one of the annotators, Recasens et al. (2012) found that 25% of coreference annotations contain disagreements that can be labeled as near-identity. After merging the annotations, they concluded that the number of near-identical links in the corpus increased from 6% to 12% for English and from 2% to 16% for Catalan, which substantiates the applicability of the proposed method for the annotation of near-identity.

To our knowledge, the same annotation scheme has subsequently been applied to annotate the Polish Coreference Corpus by Ogrodniczuk et al. (2014); however, the inter-annotator agreement scores were quite low ($\kappa = 0.22$), which confirms the results of Recasens et al. (2012) regarding the explicit annotation of near-identity in full texts.

3.1.4 Evaluation of coreference annotations

In this subsection, we will briefly introduce the most widely used scoring metrics used to evaluate the quality of coreference annotations. Typically, evaluation of coreference resolution consists of two parts: (a) evaluation of the identification of mentions and (b) evaluation of linking these mentions into coreference chains. In the following, we will first focus on the evaluation of the identification of mentions and then present the most common metrics used to compute the linking of referring expressions into coreference chains.

The **identification of mentions** can be approached by evaluating the spans of the mentions detected by a coreference resolution algorithm. In this step, only the quality of identifying referring expressions in the text is evaluated, but not their relationship between each other. Typically, to achieve that, standard information retrieval scores are used in order to compute Precision (P) and Recall (R) of a coreference resolution system, defined as follows:

$$P = \frac{\text{Number of correctly detected mentions}}{\text{Number of mentions attempted to be detected}} \tag{3.1}$$

$$R = \frac{\text{Number of correctly detected mentions}}{\text{Number of all mentions}} \tag{3.2}$$

Thereafter, the F1 score is computed as a harmonic mean between Precision and Recall using the following formula:

$$F1 = \frac{2PR}{(P + R)} \tag{3.3}$$

As one can see from the equation, the F1 score reaches its best score at 1 and the worst score at 0, and it is frequently used to assess the performance of different coreference resolution systems.

As for the **linking of referring expressions**, the most common standard to evaluate the performance of coreference resolution systems derives from the recent coreference evaluation tasks, in particular, CoNLL-2011 (Pradhan et al., 2011) and CoNLL-2012 (Pradhan et al., 2012) shared tasks on coreference resolution. Specifically, in these tasks, several scoring metrics – MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF (Luo, 2005) – as well as their average (or *MELA* score) were used to evaluate the participating systems. In our experiments, we will adopt a similar approach to the evaluation of coreference annotations, and therefore, below, we will give a detailed overview of these scoring metrics; furthermore, we will also briefly outline the most common advantages and disadvantages for each of them. Although several other metrics for the evaluation of coreference exist (e.g., BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016) etc.), in this work, we will only focus on the commonly accepted metrics that constitute the MELA score, in order to be able to fairly compare our results to the related work.

Similarly, in order to score coreference chains and assess the quality of linking referring expressions with each other, Precision, Recall and F1 measures are used, counting the number of common and distinct links in the gold standard set as compared to the system output. In the description of the metrics, we will use the following terminology:

- **key set:** manually annotated coreference dataset used as a gold standard for the evaluation;
- **(system) response set:** output of a coreference resolution system.

Importantly, the main difference between coreference metrics is the method of counting these links across the datasets, which is described below for each of the metrics.

As mentioned previously, one of the earliest and the most widely used metrics to score coreference chains is the MUC score (Vilain et al., 1995). This metric treats coreference chains as equivalence classes of mentions, counting the minimum number of links between these mentions to be inserted or deleted when mapping the system response to the key set. In particular, Precision is computed as the number of common links between the key set and the system response divided by the number of links in the system response; Recall computes the number of common links divided by the number of links in the key set. However, as pointed out by Denis and Baldridge (2008), this metric has two major shortcomings. First, it does not take into account singleton entities, since they do not involve any links. Second, it favors long coreference chains: A system that produces a single chain would obtain 100.0 Recall without a considerable drop in Precision.

Another popular metric for scoring coreference chains is B^3 (or B-cubed, (Bagga and Baldwin, 1998)), which computes Precision and Recall scores for all the mentions in a document and then combines them to obtain the final scores. In particular, Precision and Recall are computed in the following way:

$$Precision(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|R_{m_i}|} \tag{3.4}$$

$$Recall(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|K_{m_i}|} \tag{3.5}$$

where R_{m_i} is the system response chain which includes mention m_i, and K_{m_i} is the key chain that also includes mention m_i. Then, Precision and Recall averaged over all mentions are used to compute the final F1 score. Since this score is mention-based, it is capable of taking into account singletons as compared to MUC, but it does not deal with key mentions that are not in the system response as well as system mentions that are not in the key set.

To overcome this shortcoming, the CEAF metric was introduced (Luo, 2005): It uses a mention-based ($CEAF_m$) or entity-based ($CEAF_e$) similarity metric for each

coreference chain, and subsequently calculates Precision, Recall, and F-measure for the best mapping. Below we provide the formula on the entity-based similarity for a key entity K_i and a response entity R_j, since this variant is more widely used for system evaluation:

$$\phi(K_i, R_j) = \frac{2|K_i \cap R_j|}{|K_i| + |R_j|} \tag{3.6}$$

CEAF Precision and Recall are then computed for the entities that have the best total similarity. In our study, we also use the entity-based variant, and we will subsequently refer to it just as CEAF.

As already mentioned earlier, the three metrics described above were used in the CoNLL evaluations to present coreference resolution results. Typically, F1 scores for each of the metrics are reported, and, additionally, an average over these metrics is also presented, in order to account for the benefits and drawbacks of each of the metrics and demonstrate a complete evaluation picture. To compute these scores, an official CoNLL scorer[18] (Pradhan et al., 2014) was released, which includes the official implementation of the metrics described above and is used as a standard software to evaluate coreference resolution systems.

3.2 Annotation projection

A projection approach is used to automatically transfer different types of linguistic annotation from one language to another. The idea of mapping from well-studied languages to low-resourced languages was initially introduced by Yarowsky et al. (2001) who studied the induction of POS and NE taggers, NP chunkers and morphological analyzers for different languages using annotation projection. Thereafter, the technique has been used for a variety of NLP tasks. Moreover, mapping from one source language (single-source projection) was substituted by mapping from multiple sources which turned out to be beneficial for e.g., POS tagging or syntactic parsing.

In the remainder of this section, we present an overview of the recent work on annotation projection, with a focus on projecting coreference annotations. We will first present a brief overview of the available parallel resources; then, we will discuss single-source projection approaches and move forward to the most recent multi-source algorithms.

[18] http://conll.github.io/reference-coreference-scorers/ [accessed on 03.11.2017]

3.2.1 Parallel corpora

A parallel corpus is usually defined as a collection of texts in one language (source language) and their translations into some other language (target language). In general, parallel corpora can not only be bilingual (thus containing only two languages), but also multilingual (thus containing more than two languages). Importantly, parallel corpora should be aligned, for instance at paragraph and/or sentence levels (Corpus alignment is described in detail in Chapter 5). Furthermore, parallel corpora differ in the direction of translation: They can be uni-directional (containing only translations from source to target), bi-directional (containing translations from both source to target and target to source), and multi-directional (multiple translations in different directions) (McEnery and Xiao, 2007).

Currently, the biggest resource that provides access to freely available parallel corpora is the OPUS collection of parallel texts[19] – an online resource that is being constantly updated. In addition, this collection provides tools for corpus processing as well as several search interfaces to query the data. According to Tiedemann (2012), already in 2012, OPUS contained over 90 languages as well as parallel data from several domains, including over 3,800 language pairs with sentence-aligned corpora comprising a total of over 40 billion tokens in 2.7 billion parallel units (aligned sentences and sentence fragments), and it is constantly growing. At that time, the largest domains covered by OPUS were legislative and administrative texts (mostly from the European Union and associated institutions), translated movie subtitles, and localization data from open-source software projects (Tiedemann, 2012).

One of the most influential parallel corpora released relatively early and now also available at OPUS is the EuroParl parallel corpus (Koehn, 2005). This corpus contains manual translations of European parliamentary debates in 21 European languages; additionally, Koehn (2005) provided bilingual document and sentence alignments for the corpus, taking English as the common source language (alignments for other language pairs can be obtained by using specific scripts). The largest parts of the corpus for language pairs such as English-German and English-Spanish amount to around 2 million parallel sentences. Due to its size and high quality of the translated data, EuroParl has been widely used in different kinds of computational linguistic studies.

To date, parallel corpora are gaining increasing importance, in particular, in cross-lingual Natural Language Processing and Statistical Machine Translation. In our

[19] http://opus.lingfil.uu.se [accessed on 03.11.2017]

study, the use of a parallel corpus is necessary for performing annotation projection experiments: An aligned dataset is required to establish correspondences between the source and the target annotations, and a large parallel corpus is needed to estimate the alignments (see Chapter 5 for more details).

Overall, parallel corpora are important resources that are being extensively exploited not only for contrastive corpus studies but also for developing cross-lingual methods in natural language processing. Specifically, parallel corpora are used as the basis for cross-lingual annotation projection, which will be discussed in detail in the next section.

3.2.2 Single-source annotation projection

Automatic annotation projection was introduced in the work of Yarowsky et al. (2001) who investigated projection algorithms in order to induce POS and Named Entity (NE) taggers, NP chunkers and morphological analyzers for French, Chinese, Czech and Spanish. They used statistical word alignments to automatically transfer annotations and subsequently train new taggers on the projected data. An example approach of Yarowsky et al. (2001) to projection is introduced in Figure 3.1, where part-of-speech tags are automatically transferred from source (English) to target (French) via word alignment information. In this example, they used labeled English data and an aligned parallel corpus to automatically create mappings between the annotations from the source side and the corresponding aligned words on the target side: If an English word with a certain POS tag was aligned to some French word, then that French word was assigned the same POS tag as the English one. Furthermore, Yarowsky et al. (2001) presented several techniques to overcome noise in the projected data coming from statistical alignments, relying on multiple source annotations (see 3.2.3 for details).

Following the pioneering work of Yarowski, annotation projection was applied to multiple NLP tasks, including mention detection (Zitouni and Florian, 2008), named-entity recognition (Ehrmann et al., 2011), and semantic role labeling (Padó and Lapata, 2009). In addition, annotation projection has been widely exploited for cross-lingual dependency parsing, starting with the work of Hwa et al. (2005) who were the first to project dependency trees from English to Spanish and Chinese. Other notable approaches include e.g., (Ozdowska, 2006, Ganchev et al., 2009, Spreyer and Kuhn, 2009).

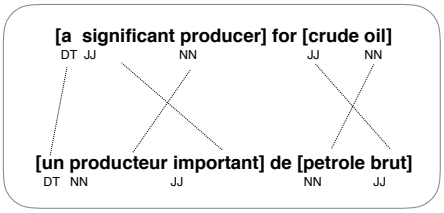


Figure 3.1: Direct projection algorithm (Yarowsky et al., 2001)

To our knowledge, the first application of annotation projection to the coreference task was done by Harabagiu and Maiorano (2000), who experimented with manually projecting coreference chains from English to Romanian using a parallel corpus. In particular, they manually translated MUC-6 and MUC-7 training corpora into Romanian (using native speakers) and marked all the translated referring expressions corresponding to the English ones. They then used the aligned antecedents to automatically resolve their anaphors using bilingual texts and a set of heuristic rules. They showed that a coreference resolver trained on a bilingual corpus can achieve better Precision than the one trained on monolingual data, with the improvement from 84% to 87% for English and from 72% to 76% for Romanian, which is due to more powerful heuristics derived from two languages rather than one.

This work was followed by the work of Postolache et al. (2006) who used automatic word alignments to automatically project coreference annotations for the same language pair. Their main goal was to use annotation projection as a preprocessing step prior to manual correction, in order to create a large Romanian corpus. Therefore, to improve the results, in addition to word alignments, they used automatic POS tagging and syntactic head information (obtained from the output of a syntactic parser) in their approach.

In their work, Postolache et al. (2006) first introduced the steps necessary for the application of a direct projection method to transfer coreference annotations:

1. Automatic word alignment between English and Romanian parts of the corpus using a Romanian-English word aligner.

2. Extraction of Romanian referring expressions corresponding to English referring expressions: For each English referring expression (RE) which spans n words

$e_1, e_2...e_n$, extract the corresponding set of Romanian words $r_1, r_2...r_m$, to which the English words are aligned. After removing the duplicates from the Romanian set of words, Postolache et al. (2006) take the span between the first and the last aligned words as the new Romanian referring expression. As head of the RE, they mark the word aligned to the head of the corresponding English referring expression.

3. Transfer of coreference chains: Since the English referring expressions are grouped into coreference chains on the source side, the same grouping is transferred into the target side, hence the transferred REs are assigned the same cluster as their corresponding source ones.

In this manner, Postolache et al. (2006) were able to transfer coreference chains from the English to the Romanian side. During the transfer, they noticed several cases that may occur given many-to-many alignments:

(a) An English RE has a corresponding Romanian RE with one head;

(b) An English RE has a corresponding Romanian RE with two or more heads;

(c) An English RE has a corresponding Romanian RE with no head;

(d) An English RE has no corresponding Romanian RE.

Postolache et al. (2006) applied several filtering heuristics to achieve higher precision. First, they discarded those referring expressions, the syntactic heads of which were not properly aligned, which correspond to cases (c) and (d) from the list above. Second, in the projected annotations, they filtered out target referring expressions with incorrect POS tags. The post-filtering results computed only for the heads of the referring expressions indeed showed high Precision (over 95%), but considerably lower Recall (around 70%)[20]. Conversely, the results for complete set of REs without filtering exhibit lower Precision (around 50%), but higher Recall numbers (around 80%). Overall, their error analysis showed that, apart from the wrong alignments, several language-specific divergences were responsible for the incorrectly mapped mentions.

Another early attempt that should be mentioned is the work of Mitkov and Barbu (2002) who enhanced anaphora resolution using a parallel English-French corpus. Technically, they did not use annotation projection as such but implemented a mutual enhancement algorithm for anaphora resolution that was able to benefit from using

[20]Postolache et al. (2006) used percentage to report their results.

information from both languages, e.g., exploiting gender discrimination in French for English anaphora resolution. Similar to Harabagiu and Maiorano (2000), they showed that coreference resolution can benefit from using information from two languages: Their bilingual approach led to an improvement in the success rate of roughly 4% for both English and French.

Thereafter, Souza and Orăsan (2011) went one step further and not only projected coreference chains in a bilingual corpus to obtain a new dataset, but also made an attempt to use this dataset to train a new coreference resolver. Specifically, they used an English-Portuguese coreference corpus and implemented a projection algorithm to automatically transfer coreference chains from English to Portuguese. As opposed to the previous approaches, they used an automatic English coreference resolver – Reconcile (Stoyanov et al., 2010) – to annotate the source texts which resulted in a poor target coreference resolution quality: The system trained on the projected annotations did not outperform a simple head-match baseline[21]. The authors claim that the poor performance is due to the use of an automatic coreference resolver with a bias towards identifying head match as coreference which resulted in low-quality projected annotations; they suggest using a better source coreference resolver to improve the quality of their approach.

The next steps in projecting coreference include several translation-based approaches. The difference is that a target text is first translated into the source language, on which coreference resolution is performed; after that, the source coreference chains can be projected back to the target side. This approach was used, for example, by Rahman and Ng (2012) to train coreference resolvers for Spanish and Italian using English as the source language. Both the translation from source to target and coreference resolution on the source side were done fully automatically. Thereafter, Rahman and Ng (2012) introduced the following projection settings:

(a) no additional linguistic taggers: In this setting, the authors assumed that there were no additional linguistic resources available for the target language;

(b) only a mention extractor available: A mention extractor is applied to the target text, and the annotated mentions are subsequently projected into the source side, where they are used as the input for the coreference resolution system;

[21]Simple head-match algorithms for coreference typically consider mentions with the same syntactic heads as coreferential.

(c) additional resources can be used: additional syntactic and semantic taggers can be used on the target side in order to generate additional linguistic features for the target coreference resolution.

After training coreference resolution systems in all settings, the authors reported on the results in each of them. While the results for setting (a) are rather poor, the systems trained in setting (b) already give around 90% of the average F-scores of a supervised resolver in experiments with both languages (average F1 of 54.9 for Italian and 46.8 for Spanish); obviously, the systems trained in setting (c) exhibit best results, with F1 improved by 5.3 points for Spanish and 0.7 points for Italian.

Similarly, Ogrodniczuk (2013) experimented with translation-based projection for English and Polish using only a mention extractor. The evaluation of the quality of the projected annotations on manually annotated data showed 70.31 F1 which suggested the promise of the projection approach for the subsequent system training for an inflectional language.

The most recent implementation of projection using a parallel corpus is due to Martins (2015) who experimented with transferring automatically produced coreference chains[22] from English to Spanish and Portuguese, and subsequently trained target coreference resolvers on the projected data. The novelty of his approach is in using softmax-margin posterior regularization that helps achieve robustness across word alignment errors, which is compared to direct projection and several other baselines[23]. The approach of Martins (2015) outperforms these baselines, with the average of 38.82 F1 for Spanish and 37.23 for Portuguese coreference systems (the results for the systems trained only using direct projection results are 34.30 and 31.77 respectively). These results are competitive as compared to the performance of fully supervised systems: 43.93 for Spanish and 39.83 for Portuguese.

In sum, most approaches to cross-lingual annotation transfer have used English as the source language because of the availability of parallel resources for English and most other languages. It has only recently been shown that projecting from multiple other languages – or *multi-source* annotation projection – prevents overfitting to the peculiarities of the source language. We provide the overview of the most recent work on that topic in the next section.

[22] using Berkeley coreference resolver (Durrett and Klein, 2014)

[23] Such as a simple deterministic baseline that selects the closest mention, or a delexicalized transfer baseline.

3.2.3 Multi-source annotation projection

The idea of using multiple sources comes from the work of Yarowsky et al. (2001), who used multiple source annotations to overcome noise in the projections. In particular, they use (a) multiple translations in the same language and (b) multiple translations in different languages to improve the quality of the projected annotations, which served as a basis for the follow-up work on annotation projection.

First, Yarowsky et al. (2001) experimented with multi-source approaches for inducing lemmatization in new languages. However, their task setting was quite specific: They applied projection to identify correct lemmas in the target language (French) by projecting a target word into the source language (English), automatically lemmatizing it and projecting it back to French (see Fig. 3.2 for illustration). Since some of the words may be missing or misaligned on the source side, they utilized multiple versions of the source English text[24], as they may exhibit different links to the target word. As a result, they concatenated different alignments together, and used the repeated words to measure the confidence in some particular alignment.

Subsequently, Yarowsky et al. (2001) exploit multiple translations into different languages to improve the lemmatization induction by projecting from English and French into Spanish. Since they only had access to one lemmatizer (for English), they first needed to apply a single-source projection to lemmatize the texts in the second source language (French) by using the approach described above. Thereafter, they used these automatically induced French annotations as an additional source for projecting into the actual target language (Spanish), which is illustrated in Fig. 3.3: The Spanish verb *creyeron* is projected into both English and French, where the corresponding lemma is found, which is then projected back to Spanish.

Overall, their results have shown that using multi-source strategies helps to improve the Precision of their annotation projection method. In particular, the projection from English into Spanish improves the Precision scores for the morphology induction from 0.966 to 0.976 using multiple versions of the source English texts, and to 0.974 using an additional source language (French).

Thereafter, the idea of using several sources has been widely used for syntactic parsing. The best unsupervised dependency parsers nowadays rely on annotation projection and were first presented in the work of (Rasooli and Collins, 2015). Rasooli and Collins (2015) train dependency parsers on the projected annotations and experiment with multi-source setting as a side task, reporting on the improvement

[24]In particular, they used Bible texts since numerous English Bible versions are freely available.

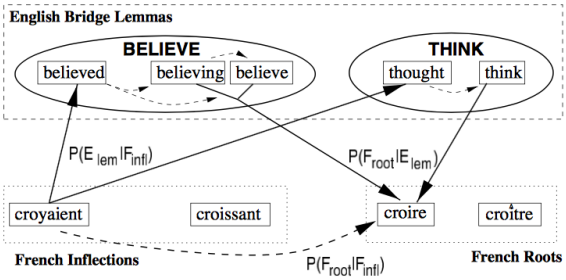


Figure 3.2: Inducing French lemmatization by using single-source annotation projection (Yarowsky et al., 2001)

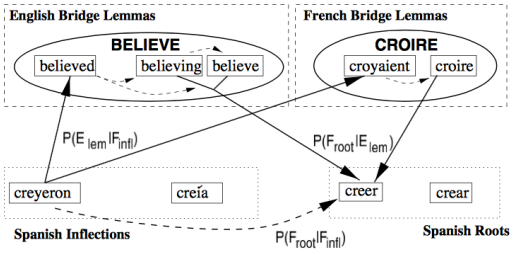


Figure 3.3: Inducing Spanish lemmatization by using multi-source annotation projection (Yarowsky et al., 2001)

achieved by using a multi-source strategy for 6 target languages present in both Google Dependency Treebanks and the EuroParl corpus[25].

In particular, Rasooli and Collins (2015) transfer dependency trees using word alignments and consider two methods for combining information from multiple source languages:

- **Method 1: Concatenation.** Data is obtained from each of the languages separately and then concatenated.

- **Method 2: Voting.** Assuming that each target languages is aligned to each of the source languages, projected annotations are obtained using a voting scheme, e.g., taking the most frequent tag across all languages.

[25]Rasooli and Collins (2015) use English as source in a single-source setting and German, Spanish, French, Italian, Portuguese and Swedish as targets; in a multi-source setting, they use $n-1$ languages as source.

Interestingly, the average accuracy of the best-performing model trained on the projected trees slightly improves from 78.89% for a single source (a) to 81.23% for multiple sources using the concatenation method and (b) to 82.18% using the voting method. It is worth noticing that German has the lowest accuracy out of all the cross-lingual models which, as the authors hypothesize, might be due to the different word order leading to lower alignment quality as compared to other languages.

Thereafter, the idea of multi-source annotation projection was further elaborated in the works of Agić et al. (2015), Agić et al. (2016), and Johannsen et al. (2016). Beginning with multi-source annotation transfer for POS tagging and building upon each other, these works present algorithms for multi-source transfer for a large set of languages, most of which are non-European and low-resource. All these works benefit from using multiple translations of the Bible which are available in hundreds of languages.

Firstly, Agić et al. (2015) present a method for learning part-of-speech taggers for low-resource languages. They applied a projection algorithm to transfer POS tags from those languages for which POS taggers exist and aggregated over these tags to learn POS taggers for 100 languages. Interestingly, most of the languages they used were non-European which posed additional difficulties for their task setting.

Secondly, the same group adapted a similar multi-source projection approach for multilingual dependency parsing. In addition to POS tags projection similar to Agić et al. (2015), Agić et al. (2016) projected dependency edges with associated weights from source to target languages. They used the data from Universal Dependency treebanks (UD) and several additional test sets that conform to UD to train source language parsers for 30 languages in total, and projected source dependency edges along with their weights from multiple source languages to the target side. Thereafter, they totaled the projected weights scaled by the corresponding alignment probability in order to rate the incoming projections.

Their method currently outperforms several commonly used baselines (delexicalized transfer, direct single-source projection, reparsing), achieving bigger improvements for non-Indo-European languages. Interestingly, German, again, has one of the lowest scores as compared to other European languages, e.g., 69.97 for POS as compared to 78.92 for English, 80.36 for French, 86.28 for Norwegian, and 86.28 for Swedish; 45.73 vs. 60.0, 56.64, 66.8 and 65.28 for parsing respectively.

Finally, the improvement of this approach was presented in the work of Johannsen et al. (2016). Not only did they project from multiple sources, but they also adapted

an algorithm for jointly projecting annotations for two mutually independent tasks – POS tagging and dependency parsing – to train supervised dependency systems.

Since the resulting dataset for a particular language may include projections from multiple source languages[26], Johannsen et al. (2016) counted the percentage to which certain sources contribute to the target. For instance, for German, the languages that contribute most of the projections (between 7 and 16%) are Norwegian, Italian, Indonesian and Swedish. On average, a German sentence has edges from 4.1 source languages.

In sum, annotation projection from multiple sources is a rapidly developing approach which has already proved its efficiency for the tasks of POS tagging and dependency parsing, but has not yet been implemented for coreference.

[26]As only the highest scoring projection is counted.

Chapter 4

Multilingual coreference corpus

This chapter introduces the multilingual corpus created for the cross-lingual study of coreference phenomena in the three languages (English, German, Russian) and for running experiments on annotation projection. In the following, we describe the process of building the corpus (Section 4.1) and the development of trilingual annotation guidelines used to annotate the corpus (Section 4.2). Furthermore, we measure inter-annotator agreement for different relations and report on the results (Section 4.3). Finally, we present the final version of the corpus, including the statistics of the corpus and the linguistic analysis of the resulting annotations (Section 4.4).

Previously published material

An earlier version of the corpus annotated only with identity coreference was described in (Grishina and Stede, 2015), and the corresponding version of the annotation guidelines was made available as an unpublished manuscript (Grishina and Stede, 2016). The bridging and near-identity annotation scheme, as well as the corpus analysis, were published as (Grishina, 2016). The complete version of the annotation guidelines is published as part of this thesis (see Appendix A).

4.1 Data collection

In this subsection, we first (a) define the requirements for the contents of our corpus, (b) provide an overview of sources offering parallel data, and (c) describe the selection process and give examples from the resulting collection.

Firstly, our general aim was to manually build a multilingual corpus that would be able to cover a wide range of coreference phenomena in the languages in question. As the study of Kunz (2010) has shown, coreference chains vary considerably in terms of their length and types of referring expressions depending on the text genre: For instance, looking at essays, fiction and popular scientific texts, Kunz (2010) already observed discrepancies in the average length of coreference chains in German, which vary between 3.5 and 7.53 markables. Subsequently, Kunz et al. (2016) demonstrated that, for the English-German language pair, differences in genre are more significant than differences between languages. Therefore, for our corpus, we decided to select texts of several genres, which would represent diverse types of referring expressions and their relationships.

Secondly, as described in Chapter 3.2, annotation projection experiments usually rely on a parallel bilingual corpus, which enables transferring the annotations from source to target using alignment information. In our case, the corpus has to be parallel and trilingual, thus enabling different directions of projection and also multi-source projection.

That being said, we define the following requirements for our corpus:

(a) it should be trilingual, thus containing texts in English, German and Russian;

(b) it should be parallel, thus the texts being parallel between all the language pairs;

(c) it should contain multiple genres, thus covering various coreference phenomena.

At present, several research projects provide collections of parallel texts in different languages acquired from the web, most of them are already aligned and publicly available for research purposes (e.g., the OPUS project[1]). However, there is a much smaller amount of texts in different genres available for Russian than for German and English; moreover, some of them contain incorrect translations that provide obstacles for the subsequent annotation. To build our corpus, we first started by collecting parallel texts of the two most common genres of written text – newswire texts and

[1] http://opus.lingfil.uu.se [accessed on 21.09.2017]

narratives – and, at a later stage, we added another genre – medicine instruction leaflets – which, however, were only available for the English-German language pair. In particular, the newswire texts were selected from the Project Syndicate newswire agency[2], which provides translations of its articles in multiple languages, while the narrative texts (short stories) were extracted from a portal distributing materials to second language learners[3]. Medical texts are from the European Medicines Agency (EMEA) subcorpus of the OPUS collection of parallel corpora. Below, we provide example fragments for newswire, narratives and medicine leaflets respectively[4]:

(47) (a) *Last month's terrorist assault in Mumbai targeted not only India's economy and sense of security. Its broader goal was to smash the India-Pakistan détente that has been taking shape since 2004.*
(b) *Daisy Hamilton was a private detective. She was thirty years old and had been a detective for the past two years. Every morning she went to her office to wait for phone calls or open the door to clients needing her services.*
(c) *Abilify is used to treat moderate to severe manic episodes and to prevent manic episodes in patients who have responded to the medicine in the past. The solution for injection is used for the rapid control of agitation or disturbed behavior when taking the medicine by mouth is not appropriate.*

Overall, the corpus consists of 38 parallel texts in English, German and Russian, with the following distribution of texts across genres: newswire articles (7 texts per language), short stories (3 texts per language), and medicine instruction leaflets (4 per language, only English-German). This choice is motivated by the fact that all the three genres are important for many NLP applications, but at the same time they exhibit many differences to each other and are therefore particularly interesting for a comparative analysis of coreference chains as well as annotation projection experiments.

Corpus statistics for the unaligned texts are provided in Table 4.1. As one can see from this table, the number of sentences and tokens across languages is quite comparable. Due to the fact that medical texts were already pre-aligned at the sentence level, we were able to select exactly the same number of sentences for English and German. Moreover, we compute the average sentence length as the number of tokens divided by the total number of sentences. Interestingly, we see that Russian

[2]https://www.project-syndicate.org [accessed on 21.09.2017]
[3]http://www.lonweb.org [accessed on 21.09.2017]
[4]Examples (47)a - (47)c are taken from the corpus developed as part of this work.

	Newswire			Stories			Medicine		Total		
	En	De	Ru	En	De	Ru	En	De	En	De	Ru
Tokens	5903	6268	5763	2619	2642	2343	3386	3002	11908	11912	8106
Sentences	239	252	239	190	186	192	160	160	589	598	431
Tokens/Sent	24.7	24.9	24.1	13.8	14.2	12.2	21.2	18.8	20.2	19.9	18.8

Table 4.1: Statistics for the raw (unaligned) corpus

texts contain shorter sentences than German and English, which is stable for both stories and newswire (24.1 token for Russian vs. 24.7/24.9 for English and German for news, 12.2 vs. 13.8, 14.2 for stories). At the same time, for English and German, the average sentence length is almost equal, with the difference ranging between 0.2 and 0.4 tokens. The only exception are medical texts, in which an average English sentence is 2.4 tokens longer than a German one.

4.2 Annotation guidelines

For the annotation of the corpus, we developed common annotation guidelines applicable to the three languages. In contrast to the usual procedure when coreference annotation guidelines are designed with one target language in mind, our goal was to have common guidelines for the three languages, in order to (i) obtain uniform nominal coreference annotations in our corpus (supporting the projection task), and (ii) facilitate extension to further languages.

In the guidelines, we focused on the following coreference relations, with each being described in a separate section of the annotation manual:

- *identity:* Two nominal expressions have exactly the same referent;
- *near-identity:* Two nominal expressions are partially the same in that they share most of the important characteristics but differ in one crucial dimension;
- *bridging:* Two nominal expressions refer to two objects that are related but are neither identical nor near-identical. The most common bridging relations are e.g., *part-whole* or *set-membership*.

In the remainder of this section, we introduce the typology of relations adopted for the annotation of the corpus, define markable types and spans, and describe the annotation flow and technical details related to the corpus creation. We also report

on the inter-annotator agreement for each of the relations. The complete annotation manual can be found in Appendix A.

4.2.1 Typology of relations

Identity. The guidelines for identity coreference present instructions for the annotation of nominal coreference chains. They were developed based on the most prominent annotation schemes for the languages in question. In particular, we relied upon the OntoNotes guidelines for English (Hovy et al., 2006), the language-neutral Potsdam Coreference Scheme (PoCoS) (Krasavina and Chiarcos, 2007), and on the parallel English-German guidelines for the annotation of nominal anaphora ParCor (Guillou et al., 2014).

While the OntoNotes guidelines distinguish between two relations – identity coreference and apposition – we include appositions into the markable span and do not annotate them separately. Moreover, in contrast to ParCor guidelines, which consider only pairwise annotation of anaphoric pronouns and their antecedents, we annotate all REs appearing in a coreference chain (i.e., that are mentioned in the text at least twice). Mentions that appear in a text only once (singletons) are not annotated according to our guidelines. For example:

(48) Much of the current trouble can be traced to [Afghanistan]$_1$, [whose]$_1$ tragedy could never have remained confined within [its]$_1$ designated borders. The dynamics of [the region]$_1$ changed when the Afghan freedom fighters of the 1980's were converted into mujahidin through a criminal enterprise.[5]

As one can see from this example, the annotated mentions of different type (*Afghanistan, whose, its, the region*) all form a coreference chain; the singletons are not annotated.

Bridging. We base our work on the main principle identified by Clark (1975): We assume that the speaker intends the listener to be able to compute the shortest possible bridge from the previous knowledge to the antecedent which is unique and determinate in the natural language discourse. Following this principle and the majority of the related work on bridging (e.g., Riester and Baumann (2017)), we assume that this uniqueness in languages such as English and German is expressed by the definite article, and therefore only definite descriptions can be annotated as bridging anaphors. By taking this decision, we aim at reducing the degree of uncertainty that

[5]The example is taken from the corpus developed as part of this work.

would pose challenges for our cross-lingual study. However, it should be noted that not all referents that appear in the text for the first time and are expressed by definite descriptions have a bridging antecedent – some of them are definite due to the common knowledge shared by the speaker and the listener.

In our pilot experiments, we identified several bridging categories, which were common across genres, and applied them to annotate the corpus. We defined these categories based on the previous work (see 3.1.2 details) as well as preliminary annotation rounds, during which we manually collected bridging examples from several texts not included in our corpus (newswire and medical texts were taken from the same sources; the narrative text was taken from a different source since we did not have any short stories from the same source available). However, the collected cases were used merely as an illustration; we did not limit the categories introduced in this thesis to the types of relations found in these texts.

Below, we describe these categories and give typical examples from different genres for each of them.

1. **Physical parts - Whole**

   One NP represents a physical part of the whole expressed by the other NP. For example:

   - *the militant organization - the offices in the whole country*
   - *the telephone - the dial pad*
   - *the knee - the bone*

2. **Set - Membership**

   Sets can be represented by multiple entities or events. One can refer to a certain subset or to a single definite element of the set and bridge from this subset or element to the whole collection. We do not distinguish between sets and collections, as is done in some of the related work[6].

   A. SET-SUBSET

   - *the European Union - the least developed countries*

[6]For instance, (Gardent et al., 2003) postulate that a collection differs from a set in that it is related to its elements based on a spatial or social connection (e.g., *the class - the student*) rather than on physical similarity (e.g., *the forest - the tree*).

- *the patients - the patients treated with Abraxane*

b. set-element

- *these studies - the main study*
- *Pakistan major cities - the most populous city*

3. **Entity - Attribute/Function**

   An entity is a person or an object that has certain attributes characterizing it and certain functions it fulfills with respect to some other entity.

   a. Entity-attribute

   - *Kosovo - their current policy of rejection*
   - *Mrs. Humphries - the monotonous voice*

   b. Entity-function

   This relation involves a bridge holding between individuals with one of the related individuals being described by his profession or function with respect to the other (Gardent et al., 2003).

   - *Trends, the shop - Mr. Rangee, the owner*
   - *Kosovo region - the government*

4. **Event-attribute**

   Core semantic frame elements of events are commonly time and place, while optional ones can include duration, participants, explanation, frequency etc. From these frame elements, one can bridge to the event itself.

   - *the regional conflict - the trained fighters*
   - *the surgical intervention - the operating room*

5. **Location - Attribute**

   As locations, we consider geographical entities that have permanent locations in the world. Such locations exhibit different semantic frames as compared to entities and events.

- *Germany - in the south*
- *Afghanistan - the population*

6. **Other**

   Other bridging relations (if any), that cannot be described using the categories presented above.

If the antecedent of a bridging markable is contained inside the same NP, we mark such NPs as BRIDGING-CONTAINED, following Riester et al. (2010). For example:

(49) [The wheel of [the bike]$_{B1}$]$_{B1}$ was completely broken.[7]

In this case, we link *the wheel of the bike* to the closest antecedent *bike*, which is a part of the same NP, and we mark it as bridging-contained. However, German compound nouns are not considered as such (e.g., one cannot split the compound *Tischbein* ('table leg') into *Tisch* and *Bein* and establish a relation between the two).

Near-identity. We used the definitions provided by Recasens et al. (2010a) and made an attempt to apply them to our texts. The annotators' goal was to extend existing annotations on top of the identity coreference. We only chose three top categories out of four described in 3.1.3, including but not distinguishing among their subtypes. Specifically, the annotation scheme provided to the annotators comprised the following relation types:

1. **Name metonymy**: The same entity is referred to via different facets, such as role, location, information realization (*Kosovo, the country - Kosovo, the people* – location);

2. **Meronymy**: The same entity is referred to by a meronym (*Kosovo - the government*);

3. **Spatio-temporal function**: The same entity is referred to via its different realizations in space and time (*Budapest - the medieval Budapest* – time).

In order to differentiate between the category of meronymy, which is common for both near-identity and bridging, we introduced the principle of primacy (see 4.2.4), according to which, in case of doubt, identity was preferred over near-identity and near-identity over bridging.

Thereafter, we merged the annotations from the first and the second annotator, since only a small number of near-identity markables was found in the corpus. For that reason, we did not compute inter-annotator agreement for near-identity.

[7]The example is taken from the work of Riester et al. (2010).

4.2.2 Types of markables and markable spans

In this subsection, we define which types of expressions can serve as markables for the three coreference relations described above. Syntactically, we define markables as phrases with nominal or pronominal heads; in contrast to OntoNotes, we exclude verbs from our annotation. According to our guidelines, the annotators should consider the following referring expressions as markables:

1. FULL NOMINAL PHRASES, e.g., *the big blue sky;*

2. PROPER NAMES AND TITLES, e.g., *Mr. Black;*

3. PERSONAL PRONOUNS (FIRST, SECOND AND THIRD-PERSON):

   Personal pronouns are only annotated if they have a referent in the text. This can be illustrated by the following examples[8]:

   (50) Hello, can [I]$_1$ help [you]$_2$? - [Daisy]$_1$ asked [the lady]$_2$.

   (51) If *you* need more information about *your* medical condition, read the Package Leaflet.

   In (50), the annotators were instructed to annotate the first-person pronoun *I* as referring to the specific antecedent *Daisy* and the second-person pronoun *you* as referring to the specific antecedent *lady*. In (51), the annotators did not have to annotate the personal pronouns *you* and *your*, since they do not have any antecedent in the text but refer to the abstract reader.

   Furthermore, the first-person pronouns were not annotated if they denoted the author of the text (and did not have an antecedent in the text), as in the following example:

   (52) *I* am sure, *our* time for standing pat, for protecting narrow interests and putting off unpleasant decisions - that time has surely passed.

4. DEMONSTRATIVE PRONOUNS, such as *this, that etc.*

   (53) You need [a camera]$_1$ [that]$_1$ works in the dark. Hm, take [this]$_1$.

[8]Examples (50)-(54), and (57) are taken from the corpus developed as part of this thesis.

In (53), the demonstrative pronoun *this* points back to its antecedent *a camera* mentioned in the previous sentence and must be annotated. Since only nominal coreference was annotated according to our guidelines, demonstrative pronouns were not annotated if they refer to a verbal phrase or to a bigger discourse unit, as in the following example:

(54) The London G-20 meeting recognized that *the world's poorest countries and people should not be penalized by a crisis for which they are not responsible*. With *this* in mind, the G-20 leaders set out an ambitious agenda for an inclusive and wide-ranging response.

In (54), *this* refers to the whole subordinate clause of the previous sentence and therefore should not be marked.

5. RELATIVE PRONOUNS, such as *who*, *whom*, *whose*, *which*, *that* etc.

6. REFLEXIVE PRONOUNS, such as *himself, herself etc.* For German, reflexive pronouns were annotated only if they were independent constituents, but not part of a reflexive verb. This was decided by applying the following test: If the position of the reflexive pronoun can be changed, then the pronoun is an independent unit, otherwise it belongs to the verb. For instance:

(55) Ich habe *mich* gestern gewundert. (*Mich habe ich gestern gewundert.)

(56) Ich habe [mich]$_1$ gestern gesehen. (Mich habe ich gestern gesehen.)

In (56), *mich* is annotated as a markable, while in (55) it is not a syntactically independent unit.

Moreover, reflexivity in German and in Russian can also be marked by other units, such as *selbst*, *selber*, *persönlich* or *своѝ* (svoj), *себя* (sebja), *сам* (sam) that also must be annotated.

7. PRONOMINAL ADVERBS: This is a special category of adverbs that are formed by combining a pronoun and a preposition; in German, these adverbs often co-refer with an NP. For example:

(57) Viele Amerikaner haben Probleme mit [Rassismus]$_1$; doch wir sind [dagegen]$_1$ immun.
Many Americans have problems with [racism]$_1$; but we are immune against [it]$_1$.

In this example, the pronominal adverb *dagegen*, which literally means 'against it', co-refers with the noun *Rassismus* ('racism').

8. S/HE, HE/SHE, HE OR SHE, HIS/HERS, HIS OR HERS: Each of these forms should be treated as a single pronoun and linked to its antecedent, as in the following example:

   (58) If [your child]$_1$ is thinking about a gap year, [he or she]$_1$ can get good advice from this website[9].

9. NPS WITH QUANTIFIERS, such as *all people*, *two people*, *105 Million euro* etc. As stated in (Krasavina and Chiarcos, 2007), the following test can be applied to identify the definiteness of an NP: If, after inserting a definite article or a demonstrative pronoun, the meaning of the noun phrase is not changed, then this NP is definite. We use the same test to check the definiteness of an NP. If an NP is definite, then it can be marked as an anaphor; otherwise, it can only begin a coreference chain.

10. NOMINAL PREMODIFIERS

    The category of nominal premodifiers is typical for English. We annotate nominal premodifiers if they refer to a named entity (*[the [US]$_1$ politicians]$_2$*, *[the [FBI]$_1$ agent]$_2$*) or are expressed by an independent noun in the genitive form (*[[creditor's]$_1$ choice]$_2$*); in all other cases, nominal premodifiers are not annotated as separate markables (e.g., *bank account* is annotated as a single markable).

11. GENERIC REFERENCE

    Generic nouns can co-refer with definite NPs or pronouns, but they cannot be linked to other generic nouns. For example:

    (59) [Computers]$_1$ are expensive. But [they]$_1$ are really useful. *Computers* cost a lot of money.

    In this case, the annotators only linked the anaphoric pronoun *they* to its antecedent in the first sentence, *computers*, but did not annotate the generic noun *computers* in the third sentence.

[9]The example is taken from the ParCor guidelines (Guillou et al., 2014).

12. TEMPORAL EXPRESSIONS, such as *this year*.

    Temporal expressions were annotated if they co-refer with each other or with a noun phrase.

Overall, all the types of referring expressions described above were annotated as markables if they have a referent in the text. Syntactically, markables are always rooted in some nominal phrase, which, however, can additionally include other dependent constituents. In order to facilitate the identification of markables by the annotators, we define markable spans as follows:

- the syntactic head of the NP;
- determiners and adjectives (if any) that modify the NP;
- deverbal modifiers (participial constructions, regardless of whether they are in pre- or postposition) that can be substituted by a subordinate clause, for example:

  (60) [Regional conflict, involving all of the region's states and increasing numbers of non-state actors]$_1$, has produced large numbers of [trained fighters, waiting for the call to glory]$_2$.[10]

  In this case, both *regional conflict, involving all of the region's states and increasing numbers of non-state actors* and *trained fighters, waiting for the call to glory* are markables.

- dependent prepositional phrases (for example, *[Queen of England]*$_1$).
- appositions, i.e., additive material that is not syntactically integrated, are included into the markable span, but are not annotated separately:

  (61) [JuD, Party of Proselytizing,]$_1$ was founded in 1972.

However, full clauses, in particular relative clauses, were not taken as parts of the markable rooted in the NP head. As opposed to OntoNotes (Hovy et al., 2006), relative pronouns were annotated separately and linked to their antecedents. Furthermore, if a nominal phrase was governed by a preposition, the preposition was excluded from the markable span, e.g., for the prepositional phrase *by the sea* only *the sea* was marked as a markable span. The only exceptions were German contractions of an

[10]The examples (60)-(61) are taken from the corpus developed as part of this thesis.

article and a preposition (such as *zum Bahnhof* ('to the train station')), which were included into the markable span. This decision is different to the PoCoS guidelines (Krasavina and Chiarcos, 2007) that treat prepositional phrases as markables.

4.2.3 Markable attributes

For each markable, a set of attributes was specified, which were used to analyze the structure and characteristics of markables in the corpus across languages and genres. In particular, the annotators had to select from a list of the following set of pre-defined properties:

a. REFERENTIALITY: the referential status of a markable, such as discourse-new, discourse-old or cataphoric.

b. DIR_SPEECH: whether a markable appears in direct/indirect speech or not.

c. PHRASE_TYPE: all markables are nominal phrases by default; used only to mark German contractions as described in Section 4.2.2.

d. NP_FORM: type of the noun phrase as described above, such as Named Entity, personal pronoun etc.

e. AMBIGUITY: whether the antecedent of a markable is ambiguous or not.

f. COMPLEX_NP: complex NPs are those containing embedded NPs, deverbal modifiers or appositions.

g. GRAMMATICAL ROLE: the grammatical role of a markable, such as subject or direct object.

h. COMMENT: any comments on the markable annotation the annotators might have.

Furthermore, for bridging and near-identity, the annotators had to select the category of the relation, such as part-whole or name metonymy respectively, which correspond to the bridging and near-identity categories defined above. Also, for bridging, there was an option of marking a noun phrase as bridging-contained.

4.2.4 Annotation process

To optimize the annotation process, in the guidelines, we described the annotation flow and postulated several principles in order to resolve controversial issues. In the following, we present the annotation process and list the annotation principles in hierarchical order.

For coreference chains, the annotators had to select only those nominal expressions ('markables') that actually appeared in a coreference chain (i.e., those that were mentioned at least two times in the text). In particular, the annotators had to check every referring expression and investigate whether it anaphorically referred to an entity that had already been mentioned. If this was the case, then they had to create a markable and link it to its nearest antecedent. When some entity was mentioned only once by some referring expression (a so-called 'singleton'), this expression was *not* a markable and therefore was not annotated. Furthermore, it should be noted that cataphoric pronouns were also annotated, but the relation was established in forward direction. For example:

(62) Before [she]$_1$ left, [Sue]$_1$ locked the door.

In this example, although 'she' precedes 'Sue', the relation had to be established from 'she' to 'Sue'.

The annotation process for bridging was similar to those for identity coreference: The annotators had to carefully examine each definite noun phrase that was not linked to any preceding antecedent. If it was definite due to common knowledge, then it was not annotated, otherwise the annotators were instructed to establish the corresponding bridging relation to its antecedent.

According to our guidelines, bridging and near-identity relations are generally directed from right to left and were annotated separately from identity relations. Each markable could have only one outgoing relation, but multiple incoming relations were allowed (e.g., a markable cannot be an anaphor of both identity and near-identity antecedent, but it can be the antecedent for a bridging and an identity anaphor). The aforementioned principles regarding the types and the size of markables hold for bridging and near-identity markables as well.

Cataphoric bridging and near-identity relations (directed from left to right) are allowed if the cataphoric antecedent is semantically closer to the anaphor than the possible anaphoric antecedent. For example:

(63) Ich kam [ins Büro] und nahm [den Hörer]$_{B1}$ ab. [Das Telefon]$_{B1}$ hat nicht funktioniert.
I came into [the office] and took up [the receiver]$_{B1}$. [The telephone]$_{B1}$ did not work.

In this example, the possible anaphoric antecedent *ins Büro* ('into the office') is less semantically related to the markable *den Hörer* ('the receiver') than the cataphoric antecedent, so we establish the relation from left to right and mark *das Telefon* ('the telephone') as its antecedent. Cataphoric antecedents can only be found in the same or in the next sentence, but not further away in the text. Consider the following example:

(64) Ich kam [ins Büro]$_{B1}$ und nahm [den Hörer]$_{B1}$ ab. Meine Kollegin hat angerufen.
< ... > [Das Telefon] hat nicht funktioniert.
I came into [the office]$_{B1}$ and took up [the receiver]$_{B1}$. My colleague called.
< ... > [The telephone] did not work.

In this case, we may want to link *den Hörer* ('the receiver') und *ins Büro* ('into the office'), but we do not link any of them to *das Telefon* ('the telephone').

Furthermore, we postulate several principles to resolve controversial issues:

A. The principle of semantic relatedness: In the case of multiple candidates, one has to pick the candidate that is semantically most closely related to the anaphoric (or cataphoric) markable.

(65) [[Das Telefon]$_{B1}$]$_{B2}$ klingelte. Ich kam [ins Büro]$_{B2}$ und nahm [den Hörer]$_{B1}$ ab.
[[The telephone]$_{B1}$]$_{B2}$ rang. I came into [the office]$_{B2}$ and took up [the receiver]$_{B1}$.

In this case, we link *das Telefon* ('the telephone') to *ins Büro* ('into the office') and *den Hörer* ('the receiver') to *das Telefon* ('the telephone').

B. The principle of primacy: In case of multiple possible relations, one has to prefer identity over near-identity and near-identity over bridging. When in doubt, establish an identity relation rather than near-identity, and a near-identity relation rather than bridging. In other words, the hierarchy is as follows:
Identity ← Near-Identity ← Bridging

(66) Last night in Tel Aviv, Jews attacked a restaurant that employs Palestinians, "[we]$_1$ want the war", [the crowd]$_1$ chanted[11].

[11]This example is taken from the work of Recasens et al. (2010a).

In this example, both near-identity and bridging are possible; however, according to the principle of primacy, near-identity is to be chosen.

C. The proximity principle: One always has to link a markable to its closest antecedent rather than establish a new relation, in accordance with the principle of primacy. For example[12]:

(67) Today [the right knee]$_{B1}$ is markedly swollen and there is a deformity overlying [[the patella]$_1$]$_{B1}$. [The patella]$_1$ appears to be high riding at this time.

In the first sentence, markable *the patella* is bridged to *the right knee*. However, in the second case, we do not have to bridge *the patella* to *the right knee* again, for there is a preceding markable that *the patella* can be linked to with the identity relation according to the proximity principle.

4.2.5 Difference to other major annotation schemes (OntoNotes, RefLex)

In this subsection, we compare our guidelines for identity coreference to two major annotation efforts – to the OntoNotes (Hovy et al., 2006) and RefLex guidelines (Riester and Baumann, 2017) that are publicly available. In the following, we focus on several deviations that differentiate our annotation scheme from the previous work, and we also provide example annotations for each of the annotation schemes.

- Appositions
  a. OntoNotes: Appositions are linked with the head phrase by a separate relation (APPOS): *[Daisy]$_{HEAD}$, [a private detective]$_{ATTR}$, was working on a case in a faraway neighborhood.*[13]
  b. RefLex: Appositions are included into the span of the head phrase: *[Daisy, a private detective]$_1$, was working on a case in a faraway neighborhood.*
  c. Our guidelines: similar to (b).
- Generic NPs
  a. OntoNotes: Generic expressions can only serve as antecedents to pronouns or definite mentions with the same referent: *[Computers]$_1$ are expensive. But [they]$_1$ are really useful. Computers cost a lot of money.*

[12]The example is taken from the corpus developed as part of this work.

[13]The example is taken from the corpus developed as part of this work.

b. RefLex: Generic expressions can form coreference chains: *[Computers]*$_1$ *are expensive. But [they]*$_1$ *are really useful. [Computers]*$_1$ *cost a lot of money.*

c. Our guidelines: similar to (a).

- Nominal premodifiers

a. OntoNotes: Nominal premodifiers that are represented by proper nouns are annotated as separate markables: *[the [FBI] spokesman].* The only exception is nationality acronyms (such as *U.K.*) which are not annotated.

b. RefLex: Nominal premodifiers are not annotated as separate markables: *[the FBI spokesman].*

c. Our guidelines: similar to (a), except that *all* nominal premodifiers (including nationality acronyms and independent nouns in the genitive form) are annotated as separate markables (see Section 4.2.2 for details).

- Prepositions

a. OntoNotes: Prepositions are not included into the markable span: *I have never been to [France].*

b. RefLex: Prepositions are included into the markable span: *I have never been [to France].*

c. Our guidelines: similar to (a), with the only exception being German prepositions if they are contracted with an article, such as *[zum Bahnhof].*

- Abstract anaphora

a. OntoNotes: Abstract anaphors corefer with the verbal head of the abstract antecedent: *Sales of passenger cars [grew]*$_1$ *22%. [The strong growth]*$_1$ *followed year-to-year increases.*[14]

b. RefLex: Abstract anaphors can have a verb phrase or a full clause as their antecedent: *[Sales of passenger cars grew 22%]. [The strong growth]*$_{r-given}$ *followed year-to-year increases.*

c. Our guidelines: Abstract anaphors are not annotated.

[14]The example is taken from the OntoNotes guidelines (BBN-Technologies, 2006).

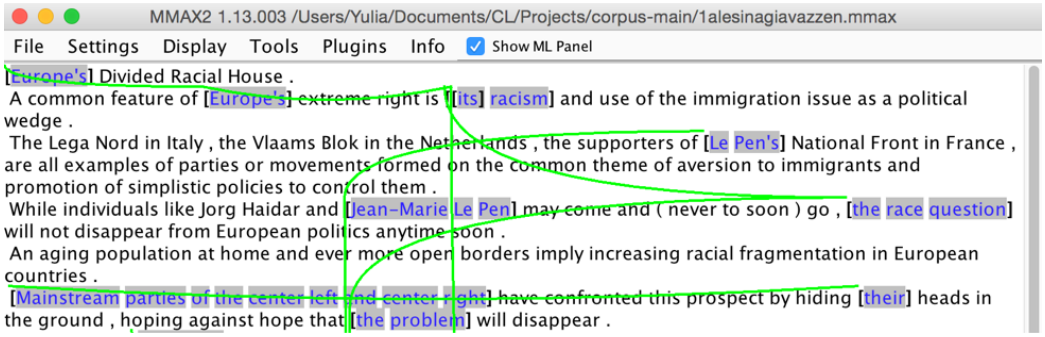


Figure 4.1: Sample annotation of coreference chains in MMAX-2

In sum, we have demonstrated that our guidelines are generally compatible to the recent large-scale annotation efforts – RefLex and OntoNotes – having slightly more overlap with the latter. However, our guidelines still contain a couple of deviations as compared to both schemes.

4.2.6 Technical details

For the annotation, we used publicly available MMAX-2 annotation tool[15] (Müller and Strube, 2001). The user interface of MMAX-2 allows for selecting markable spans and pairwise connecting anaphors and their antecedents, which is illustrated in Fig. 4.1: the highlighted markable spans are linked together if they co-refer, such as *Europe's* and *its*. Moreover, MMAX-2 allows for selecting the so-called *group markables* (if an anaphoric pronoun has several referents simultaneously), such as:

(68) Did [your husband]$_1$ buy Lorna, [Mrs. Humphries]$_2$? - No, [we]$_{1+2}$ bought her together.[16]

In this case, it is possible to create a group markable consisting of the set elements (*your husband, Mrs. Humphries*) and then link the anaphoric pronoun *they* to it.

Additionally, the annotators were offered to select a set of attributes described above (see Fig. 4.2). The annotations are subsequently saved in a stand-off XML format provided by MMAX-2, which can be used for querying and retrieving necessary information.

Furthermore, we converted the corpus into the CoNLL-2012 format, which is a standard format typically used in the NLP community for operating with different

[15] http://mmax2.sourceforge.net [accessed on 21.09.2017]

[16] The example is taken from the corpus developed as part of this work.

Figure 4.2: Sample annotation of markable properties in MMAX-2

types of linguistic annotations. In this format, text tokens and linguistic annotations are saved in a tab-separated file, with each token being on a separate line and sentence boundaries represented by blank lines. The conversion of the corpus is necessary for the subsequent experiments with annotation projection and, in particular, for a convenient representation of different layers of linguistic annotation.

To convert our corpus, we used the publicly available discoursegraphs library developed by Neumann (2015). The resulting files contained tab-separated columns with token IDs, tokens, and coreference chain IDs. At this stage, all the other fields were left blank, and the necessary linguistic annotations (e.g., POS information, syntactic trees) were added to the corresponding columns later.

4.3 Inter-annotator agreement

In this subsection, we describe the inter-annotator agreement metrics and the process of computing the inter-annotator agreement scores. Specifically, we report on the

inter-annotator agreement scores for identity coreference and bridging. For near-identity, we do not report on inter-annotator agreement due to the low number of links found in the corpus as already mentioned in Section 4.2.1.

The inter-annotator agreement scores for corpus annotation are typically computed using statistical measures which build upon the joint probability of agreement[17] corrected for chance agreement. One of the most used metrics is Cohen's κ, which is defined by the following formula (Cohen, 1960):

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \tag{4.1}$$

where P_0 is the relative observed agreement among raters, and P_e is the hypothetical probability of chance agreement. It should be noted that Cohen's κ is designed to measure agreement between two raters; to measure agreement among more than two raters, other κ measures can be applied[18] (such as e.g., Fleiss κ (Fleiss, 1971)).

Texts in the English-German part of the corpus were partially double-annotated with identity coreference: by the author of this thesis and two independent annotators (one annotator per language). For Russian, we only had one annotator available (the author of this thesis) and therefore were not able to compute the agreement scores. The independent annotators were students of linguistics, who had some previous experience in corpus annotation, but not in the annotation of coreference. Before starting the annotation process, the annotators received preliminary training on the annotation task: Having studied the guidelines, they had to complete test annotations on a corpus sample. The annotation guidelines were developed on 7 documents, and 4 of them were given to the annotators for training. The understanding of the guidelines and the quality of the resulting annotations were ensured by weekly meetings with the supervisor (author of the thesis) and discussions over the controversial issues. During the pilot annotation round, the annotators discussed the disagreements, and necessary changes to the guidelines were made.

In the following, we describe the metrics used to compute the inter-annotator agreement for different relations and report the agreement scores. We start by computing agreement scores for identity coreference for both languages. Thereafter, we compute agreement for bridging, but only for German texts. This choice was motivated by two reasons: (a) the agreement numbers for identity coreference did not vary

[17]Measured as the number of times each rating is assigned by each rater divided by the total number of ratings.

[18]Since we only have two annotators, we are not interested in these measures in the scope of this work.

	EN	DE
A1	1439	1446
A2	1350	1395

Table 4.2: Number of markables annotated by the first (A1) and the second (A2) annotator

considerably for English and German, as the identity coreference annotation round showed, therefore we decided to concentrate on one language, and (b) the fact that bridging is a more complex category than identity, thus we wanted to have native speakers as both annotators (and we did not have English native speakers available).

As already mentioned in the beginning of this section, the annotations of our corpus exhibited a small number of near-identity markables, which was not sufficient to compute inter-annotator agreement. For this reason, we merged the annotations from the first and the second annotator and then analyzed their distribution according to the near-identity types in Section 4.4.

4.3.1 Inter-annotator agreement for identity coreference

For identity coreference, we were interested in assessing inter-annotator agreement for the following annotation subtasks: (a) the identification of markables, (b) the assignment of coreference relations, and (c) the selection of markable attributes. Below, we give details on computing agreement scores for each of the subtasks.

First, we focused on the inter-annotator agreement for the identification of markable spans, since our goal was to see how well our guidelines were suited for a manual identification of markable borders (that, as opposed to some of the related work, were not present in the corpus data and had to be marked by the annotators). To achieve this, we scored the agreement using Cohen's κ (see Eq. 4.1), following the approach of Sidarenka (2016), who had a similar task of computing inter-annotator agreement (IAA) for the markable spans in a sentiment annotation. In particular, we computed the observed agreement as the ratio of tokens with matching annotations to the total number of tokens, and we scored the chance agreement based on the proportions of tokens annotated by the first and the second annotator. Statistics for the overall number of markables annotated by the first (A1) and the second (A2) annotator are presented in Table 4.2.

However, similarly to Sidarenka (2016), we observed the following two issues when computing the agreement on markable spans: (a) whether tokens need to be counted several times if they belong to several overlapping markable spans and (b) how many tokens should two markables have in common in order to be considered identical. Therefore, in order to account for these issues and following Sidarenka (2016), we computed the agreement on the identification of markable spans in two different settings:

a. binary overlap, considering two markables as "agreed" if they overlap by at least one token;

b. proportional overlap: measuring the extent to which annotators agree on the identification of spans (number of overlapping tokens).

In this manner, we were able to measure the chance-corrected percentage of every overlapping match as well as the percentage of tokens in overlapping markables. The results for both binary overlap and proportional overlap are shown in Table 4.3. As one can see from the table, the results for the identification of markable spans are quite reliable: We were able to achieve a $\kappa = 0.82/0.81$ for English and German respectively in the binary overlap setting, and a $\kappa = 0.74$ for both in the proportional overlap setting. Furthermore, we did not notice any considerable differences across languages, hence our annotation scheme is stable for both English and German.

Secondly, for the coreference annotation, similar to OntoNotes, we chose MUC score as a measure to compute the agreement, which is one of the metrics typically computed for the evaluation of coreference resolution systems (for more details, see Sect. 3.1.4). As opposed to other scores – specifically, B-cubed and CEAF, – this metric is highly representative for our purposes, since it focuses on a pairwise evaluation of coreference links, thus counting the minimum number of links between mentions to be added or deleted when mapping system results to a gold standard. We computed MUC score with strict mention matching, as we already evaluated the identification of mentions in the first step.

The results for the agreement on coreference chains are also shown in Table 4.3. Similar to the previous computation, the results for the annotation of coreference chains exhibit reliable agreement numbers: F1 of 73.64 for English and F1 of 71.96 for German. Importantly, we again obtain reliable results for both languages that do not vary considerably.

	EN	DE
Binary overlap κ	0.82	0.81
Proportional overlap κ	0.74	0.74
MUC F-score	73.64	71.96

Table 4.3: Inter-annotator agreement for identity coreference annotations

Finally, for the markable attributes, we computed Cohen's κ for the texts from all genres. The average agreement on markable attributes is $\kappa = 0.94$, which we also consider as reliable.

Analysis of the most frequent disagreements between annotators exhibited a number of cases that allowed multiple interpretations in terms of coreference relations and therefore posed difficulties for the annotators. Firstly, in the newswire texts, most contradictions occurred due to a number of borderline cases for which the referent could not easily be determined, such as collective nouns. For instance:

(69) On orders from Serbia's government, *Kosovo Serbs*, who represent some 5% of the population, refuse to cooperate with Kosovo's government and the EU mission. In doing so – and this is the irony of the matter – *Serbs* themselves are preventing the early implementation of the wide-ranging community rights foreseen in the Ahtisaari Plan, <...> Only a unified EU position, combined with the knowledge that EU accession for Serbia is unthinkable as long as this conflict has not been fully resolved, may over time lead to a change of attitude on the part of both *ordinary Serbs* and their government.[19]

In this example, the possible markable candidates for a coreference chain are *Kosovo Serbs, Serbs* and *ordinary Serbs*. Although these nominal phrases are indefinite, they are still among the central entities in the text and therefore could not be left unnoticed by the annotators. Here, two questions arose: (a) whether these nominal phrases have a real-world referent and as a result can be treated as definite, and (b) whether these mentions actually refer to the same entity (e.g., whether *Kosovo Serbs* represent the same set or simply a subset of *Serbs*, or whether *ordinary Serbs* have the same meaning as just *Serbs*). As we have seen from the subsequent annotation rounds, these cases represented near-identity and bridging relations and had to be annotated as such.

[19]Examples (69)-(71) are taken from the corpus developed as part of this work.

Another type of nominal phrases that were difficult to annotate were those referring to processes or events. Similarly, the identity of referents for those noun phrases was not completely clear to the annotators. For example:

(70) *Interethnic violence* – which many feared – has largely been avoided, and the mass exodus of Serbs that some also predicted has not occurred. <...> Unfortunately, however, we still cannot turn the page on *this pernicious conflict*, which has led to so much tragedy and has been a cause of instability in the Balkans for far too long.

In this case, *interethnic violence* was linked as identical to *this pernicious conflict* by one of the annotators; however, the identity of the referents for these two mentions cannot be clearly determined.

Secondly, in the short stories, the problematic cases were those that included a change of perspective, that occurred in dialogue speech. For example:

(71) You bought [Lorna]$_1$ in India? - Yes indeed! And [she]$_1$ always keeps me great company, you know. - With horror Daisy saw [a wiggling creature]$_2$ come out of that bag. Mrs. Humphries. [This]$_2$ is Lorna?

In this example, both mentions in both coreference chains have the same referent, but they are referred to from different perspectives: While the first speaker is familiar with the referent, for the second speaker the actual referent is not known and is therefore referred to by an indefinite description (*a wiggling creature*). In this case, since an indefinite description cannot have an antecedent according to the guidelines, we decided to establish two coreference chains.

Overall, after discussing the inter-annotator disagreements, we added such cases to the corresponding parts of the guidelines, also providing detailed instructions on how to annotate them in parallel texts. Collecting difficult issues found in the texts during subsequent annotation rounds and describing the strategies on how to deal with these issues helped to facilitate the annotation process as well as enrich our annotation guidelines.

4.3.2 Inter-annotator agreement for bridging

The annotation of bridging relations was made over the already existing identity coreference annotations. Since the inter-annotator agreement on markable selection was already measured during the identity annotation round, this was not computed separately for the bridging relation. Therefore, to facilitate the annotation process,

Relation	A1: #	A1: %	A2: #	A2: %
Part-Whole	20	9.09	18	8.57
Set-Membership	2	0.92	19	9.05
Entity-Attr/F	146	66.36	109	51.91
Event-Attr	20	9.09	29	13.81
Location-Attr	29	13.18	33	15.71
Other	3	1.36	2	0.95

Table 4.4: General distribution of bridging relations for the first (A1) and the second (A2) annotators

the rest of the markables in the texts were manually pre-selected by the author of this thesis. We computed the inter-annotator agreement for the following cases:

(a) anaphor recognition: the number of bridging anaphors both annotators agreed upon;

(b) antecedent selection: the number of anaphor-antecedent pairs both annotators agreed upon;

(c) individual bridging categories for those pairs that both annotators agreed upon.

For both anaphor recognition (a) and antecedent selection (b), we measured the standard F1 score that takes into account Precision and Recall scores. For the individual bridging categories (c), we computed Cohen's κ in order to measure the chance-corrected agreement on the assignment of bridging categories. The agreement scores were measured solely on the anaphor-antecedent pairs both annotators agreed upon.

Inter-annotator agreement was measured on 5 documents, with the first annotator marking 220 and the second annotator marking 210 pairs as bridging. Table 4.4 shows the distribution of the types of relations for the first (A1) and the second annotator (A2). As one can see from this table, the most frequent relation marked by both annotators was Entity-Attribute/Function, followed by Location-Attribute and Event-Attribute.

Table 4.5 shows agreement results for the cases (a) - (c) described above. We consider these scores as overall reliable for bridging when compared to related work

F1 anaphor recognition	64.0
F1 antecedent selection	79.0
κ Part-Whole	1.0
κ Set-Membership	N/A
κ Entity-Attr/F	0.97
κ Event-Attr	0.96
κ Location-Attr	1.0

Table 4.5: Inter-annotator agreement for bridging

on extended coreference. We were able to achieve even higher agreement scores on bridging categories (average $\kappa = 0.98$), introducing a wider range of relations than Nedoluzhko et al. (2009) (for details regarding this study see 3.1.2). We do not give an agreement score for set-membership, the reason for that being data scarcity and the preference of the first annotator towards other relations: The first annotator marked only about 0.1% of all bridging pairs as set-membership, and did not agree on antecedent selection with the second annotator for any of them, therefore it was not possible to measure agreement for this category.[20]

Table 4.6 shows the distribution of types for those pairs that were labeled differently by both annotators. The most controversial category is Entity-Attribute/Function, which correlates with this category being the most frequent one; the other types are almost equally disagreed upon. Particularly interesting is that only 3% of all the different bridging pairs are marked as near-identity pairs by the other annotator; accordingly, these categories in general do not intersect.

4.4 Corpus analysis

After computing the inter-annotator agreement, the annotations for identity, near-identity, and bridging created by the author of this thesis were selected for the final version of the corpus, and the rest of the corpus was subsequently annotated by the same annotator. Similar to the procedure described above, we first annotated identity

[20]One of the possible reasons for the low number of set-membership pairs could be the fact that our scheme for identity coreference includes discontinuous group markables, therefore some of the potential set-membership pairs were marked as identity coreference links (see 4.2.6 for details).

Relation	#	%
Part-Whole	32	14.95
Set-Membership	21	9.81
Entity-Attr/F	95	44.39
Event-Attr	30	14.03
Location-Attr	32	14.95
Other	4	1.87

Table 4.6: Distribution of inter-annotator disagreements for bridging

	Total		
	En	De	Ru
Tokens	11908	11912	8106
Sentences	589	598	431
REs	1350	1395	1085
Chains	259	273	188

Table 4.7: Statistics of the final version of the corpus for identity coreference

coreference, and thereafter, we added bridging and near-identity links. These annotations formed the final parallel corpus that was subsequently used for the annotation projection experiments and for the evaluation of the projection method.

In the following, we present the statistics of the final version of the corpus and analyze the distribution of different types of bridging and near-identity relations. Furthermore, we investigate the interplay between the identity, near-identity, and bridging annotations.

4.4.1 Corpus statistics

First, we computed corpus statistics for the identity coreference annotations. Statistics of the final version of the corpus for the three languages are presented in Table 4.7.

As one can see from the table, the German side of the corpus exhibits a slightly larger number of referring expressions and coreference chains in the corpus as com-

	Newswire			Stories			Medicine	
	En	De	Ru	En	De	Ru	En	De
Tokens	5903	6268	5763	2619	2642	2343	3386	3002
Sentences	239	252	239	190	186	192	160	160
REs	558	589	606	470	497	479	322	309
Chains	124	140	140	45	45	48	90	88
Av. chain length	4.5	4.2	4.3	10.4	11.0	10.0	3.6	3.5

Table 4.8: Statistics of the final version of the corpus for identity coreference across genres

	Newswire	Stories	Medicine
Named Entities	39.3	27.5	48.0
Personal pronouns	15.9	51.4	8.2
Definite NP	30.1	16.1	16.9
Relative pronouns	9.9	1.1	14.4
Indefinite NP	4.7	3.5	12.3
Other	0.1	0.4	0.2

Table 4.9: Distribution of markable types across genres (%)

pared to English (1395 vs. 1350 REs, 273 vs. 259 coreference chains respectively). Also, it should be noted that the numbers for Russian are lower than those for English and German due to the absence of medical texts in the corpus that were not available for this language.

Furthermore, we show the total number of markables and coreference chains for each of the genres in the corpus in Table 4.8. Looking at the distribution of markables across the genres, we can see that Russian exhibits the largest number of markables in the newswire texts, while German exhibits the largest number of markables in stories. As for the number of coreference chains, we see that these numbers do not differ considerably across genres, except for the newswire texts. Also, for all the languages, we notice that stories contain the longest coreference chains (with the average chain length of 10.5 markables), while the newswire texts contain the shortest chains (with

	# DE
Bridging pairs	432
Near-identity pairs	107

Table 4.10: Statistics of the final version of the corpus for bridging and near-identity (German)

the average chain length of 4.3 markables).

In addition, we present the distribution of markable types across different genres: Table 4.9 shows a breakdown of NP types of our markables for the three genres. Interestingly, newswire and medical texts exhibit the highest percentage of Named Entities (39.3% and 48.0% respectively); also, newswire texts contain a large proportion of definite noun phrases (30.1%), while medical leaflets exhibit the highest number of indefinite NPs (12.3%). Stories, contrastively, contain the largest number of personal pronouns (51.4%).

Second, we computed corpus statistics for bridging and near-identity pairs. In Table 4.10, we present statistics for the original German annotations, including the number of near-identity and bridging links. As one can see from the table, the final version of the corpus contained 432 bridging and 107 near-identity pairs.

Moreover, Table 4.11 shows the percentage of near-identity and bridging pairs across different genres. Interestingly, for bridging, all of the genres exhibit a high proportion of Entity-Attribute/Function relations. However, while in the newswire texts all of the relations are present, medicine leaflets and narratives lack some of the relation types, such as Event-Attribute (narratives) or Location-Attribute (both). In the narratives, we also encounter a lot more Part-Whole relations than in the other genres (37.14% vs. 9.77%, 16.66%).

As for near-identity, it is worth noticing that the annotations of medical texts exhibit a very high percentage (71.43%) of spatio-temporal relations, the reason for that being the specificity of the texts (instruction leaflets). In the narratives, we only find metonymic relations, while medical texts do not contain them. In the newswire texts, all types of relations are found, with meronymy being the most common one (76.32%).

Relation	News	Narrative	Med. leaflets
Part-Whole	9.77	37.14	16.66
Set-Membership	3.9	0.0	10.0
Entity-Attr/F	58.3	62.85	72.22
Event-Attr	12.08	0.0	1.12
Location-Attr	14.33	0.0	0.0
Other	1.62	0.0	0.0
Metonymy	15.79	100.0	0.0
Meronymy	76.32	0.0	28.57
Spatio-temporal func.	7.89	0.0	71.43
Other	0.0	0.0	0.0

Table 4.11: Distribution of bridging and near-identity relations across genres

4.4.2 The interplay between direct and indirect coreference annotations

In this subsection, we investigate the interplay between direct and indirect coreference annotations. In particular, we focused on the following issues:

(a) the correlation between bridging pairs and coreference chains;

(b) the interplay between bridging pairs and the most prominent coreference chains;

(c) the distance between bridging anaphors and antecedents.

Firstly, we were interested in investigating the correlation between bridging pairs and coreference chains in the corpus. To reach our goal, we first looked at the number of bridging anaphors that actually start a new coreference chain further in the text. On average for all the texts, only 17% of all the bridging anaphors are being referred to later on. These chains are on average 3.28 markables long, which is 1 markable shorter than the average length of coreference chains in the corpus (4.05). The most frequent relation that starts a new chain is Entity-Attribute/Function (44%), followed by Location-Attribute (21%), and Event-Attribute (18%).

Secondly, we were interested in whether bridging markables correlate with the prominent coreference chains in the text. Our study showed that 56% of all the

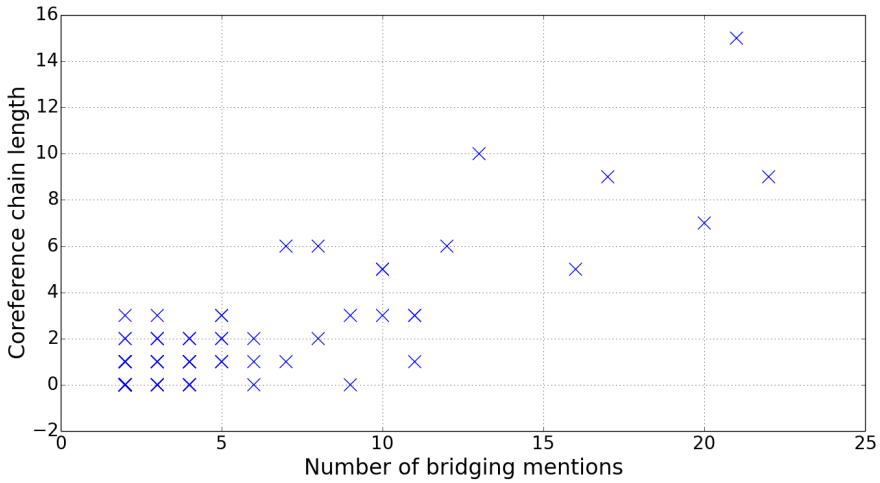


Figure 4.3: Length of identity chains and number of their bridging markables with Spearman's $\rho = 0.6595$

chains have bridging markables connected to them. We computed the average lengths of a target chain and a non-target chain for bridging, which is 6.1 markables and 2.4 markables, respectively. These numbers show that a target 'bridging' chain is usually longer than an average chain in the text (see above) while a 'non-bridging' chain is shorter. The longest 'bridging' chain can reach up to 22 markables, while the longest 'non-bridging' chain can only reach up to 9 markables.

We computed the correlation between the length of identity chain and the number of bridging markables that are linked to this chain. Using Spearman's rank correlation coefficient, we found that there is a strong correlation between the chain length and the number of its bridges: 0.6595, with p-value of 1.35E-008. Figure 4.3 shows the relation between the chain length and the number of its bridging markables.

Finally, we investigated the minimal and maximal distance between bridging anaphors and antecedents in different text genres. In general, our guidelines do not limit the scope of the study at any point, allowing annotators to bridge back over an unlimited number of sentences if they find the antecedent semantically close to the anaphor. However, it should be noted that we postulated several principles in order to set priorities and help annotators resolve controversial issues, one of them being the principle of SEMANTIC RELATEDNESS: In the case of multiple antecedent candidates, pick the one that is more semantically related to the anaphoric (or cataphoric) markable. This principle wins over the principle of PROXIMITY, according to which one has to bridge to the nearest semantically close antecedent in the text (see 4.2.4

for details and illustrations). With these principles in mind, we computed the average bridging distance (anaphora + cataphora), which is 20.55 tokens for all texts,[21] with the average sentence length being 24.87. The average distances for anaphora and cataphora, if computed separately, are 30.96 and -3.6 tokens, respectively. Also, our study has shown that distance does not seem to correlate with prominence: Both longer and shorter chains can have close and long-distance bridging anaphors.

4.5 Corpus delivery

The corpus was made publicly available online[22], distributed under a Creative Commons BYNC- SA 4.0 International License[23]. The annotations were distributed in the MMAX-2 format described above. Additionally, identity coreference annotations were distributed in the standard tab-separated CoNLL format, which is useful for system training and evaluation.

[21] We excluded bridging-contained markables from this computation.

[22] https://github.com/yuliagrishina/corefpro

[23] https://creativecommons.org/licenses/by-nc-sa/4.0/

Chapter 5

Corpus alignment

In this chapter, we will focus on the corpus alignment that is required for the subsequent annotation projection experiments. As already explained previously (e.g., see Chapter 3.2), annotation projection usually relies upon different types of alignments that enable automatic resource transfer.

Bitext alignment is defined as a step-by-step process of establishing correspondences between the source language and the target language. These correspondences could be established at different levels, for instance, paragraph, sentence, or word levels. Furthermore, different source units (e.g., sentences, words) can be aligned to one or many target units depending on the alignment type and algorithm. In particular, the following alignment scenarios are possible (for the illustration, see Fig. 5.1, 5.2):

a. one-to-null: one or more source units are aligned to the NULL unit;

b. one-to-one: one source unit is aligned to one target unit;

c. one-to-many: one source unit is aligned to many target units.

d. many-to-one: many source units are aligned to one target unit.

Typically, there are several steps that are performed to create a parallel aligned corpus from raw translation equivalents:

1. Corpus preprocessing: cleaning of the raw text data from unnecessary symbols, tokenization, etc. (see 5.1 for details).

2. Sentence alignment: establishing correspondences between source and target sentences that are translation equivalents of each other.

3. Word alignment: establishing correspondences between source and target words that are translation equivalents of each other.

In the following, we will explain each of these steps in more detail and describe the process of building our parallel corpus.

Previously published material

Some parts of Section 5.3, including the evaluation of the word alignment quality, have previously been published as (Grishina and Stede, 2015) and (Grishina and Stede, 2017).

5.1 Corpus preprocessing

As already stated in the previous section, raw text data first needs to be preprocessed before performing any kind of corpus alignment. The necessary preprocessing steps include cleaning the corpus and breaking text down into smaller units that could be aligned at a later stage. Overall, there are several steps that are typically performed for corpus preprocessing:

a. corpus cleaning: removing any malformed symbols, XML/HTML tags etc.;

b. sentence splitting: breaking down bigger discourse units into sentences;

c. tokenisation: splitting text into single tokens;

d. lower-casing: making all the uppercase letters lowercase.

The last step is usually done before automatic word alignment since the distinction between upper- and lowercase letters supports splitting text into sentences, but can hinder automatic word alignment.

Firstly, we thoroughly checked the corpus by using our script to detect malformed characters (that result from encoding incompatibilities, such as Russian quotation marks). To preprocess our corpus, we used EuroParl tools (Koehn, 2005) that were initially developed for the preprocessing of the EuroParl corpus and are therefore suitable for multiple languages. These tools include a sentence splitter, a tokenizer, and a sentence aligner. Specifically, we split raw texts into sentences and tokenized them using language-specific sets of non-breaking prefixes[1]. Since such a set was not available for Russian, we had to use one that is freely available with MOSES software (Koehn et al., 2007). The procedure was fully automatic, hence we did not perform any manual checks on the preprocessed data.

5.2 Sentence alignment

Sentence alignment is a task of computing the distances between the source and the target sentences and establishing the most probable correspondences between the sentences with minimal distance. Typically, the distance measure is based on the sentence length, which can be calculated as a number of tokens or a number of characters. The latter was proven to be a more successful approach that was used

[1]Abbreviations that should not be separated from their punctuation, such as *Mr.* or *St.*

by Gale and Church (1993) to develop their sentence alignment algorithm, which became standard for computing sentence alignments. In particular, it is based on the assumption that sentences that correspond to each other should also roughly correspond to each other in their length in terms of number of characters. In other words, sentences with a similar length are most likely to be translation equivalents of each other.

Based on this principle, Gale and Church (1993) implemented an algorithm to compute the best corresponding source and target sentences given a parallel corpus and based on a statistical model of sentence length in terms of number of characters. Specifically, they assumed that sentence lengths follow a normal distribution, and they used the following formula to estimate the length difference δ between a pair of sentences:

$$\delta = \frac{(l_t - c \cdot l_s)}{\sqrt{l_m \cdot s^2}} \tag{5.1}$$

where l_t and l_s are the lengths of the source and target sentences respectively, l_m is their average length, c is the scaling factor[2] and s^2 is the variance for all the values of δ in a typical aligned corpus[3]. This length difference metric is subsequently used by the authors to establish a conditional probabilistic model that is used to estimate sentence alignment likelihood. The final alignment is then found by choosing the alignment with the highest likelihood.

We used the HunAlign implementation of the Gale and Church algorithm (Varga et al., 2007) and its wrapper LF Aligner[4] to process the corpus. Apart from the Gale-Church algorithm, HunAlign additionally makes use of bilingual dictionaries for the languages in question. If no dictionary is available, it first runs the Gale and Church algorithm and then automatically creates a dictionary based on the output of the first alignment round, subsequently using the dictionary to realign the text in the second pass. Using this method, HunAlign is able to achieve a reasonably high performance: Varga et al. (2007) showed that it was able to achieve over 97 F1 score on an English-Hungarian dataset, as compared to manually produced gold alignments.

Fig. 5.1 illustrates one-to-one and one-to-many alignments in the English-German-Russian parallel corpus. While the first example is quite straightforward, illustrating a case where one German sentence has its exact equivalents in both languages, the second example is more complex: Three German sentences are translated as one

[2]The scaling factor is computed as the average ratio between the source and the target text sizes.

[3]Gale and Church (1993) provide an average variance of 6.8, which they use across languages.

[4]http://aligner.sourceforge.net [accessed on 31.07.2017]

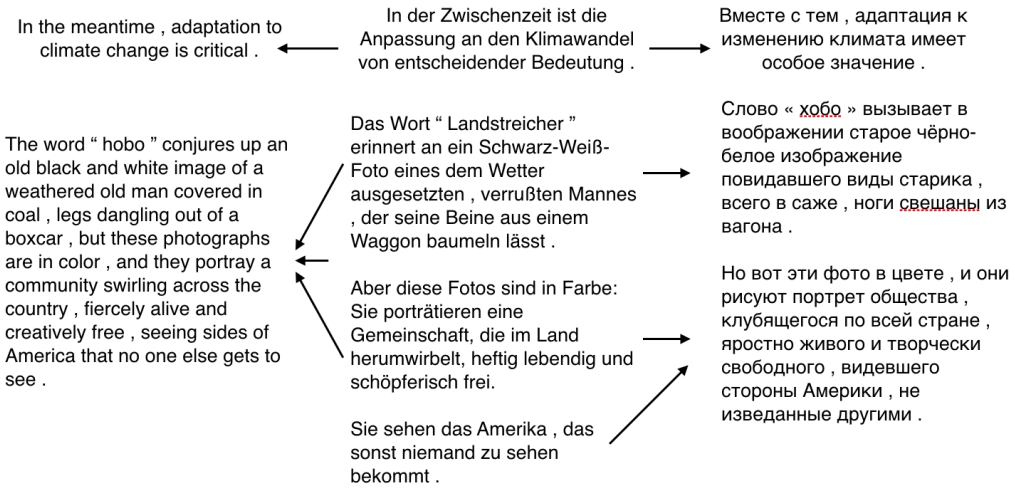


Figure 5.1: An example of one-to-one (top) and many-to-one (bottom) sentence alignment between English, German and Russian text

sentence in English and as two sentences in Russian. This example provides an illustration of translation divergence in a parallel corpus, where the translator's choice of splitting or combining sentences can influence the quality of the resulting alignments.

Since the bilingual dictionaries for our languages were available, we ran HunAlign for English-German, English-Russian and German-Russian language pairs. For the multi-source approach, we selected only those parallel sentences that were present in all the three languages which resulted in a slight reduction of the corpus size. In particular, this method reduced the average number of sentences per language by 5% and the average number of coreference chains per language by 6%.

5.3 Word alignment

Bitext word alignment is a task of identifying translation equivalents among words (and sometimes multi-word units) in parallel sentences. In other words, given a pair of sentences from an aligned bilingual corpus in the source (L1) and target (L2) languages, the goal of a word alignment system is to determine which word in the given sentence of language L1 corresponds to which word in the given sentence of language L2.

More formally, statistical word alignment can be defined as finding the best mapping between source and target words for pairs of sentences in a bilingual sentence-aligned corpus using language-independent statistical methods. Word alignments are a by-product of translation probabilities, which are estimated based on the initial assumption that any word in the source sentence can be aligned to any word in the target sentence and subsequently calculating the cooccurrences of the source and target words. For instance, if a certain source word frequently appears in aligned sentence pairs together with a certain target word, then they are likely to be translation equivalents of each other. To exemplify, given a large parallel corpus, one can notice that the German word *Katze* often appears in the same sentence pair as the English word *cat*, therefore they could be considered equivalent to each other. However, a considerable amount of training data is typically required to estimate these correspondences from a parallel corpus.

Words in parallel sentences are aligned based on statistical models. These models explain the relationship between a source language sentence f_1^J that contains tokens $f_1, ..., f_j, ..., f_J$ and a target language sentence e_1^I that contains tokens $e_1, ..., e_i, ..., e_I$:

$$Pr(f_1^J|e_1^I) = \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) \tag{5.2}$$

where a_1^J is the word alignment between words in the sentence pair in question[5] (it should be noted that the alignment may contain zero alignments representing cases when a source word is not aligned to any of the target words). Various alignment models differ in terms of the decomposition of $Pr(f_1^J|e_1^I)$ (Och and Ney, 2003); we will review some of them below.

An example of bidirectional alignment for English and German is presented in Fig. 5.2[6]: (i) shows alignment links from English to German, and (ii) shows alignment links from German to English. As one can see from this figure, some of the words are left unaligned (for example, *up, to, get* in (i)).

There are several issues that may arise while aligning our language pairs. The alignment between two sentences might include reorderings, omissions, insertions, or word-to-phrase alignments (Och and Ney, 2003), which pose difficulties for the alignment algorithms. For that reason, various alignment models were developed, to

[5]Although for a given sentence pair a large number of possible alignments may exist, the best alignment can always be found by taking the most probable predicted alignment (the so-called *Viterbi* alignment). In the rest of this thesis, we will only be using best predicted alignments.

[6]The example is taken from the corpus developed as part of this work.

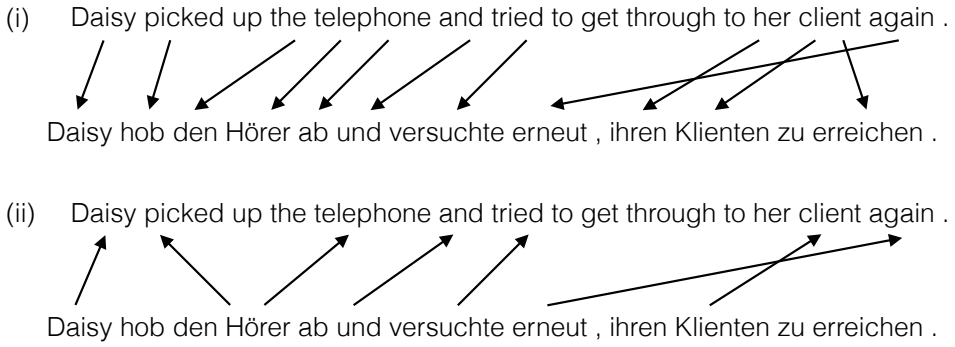


Figure 5.2: English-to-German and German-to-English alignments

overcome the mismatches between source and target sentences. In particular, these models differ in respect to the alignments they are able to model.

The IBM Models are among the most widely used alignment models which differ in their complexity. While IBM Models 1 and 2 rely on lexical translation and position of a token in a sentence respectively and can only model one-to-one alignments between source and target, the IBM Models 3-5 (Brown et al., 1993) also allow many-to-one alignments. These models take into account the parameter called *fertility*: Fertility of a target language word is defined as the number of source language words that can be aligned with it under a given alignment. Also, it is worth pointing out that the IBM Model 4 implements a relative reordering model, in which the alignment of a current target word depends on the alignment of the preceding target word.

Another approach widely used is the Hidden Markov Model (HMM) approach to word alignment (Vogel et al., 1996), which relies on the lexicon probability as well as the HMM alignment probability. While alignment probabilities for the former depend only on the alignment of the current word, those of the latter, on the contrary, are dependent on the alignment probabilities of the preceding word. This model in combination with the IBM Model 4 forms the IBM Model 6 (Och and Ney, 2003).

In our experiments, we used GIZA++ word alignment tool (Och and Ney, 2003) to align our data. It uses multiple alignment models, including the models described above, to compute alignment probabilities in a large parallel corpus. These alignment models provide only one-to-zero, one-to-one and many-to-one alignment links, but they do not allow a source word to be aligned with more than one target word (one-to-many alignments). Therefore, some target multi-token units cannot be properly aligned: For example, source German compound nouns (e.g., *Zugverspätung*) cannot

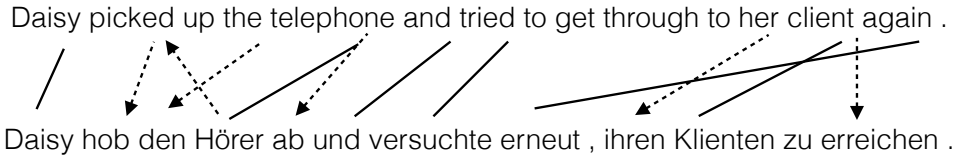


Figure 5.3: Intersection (solid links) and union (all links) of bidirectional alignments

be aligned to the corresponding English noun phrases that contain several words (*train delay*). To overcome this limitation, alignment in both directions (source-to-target and target-to-source) can be performed and, subsequently, several symmetrization methods can be exploited, such as (Och and Ney, 2003):

- `union`: all alignments that are present in source-to-target and target-to-source alignments are selected;
- `intersection`: only those alignments that are present in both source-to-target and target-to-source are selected;
- `grow`: firstly, the alignments from the intersection step are added. In the next step, different neighboring alignment points from the union are added, depending on the grow method (such as `grow-diag, grow-final`, etc. (Wu and Wang, 2007)).

The first two methods (union and intersection) are illustrated in Fig. 5.3: Intersective alignments are those presented by solid lines, and all the links show the union of alignments. These alignments result from Fig. 5.2 by applying the symmetrization methods described above.

Annotation projection often can be done using the intersection of the source and the target alignments, since it provides a higher alignment quality in terms of Precision (Spreyer, 2011); however, in order to achieve higher Recall, all the alignments need to be used. In the following chapters, we will describe 3 experiments on annotation projection that use different types of word alignments depending on the projection task. Thus, in our first experiment (Chapter 6), we used a single-source projection method and computed both bidirectional alignments and the intersection of the source-target and target-source alignments, using the latter to transfer the annotations with maximal Precision. However, for the second and the third experiments (Chapter 7 and 8), we were not interested in obtaining higher Precision scores

at the cost of lower Recall, therefore we only computed standard alignments and used a dependency parser to extract the target mentions and overcome the noise in the resulting annotations coming from the low-quality alignments. In sum, we performed word alignment at two stages in the following directions:

(a) intersection: English-German ∩ German-English, English-Russian ∩ Russian-English;

(b) bidirectional alignments: English-German, English-Russian, German-Russian.

Since the texts in the training parallel corpus should be of the same domain, we used two training sets to align our corpora. In our case, we had access to two large collections of parallel texts – newswire texts and medical instruction leaflets, that we chose as training sets to align news and stories, and medical texts respectively[7]. Therefore, our word alignment was performed in two rounds: (a) for the newswire texts and stories and (b) for the medical texts. As for the first training set, we used a collection of bilingual newswire texts taken from the Project Syndicate website. As for the second one, we relied upon the complete EMEA collection of texts taken from the OPUS website[8].

Overall, the training set consists of the following amount of parallel sentences:

- English-German: 232 179 parallel sentences (Project Syndicate texts), 1 108 752 parallel sentences (EMEA texts);
- English-Russian: 156 819 parallel sentences (Project Syndicate texts);
- German-Russian: 230 261 parallel sentences (Project Syndicate texts).

Evaluation of automatic word alignment is an important step which helps to estimate the quality of annotation projection to a certain extent, which, however, is a challenging task, since manually annotated test sets are rarely available. Therefore, in order to estimate the coverage of our word alignments, we computed the percentage of unaligned target words for each of the language pairs. The results for all the alignment directions are presented in Table 5.1. As one can see from this table, the English-Russian alignments exhibit the smallest percentage of unaligned

[7]Since we did not have access to any large parallel collection of stories, we decided to align them together with newswire texts (rather than with medical texts that contain domain-specific vocabulary).

[8]The description of sources is presented in Chapter 4.1.

	English	German	Russian
English	x	17.84	14.96
German	15.08	x	17.03
Russian	18.24	21.01	x

Table 5.1: Percentage of unaligned words (%)

	Bisentences	P	R	F1
Padó (2007)	1 029 400	98.6	52.9	68.86
Spreyer (2011)	1 314 944	94.88	62.04	75.02
Our alignment	205 208	92.95	51.23	66.05

Table 5.2: Evaluation of the automatic word alignment

words (14.96%), while the Russian-German alignments contain most of the empty alignments (21.01%). Obviously, this difference in the alignment coverage does not necessarily indicate the lower alignment quality (since some language units such as articles can simply not be present in the other language and are therefore left unaligned) but provides some insight for further investigation. In particular, in the next chapters, we will examine the types of unaligned units in order to evaluate the quality of the annotation projection methods.

Furthermore, for English-German, we evaluated our word alignment quality using a set of manually annotated parallel sentences made available by Padó (2007)[9]. This dataset consists of 1000 sentences, annotated according to the Blinker project guidelines (Melamed, 1998), and is publicly available.

The results of the quality of automatic word alignments as compared to the gold dataset are given in Table 5.2. Following Padó (2007), we evaluated only the resulting intersective alignments. We compared our results to the similar evaluations of Padó (2007) and Spreyer (2011), who used the English-German part of the EuroParl dataset. As one can see from the table, our results are slightly lower, the reason probably being a much smaller training set. Since our aim was to experiment in a low-resource scenario for both languages in the same setting, we deliberately did not include any additional parallel data for English-German. As for the other lan-

[9]http://nlpado.de/ sebastian/data/srl_data.shtml [accessed on 31.07.2017]

guage pairs (English-Russian, German-Russian), we are not aware of any similar gold alignments and thus did not evaluate.

Chapter 6

Single-source projection of coreference chains

In the following chapters, we report on three experiments on projecting coreference chains in different settings. All the experiments below are based on the direct projection method, but differ in the types of alignments and projection strategies used. Furthermore, each of the experiments is performed in two settings: a knowledge-lean setting (relying only on automatic word alignments) and a more linguistically informed one (using the output of a syntactic parser to identify markable borders and therefore improve the quality of the mention identification).

Thus, the first experiment describes a study of projecting coreference chains using only one source, resembling the work of Postolache et al. (2006). In this experiment, we use a 'classical' projection direction (from English to other languages) and intersective word alignments to maximize the projection quality. In particular, our goal is to compare how well the projection algorithm works for two relatively similar languages (English-German) and for less similar languages (English-Russian), and we are also interested in differences incurred by the text genre.

Thereafter, the second experiment (see Chapter 7) presents a broader scenario: multi-source annotation projection using *all* the alignments and additional projection directions, such as German-Russian and Russian-German. In this experiment, we implement several multi-source projection strategies based on the concatenation and intersection of the projected mentions and compare them to each other.

In the third experiment (see Chapter 8), we adopt a fully automatic pipeline and use automatic source annotations produced by two state-of-the-art coreference

systems. We combine the output of our projection method for two source languages (English and German) to obtain target annotations for a third language (Russian), and we compare these results to the projection of manual coreference annotations. At the end of each chapter, we evaluate the approaches and discuss the results.

In this part of the work, our general aim is to explore the limitations of a knowledge-lean approach to the problem, so that it is easy to generalize to other low-resourced languages, as well as to investigate whether introducing limited linguistic information can be beneficial for our approach. Thus, at this point, we implement two projection settings and discuss the benefits and drawbacks of each of the approaches.

Previously published material

The experiment presented in this chapter was previously published in a more compact version as (Grishina and Stede, 2015).

	Newswire			Stories			Medicine		Total		
	En	De	Ru	En	De	Ru	En	De	En	De	Ru
Tokens	5903	6268	5763	2619	2642	2343	3386	3002	11908	11912	8106
Sentences	239	252	239	190	186	192	160	160	589	598	431
REs	558	589	606	470	497	479	322	309	1350	1395	1085
Chains	124	140	140	45	45	48	90	88	259	273	188

Table 6.1: Statistics for the annotated gold corpus

6.1 Experimental setup

In this section, we describe our implementation of the direct projection method for transferring coreference chains and then give details on the two settings that we use in our experiments. We use our parallel sentence- and word-aligned corpus (described in Chapter 4 and 5) to transfer manual English annotations to German and to Russian, and we use the gold standard corpus for the evaluation of our method. Table 6.1 (as also already shown in 4.4) presents statistics of the gold corpus used in this experiment, showing the number of referring expressions and coreference chains for the three languages and the three genres.

6.1.1 Projection method

Following the approach of Postolache et al. (2006), we define the experimental setup for our annotation projection experiments. The experimental setup consists of the following stages:

1. **Automatic alignment:** We perform sentence and word alignment of the corpus as described in Chapter 5;

2. **Extraction of referring expressions (REs):** For each annotated RE in the source language we extract the corresponding RE in the target language. Specifically, for each word span representing an RE in the source language, we extract the corresponding set of aligned words in the target language. The resulting target RE is the span between the first and the last extracted word. The resulting target mentions are assigned the same coreference chain IDs as the source mentions.

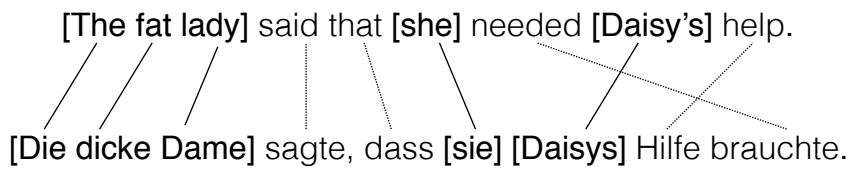


Figure 6.1: Example of automatic annotation transfer from English to German using word alignments

Fig. 6.1[1] illustrates the transfer of coreference annotations from English to German. Using a pair of parallel sentences, we first extract source annotated mentions *'the fat lady', 'she', 'Daisy's'*. Then, for each of the mentions, we extract the corresponding words that are aligned to these mentions: *'die, dicke, Dame', 'sie', 'Daisys'*. At this step, we extract the target mentions as the spans between the first and the last aligned word. Finally, we add these annotations to the target text with the same coreference chain numbers as the source mentions. The resulting target mentions are *'die dicke Dame', 'sie', 'Daisys'*.

6.1.2 Projection settings

As already mentioned previously, in the first setting, no additional linguistic information is used to support the projection method. In this setting, we use only bidirectional word alignments as computed by GIZA++ to transfer information from one language to the other. In contrast, in the second setting, there is a mention extractor for both German and Russian available to support the recovery of the projected coreference mentions on the source side. We define the projection settings as follows:

a) Setting 1: no additional linguistic resources are available, and we only rely upon the automatic word alignments to extract the mentions in the target side.

b) Setting 2: a mention extractor is available. In particular, we rely upon the output of the MATE dependency parser[2] (Bohnet, 2010) for German and the MALT dependency parser[3] (Nivre et al., 2006) for Russian. For the latter, we used the model provided by Sharoff and Nivre (2011). Thereafter, we automatically extracted all mentions from the target sides that have nouns, pronouns

[1]The example is taken from the corpus developed as part of this work.

[2]https://code.google.com/archive/p/mate-tools/ [accessed on 31.07.2017]

[3]http://www.maltparser.org [accessed on 31.07.2017]

or pronominal adverbs (for German) as their heads, and we subsequently map the output of the projection algorithm to the extracted mentions.

Regarding the mapping strategy, we consider the work of Rahman and Ng (2012), who also used a mention extractor to map target coreference annotations to the mentions automatically extracted by the Reconcile coreference resolution platform (Stoyanov et al., 2010). They suggest the following order of mapping a projected mention m_P to a mention from the automatically identified set of mentions M_R (Rahman and Ng, 2012):

1. mapping of m_P to a mention in M_R that shares the same right boundary;
2. if it fails, mapping of m_P to a mention in M_R that covers its entire text span;
3. if it still fails, mapping of m_P to a mention in M_R that has a partial overlap with it;
4. otherwise assume that m_P is not present in M_R and add it to the target annotations.

In our case, given the differences in syntactic taggers and hence automatic mention extraction, we map the markables in a slightly different way. Relying upon the steps (1) and (2) from the above, we add an extra step to our mapping heuristic and change the mapping order. In particular, we adopt the following strategy:

1. first, we map a projected mention to an extracted mention that is identical to it;
2. if it fails, we map a projected mention to an extracted mention that shares the same right boundary with it;
3. if it still fails, we map a projected mention to an extracted mention that spans this mention;
4. otherwise we assume that no corresponding mention is found and add the projected mention to the target annotations.

Once a target markable is mapped to some automatically extracted mention, we discard this mention, to ensure that it is not mapped to any other markable. Taking into account the differences between the syntactic taggers as well as the differences in NPs for both languages, we experimentally chose to skip step (2) for Russian annotations.

6.2 Evaluation

The evaluation of coreference resolution quality usually consists of two steps: the evaluation of the identification of mention spans and the evaluation of the linking of these mentions into coreference chains. Regarding the spans of the mentions, the most common strategies use strict mention matching (where the gold and the system RE spans should exactly correspond to each other) and minimal span matching (with minimal spans being, for example, syntactic heads of the referring expressions). The latter can be useful when gold and system annotations adopt different notions of mentions (e.g., whether a subordinate clause can be part of a mention or not) and therefore cannot be compared fairly in terms of identified mentions using the former strategy.

As for coreference chains, there are several standard coreference metrics that are typically used to score the results: MUC (Vilain et al., 1995), B-cubed (Bagga and Baldwin, 1998) and CEAF (Luo, 2005).[4] Since each of these metrics has its own advantages and disadvantages, during the CoNLL-2011 evaluation campaign, it was proposed to combine their unweighted average as the final evaluation score (CoNLL score) (Pradhan et al., 2011). Following this evaluation methodology, we also score the projected annotations as compared to the gold annotations using the three metrics as well as their unweighted average.

In the following, we perform the evaluation of our method in two steps. Firstly, we evaluate the identification of mentions and the quality of the projected coreference chains on the whole corpus using strict mention matching. At this stage, our goal is to evaluate the quality of annotation projection for the two languages and to compare the projection settings. In the second step, we evaluate the quality of the projected annotations for different text genres and explore the differences between them. Therefore, we compute the micro-average coreference scores for each of the genres separately, and we also use the minimal span strategy to score coreference chains. This evaluation method also allows us to fairly compare our results to the work of Postolache et al. (2006), who also additionally scored their method using the minimal span strategy.

	News		Stories		Med
	De	Ru	De	Ru	De
Transferred REs	465	493	329	357	214
Transferred coreference chains	122	122	44	44	82

Table 6.2: Number of REs and coreference chains transferred through bilingual projections

	P	R	F1
EN→DE	64.5	48.5	55.2
EN→RU	80.3	63.7	70.8
EN→DE+*ment*	71.8	54.4	61.7
EN→RU+*ment*	78.1	61.6	68.6

Table 6.3: Results for German and Russian: identification of mentions

6.2.1 Macro-averaged evaluation of the projection method

In this subsection, we evaluate the quality of the identification of mentions and of the projection of coreference chains for the full corpus. As the starting point, we compute the overall number of transferred mentions and coreference chains: Table 6.2 shows the number of REs and coreference chains projected through word alignment from English to German and Russian. This table gives us a rough estimation of how many REs and chains were transferred from the source language to our targets, and how many of them were lost due to the wrong alignments or mismatches between the source and the target texts. Interestingly, we see that for news and stories, the number of transferred chains is the same for both German and Russian, and there is only a small difference in the number of mentions projected.

Subsequently, we evaluate the quality of the identification of mentions and of the projection of coreference chains using the official CoNLL scorer[5]. We evaluate our algorithm using the gold annotations of German and Russian texts, and we compute the macro-averages scores for coreference chains in each of the settings. Table 6.3

[4]A detailed overview of coreference metrics is presented in Chapter 3.1.4.

[5]http://conll.cemantix.org/2012/software.html [accessed on 31.07.2017]

	MUC			B^3			$CEAF_m$			Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EN→DE	63.1	44.5	52.0	52.6	34.2	41.0	62.6	47.1	53.6	59.4	42.0	48.9
EN→RU	78.1	60.5	67.9	70.9	50.1	58.1	77.9	62.0	68.8	75.6	57.5	64.9
EN→DE+*ment*	68.1	48.5	56.5	58.0	38.5	45.8	68.5	51.9	58.9	64.9	46.3	53.7
EN→RU+*ment*	75.1	57.9	65.1	67.7	47.5	55.3	75.5	59.8	66.5	72.8	55.0	62.3

Table 6.4: Results for German and Russian: projection of coreference chains

presents the results for the identification of mentions, and the results for the coreference scores are given in Table 6.4.

Comparing the results for the projection of coreference chains across languages, one can see that the overall results for the English-Russian language pair are higher than the results for the English-German language pair in the knowledge-lean setting (64.9 F1, 48.9 F1 respectively). However, applying a mention extractor can considerably improve the results for English-German (from 48.9 F1 to 53.7 F1), but not for English-Russian. For this language pair, using the output of a syntactic parser leads to a drop in performance (from 64.9 to 62.3 F1). We assume that the reason for this drop in performance is the better quality of word alignment for Russian nominal phrases than for German, and the noise added by the Russian mention extractor, which relies on the automatic syntactic annotations. In section 6.3, we investigate this issue in more detail.

6.2.2 Micro-averaged evaluation according to the text genre

In order to study the differences in projection quality across the text genres, we compute the micro-averaged coreference scores for each of the genres separately. Moreover, to fairly compare our results to the most closely related work of Postolache et al. (2006), we compute the scores with both strict and minimal span matching techniques, with the latter considering only the matching of the syntactic heads of the referring expressions in question. The benefit of this strategy is that it indicates how well the REs can be projected, not punishing the algorithm for detecting only partially correct REs. To achieve this, we manually annotated syntactic heads of the gold and projected REs. Following the approach of Postolache et al. (2006), we select the leftmost noun, pronoun or numeral as head; otherwise, the RE is discarded.

	Mentions			MUC			CEAF			B^3			Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
de-News	61.5	48.6	54.3	55.9	43.2	48.7	58.6	46.7	51.9	45.8	34.2	39.1	53.4	41.4	46.6
de-Stories	82.0	54.5	65.5	81.9	51.6	63.3	81.7	53.7	64.8	71.6	32.5	44.7	78.4	45.9	**57.6**
de-Med	61.2	44.7	51.7	66.2	42.7	51.9	59.1	43.3	50.0	53.4	35.2	42.4	59.6	40.4	48.1
de-News$_{min}$	89.9	71.2	79.4	87.3	66.2	75.3	85.5	67.5	75.5	80.4	58.1	67.5	84.4	63.9	**72.8**
de-Stories$_{min}$	95.4	62.2	75.3	94.4	58.5	72.2	95.1	61.2	74.5	90.9	40.2	55.7	93.5	53.3	67.5
de-Med$_{min}$	79.9	58.4	67.5	84.2	54.4	66.1	77.7	56.9	65.7	73.3	47.2	57.4	78.4	52.8	63.1
ru-News	79.3	64.5	71.2	76.3	60.7	67.6	76.3	62.0	68.4	69.0	52.2	59.4	73.9	58.3	65.1
ru-Stories	87.4	65.1	74.6	87.9	64.4	74.3	86.1	64.6	73.8	79.7	47.9	59.8	84.6	59.0	**69.3**
ru-News$_{min}$	90.9	72.6	80.7	89.6	69.8	78.5	87.3	69.7	77.5	83.7	61.4	70.9	86.9	67.0	75.6
ru-Stories$_{min}$	94.3	72.4	81.9	94.0	70.9	80.9	93.6	71.7	81.2	90.2	57.3	70.1	92.6	66.6	**77.4**

Table 6.5: Results for German and Russian: projection of coreference chains for different genres

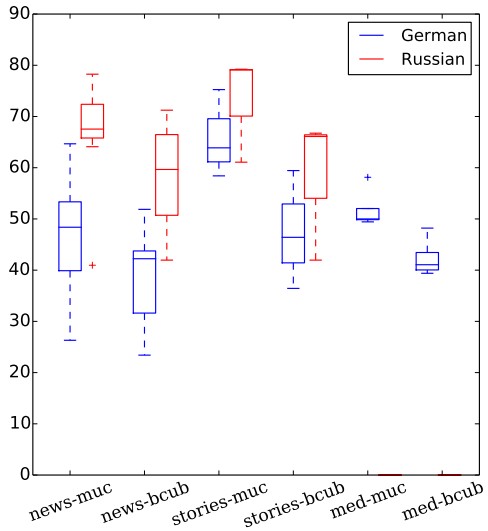


Figure 6.2: Comparison of English-German and English-Russian projections: boxplots of the macro-averaged F1 scores (MUC and B-cubed) for different genres

Table 6.5 presents the scores for the quality of the identification of mentions and projection of coreference chains with both the strict mention matching and the minimal span strategy (with the tag *'min'*). As one can see from this table, with strict mention matching, the projection method was able to achieve the highest scores of 57.6 F1 for German and 69.3 F1 for Russian. Both scores are observed within the narrative genre, which tends to be easier for the projection method as compared to the newswire and medical texts. Furthermore, we notice the same tendency when comparing the scores obtained using the minimum span strategy for Russian (the highest F1 of 77.4 for stories), but not for German (the highest F1 of 72.8 for news). It also should be noted that, using this strategy, we obtain P>90 for stories and P>80 for news in both languages, which shows that target mentions could successfully be found in in the projected annotations. The different results for the two settings show that a better mention extraction strategy is required. Interestingly, the overall results for Russian are higher for all genres as compared to the results for German; as for the difference across genres, stories seem to be the 'easiest' genre for the projection, which we attribute to their NP structure[6]. This is also illustrated in Fig. 6.2 which presents the boxplots of the macro-averaged F1 scores (MUC and B-cubed) for the two languages and for the three genres.

[6]See Section 6.3 for a more detailed analysis.

According to Table 6.5 and Figure 6.2, we see that newswire texts get the lowest scores, the reason most likely being the more complex NPs. In particular, as already shown in Chapter 4.4.1, newswire texts contain the highest percentage of definite NPs that are more difficult for the projection since they often consist of multiple tokens (which poses difficulties for the word alignment), while stories exhibit a much larger number of personal pronouns that are usually single-token mentions and are therefore easier for the projection (see Table 4.9 for details). In setting 2 (evaluation of minimal spans), both newswire texts and stories obtain closer F1-scores, but the stories still have better Precision scores. The medicine instruction leaflets in the minimal span setting have the worst results, and we observe lower improvement for Precision between two settings compared to the newswire texts. This indicates that the quality of coreference resolution for medical texts depends to a higher degree on the coreference relations, than on the identification of mentions. In these texts, we frequently find borderline cases of non-/reference, when diseases, parts of the body, etc. are being mentioned.

In the next subsection, we will look at the projection errors from the linguistic point of view and empirically classify difficult cases.

6.3 Error analysis

In this subsection, we will perform an error analysis of the projected annotations by investigating their false positives and false negatives. From a formal viewpoint, there are three categories of projection problems:

1. An RE is present in both source and target text, but it is not projected correctly, or not at all, on the grounds of mistakes in the word alignment phase.

2. An RE is present in the source text and correctly projected into the target text, but it does not show up in the gold standard, because the target language text does not have a corresponding RE *pair* for one in the source language.

3. An RE in the gold standard is not present in the target text and therefore cannot be projected (the dual problem to (2): the source text does not have an RE pair that would correspond to one in the target text).

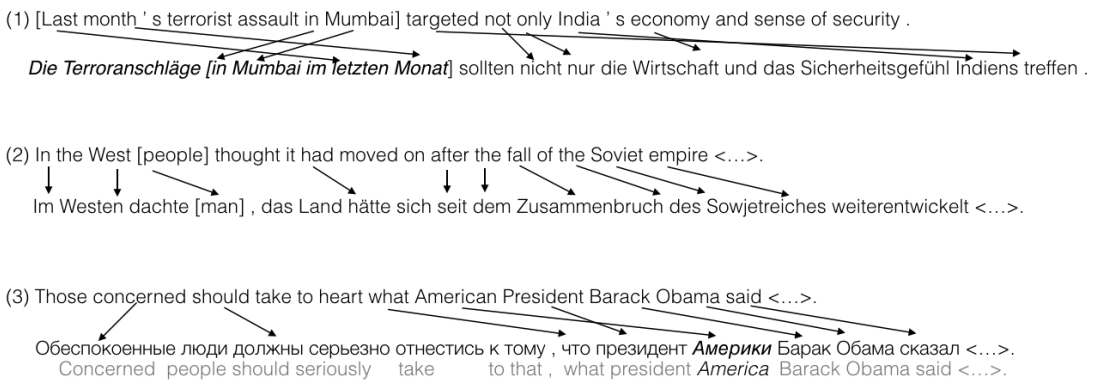


Figure 6.3: The most typical projection problems

These categories are illustrated in Fig. 6.3, where we provide examples for each of the cases[7]. In (1), the source markable *[last month's terrorist assault in Mumbai]* is not correctly projected to the target side due to missing alignment links, therefore its German counterpart *[die Terroranschläge in Mumbai im letzten Monat]* cannot be fully recovered on the target side (only the second part of the markable is projected). In (2), the source markable *[people]* is correctly projected to the target side (*man*). However, the pronoun *man* is not a markable on the target side, since it is impersonal and therefore cannot have an antecedent. The source markable *[people]*, on the contrary, is a noun and serves as an antecedent to an anaphoric pronoun later in the text (and was therefore annotated). In (3), the target markable *Америки* (America) cannot be projected from the source side, since it is not marked as a markable there, being expressed as an adjective (*American*).

Firstly, we aimed at assessing the number of errors caused by missing alignments (1). As seen from the previous section, the results obtained in the evaluation of REs with minimal spans are higher than of those with full REs, which means that the REs corresponding to one another are present in the source and target texts, but their spans are not correctly transferred due to wrong alignments and lack of word-to-word correspondence.

In general, the number of errors caused by wrong word alignment can be estimated on the basis of the alignment evaluation (Section 5.3). In order to be able to estimate the quality of the automatic word alignments for both language pairs, we look at

[7]In this example, for the target sentences, we use square brackets to show the actual spans of the projected markables and italics to show the correct spans that should have been projected.

the number of unaligned units. As one can see from the corresponding table (Table 5.1), the number of unaligned units for English-German constitutes 17.84% of the corpus, while the number of unaligned units for English-Russian is only 14.96%. Furthermore, while the Precision scores for the aligned units in German is quite high (92.95, see 5.3, Table 5.2), the Recall numbers are considerably lower (51.23, Table 5.2); however, since we do not have access to any gold alignments for English-Russian, this kind of evaluation is not available for that language pair. In sum, we can conclude that English-German alignments exhibit moderate Recall scores and contain higher percentage of unaligned units, therefore the use of a mention extractor is able to improve the projection quality of target coreference mentions; however, the English-Russian alignments are more precise and therefore more reliable than the output of a syntactic parser.

Secondly, we are interested in the structural differences between coreference chains in the two languages that are responsible for the rest of the projection errors (2, 3): These problems are the more interesting ones for a qualitative error analysis. For this purpose, we visualized the projected files and the gold standard using the coreference module of the ICARUS corpus analysis platform (Gärtner et al., 2014). 50% of the data was randomly selected for the detailed analysis, and we determined the most frequent projection errors and categorized them into three different groups. Thereafter, we tried to verify our resulting hypotheses about variation in pronominal coreference in the three languages using a larger external corpus: InterCorp[8] (Čermák and Rosen, 2012). This corpus offers an online interface for searching parallel corpora in different languages and sub-corpora. For our study, we performed both monolingual and multilingual queries (e.g., querying one side of a parallel corpus vs. querying parallel data) to verify our results.

Furthermore, we were interested in comparing our findings to available studies on multilingual nominal coreference in Contrastive Linguistics. However, the only work we found on this topic is a comparative study of nominal referring expressions for newswire texts in English and German started by Kunz (2010) and continued in several other works of the same author and her colleagues (for instance, see (Kunz et al., 2016)). Since we have already summarized the main outcomes of their research in Section 4.1, we will only focus on comparing their findings to ours at the appropriate places further in the text.

In our data, the problematic cases are those where the source language referring expression is missing or reformulated in the target text, and therefore is not being

[8]www.korpus.cz/intercorp.[accessed on 21.09.2017]

projected. We identified three categories of errors caused by structural differences among the three languages:

a) morphological differences: differences in the morphological structure of single nouns

b) differences in NP syntax: differences in the syntactic structure of noun phrases

c) non-equivalence in translation: translation divergences such as reformulations, omissions etc.

In the following, we will analyze these error classes in more detail and provide examples for each of them.

Morphological differences.
These are cases of German contractions and compound nouns. For example, as in the case of *policy towards [minorities]* and *[Minderheiten]politik*, the source language markable is not present in the target language as a separate unit, since we cannot split compound nouns and mark only one part of it. Also, cases like *zum Bahnhof* (short for *zu dem Bahnhof* ('to the station')) cause errors in the identification of spans, because we do not annotate prepositions as parts of markables on the English side. However, such cases are frequent in the German data, where, in general, the prepositions *an, bei, in, von, zu* can be contracted with subsequent determiners in written text. Our corpus study has shown that for the preposition *zu* ('to') the frequency of the contraction is 16 times higher than for the full form (InterCorp, measured in items per million (henceforth *i.p.m.*)).

Differences in NP syntax.
1: The use of articles. Some NPs are more frequently used with a definite article in German than in English, which resulted in the misidentification of spans. According to Kunz (2010), English allows the use of nouns with zero article more frequently than German. In our guidelines, nouns with zero article can only be linked to anaphoric pronouns (if any), but not between each other. This resulted in mismatching chains: English NPs with zero article do not form chains and therefore cannot be projected, while the same NPs actually form a chain in German. For example:

(72) a. Lastly, the G-20 could also help drive momentum on *climate change*. < ... >
We also have to find a way to provide funding for adaptation and mitigation - to protect people from the impact of *climate change* and enable economies

to grow while holding down pollution levels - while guarding against trade protection in the name of *climate change* mitigation.[9]

b. Schließlich könnten die G-20 auch für neue Impulse im Bereich [des Klimawandels]$_1$ sorgen. Ebenso müssen wir einen Weg finden, finanzielle Mittel für die Anpassung an [den Klimawandel]$_1$ sowie dessen Eindämmung bereitzustellen - um die Menschen zu schützen und den Ökonomien Wachstum zu ermöglichen, aber den Grad der Umweltverschmutzung trotzdem in Grenzen zu halten. Außerdem gilt es, sich vor handelspolitischen Schutzmaßnahmen im Namen der Eindämmung [des Klimawandels]$_1$ zu hüten .

The query of InterCorp data has shown that German exhibits a higher number of NPs with definite articles (57.928,55 i.p.m.) compared to English (31.405,22 i.p.m.). We also noticed that article use with named entities can vary in both languages (for example, the English *Hamas* corresponds to the German *die Hamas*). However, our corpus queries did not show any regularities yet; this issue requires a more detailed study regarding the types of Named Entities (which we assume to be the reason for the different use of articles). In the case of Russian, the absence of articles led to better results in the identification of REs, since in general, shorter spans increase the chance for a correct alignment.

2: The use of reflexive pronouns. According to our annotation scheme, we annotated reflexive pronouns only when they are independent constituents (rather than verb particles), but we observe differences in the use of these pronouns for the three languages, so that in most cases these are non-parallel, for example:

(73) a. Du hast [*dich*]$_1$ in meiner Tasche versteckt!
b. You hid in my bag!

These differences have to do with the form and distribution of reflexive pronouns. In English, we only have *-self* to express reflexivity, while in German and Russian a wider range of reflexives can be used. Furthermore, in German and Russian, it is possible to use more than one reflexive in a sentence to emphasize the action, which is not possible in English. As a result, there are fewer reflexives to be transferred from English to the target (German and Russian) sides of the corpus, which led to errors in the projection.

[9]Examples (72)-(75) are taken from the corpus developed as part of this work.

3: Pre- and post-modification. In general, we noticed that German NPs allow more complicated premodification than English and Russian. According to Kunz (2010), English tends towards postmodification, while German is less restrictive with premodification. These variations result in syntactical differences in markables and non-parallelism.

Regarding the participial constructions, one of the complications is that in German, they occur only in preposition, while in English and Russian they can be placed in both pre- and postposition. For example:

(74) a. Pakistan needs international help to bring hope to [*the young people living there*]$_1$.
b. Pakistan braucht internationale Hilfe, um [*den dort lebenden jungen Menschen*]$_1$ Hoffnung zu bringen.

Non-equivalences in translation. The following cases of non-parallelism resulted in projection errors in our dataset; however, we could not find enough evidence to characterize them as systematic.

- Personal pronouns vs. indefinite pronouns.

  (75) a. [*It*]$_1$ was pursuing a two-pronged strategy.
  b. *Man* verfolgte eine Doppelstrategie. ('One followed a two-pronged strategy.')

  The German indefinite pronoun *man* is the target of the projected annotations, but it is not a markable according to our guidelines: It is non-referring and thus unable to participate in RE chains.

- Possessive NPs vs. adjectives. Some possessive NPs in the source language (for example, *the government of [India]*$_1$) can be expressed through adjectives in the target language (*die [indische] Regierung* or *[индийское] правительство* and therefore are not markables.

- Determiners vs. possessive pronouns. Personal pronouns in English can be translated as articles in German (for example, *[its]*$_1$ *broader goal = das weiter gefasste Ziel*), so that the source RE has no correspondent in the target language. For Russian, in this case, a possessive form of a reflexive pronoun *свой* can be used, or the possessive pronoun can be omitted.

- Relative clauses in one language can correspond to participle constructions or PPs in another. Examples:
  a. [*a fat lady*]$_1$ [*who*]$_1$ wore a fur around her neck
  b. [*eine dicke Dame mit einer Pelzstola*]$_1$ ('a fat lady with a fur')

In sum, we performed a qualitative analysis of the projection errors according to their types and observed several major challenges that pose difficulties for the projection algorithm. In particular, there are several outcomes from the error analysis that need to be considered in the next experiments. Firstly, noisy word alignments result in poor recovery of coreference mentions on the target side, therefore more sophisticated techniques for mention detection to overcome alignment noise should be integrated. Secondly, translation divergences and morphological and syntactic differences between languages result in a mismatch between source and target mentions that, as a result, cannot be projected. Finally, we observed differences in the structure of coreference chains between genres, which should be studied in more detail (see Chapter 7).

6.4 Discussion

In this experiment, we observed that the annotation projection for the English-Russian language pair performs better than the annotation projection for the English-German language pair. Interestingly, we found that using a mention extractor to a high degree supports the recovery of the target mentions for the German NPs, but not for Russian, the reason being the lower quality of the word alignment for noun phrases between English-German than English-Russian. In particular, as already investigated in Section 6.3, the alignment of articles and their use in the two languages poses additional difficulties for the alignment algorithm and therefore impedes the identification of the target markables. However, due to the lack of articles, the projection to Russian is easier and therefore exhibits better scores, while using a mention extractor only adds additional noise to already precisely identified noun phrases.

The most closely related work is the approach of Postolache et al. (2006), but some differences are noteworthy. In contrast to Postolache and colleagues, we do not focus on maximizing Precision; instead, our goal is to assess how well projection can work for all the annotations. Moreover, we use two settings to test our approach: In setting 1, we use neither language-dependent software nor any additional linguistic

information about the target language in the coreference projection and evaluation, while in setting 2 we only use a mention extractor to determine the target mention boundaries. Postolache et al. (2006), in contrast, applied a dedicated Romanian-English word aligner[10] (which achieves an F-score of 83.3 compared to our 66.05 of the language-independent GIZA++) and used special rules that rely upon the POS information and syntactic heads to produce their annotations, and then discarded the incorrectly projected ones (we used such rules only in the evaluation of the projected heads of REs). These rules reduced the number of gold and projected REs in the English-Romanian corpus considerably: from 3422 to 2491 (Postolache et al., 2006).

In our case, we use all REs to evaluate the spans of the projected annotations and the resulting coreference chains. Comparing our evaluation to Postolache's evaluation of all REs[11], we can see that, already in setting 1, our results yield a higher MUC Precision for all of the genres (average 63.1 for English- German, 78.1 for English-Russian vs. 52.3 for English-Romanian), but a lower Recall for both languages (44.5/60.5 vs. 82.04), which results in different F-measure (Postolache et al. (2006) obtained an average F1 of 63.9 compared to our F1 of 52.0 for German and 67.9 for Russian). As for the B-cubed scores, we were able to achieve similar or higher Precision for stories (71.6 for German, 79.7 for Russian vs. 73.75), but not for other genres; similar to the MUC evaluation, our Recall numbers are lower than those of Postolache et al. (2006). Overall, using both metrics, we conclude that our method outperforms the method of Postolache et al. (2006) in terms of Precision for English-Russian (average Precision of 71.4 in setting 1 vs. 63.04 of Postolache et al. (2006)), and achieves comparable scores for English-German (average Precision of 63.05 in setting 2 vs. 63.04 of Postolache et al. (2006)), but not in terms of Recall. This can be explained by the lower quality of our automatic English-German alignments compared to the more precise English-Romanian alignments produced by a language-specific software; the Russian REs were extracted slightly more accurately due to the structural differences in NPs. We also observed different scores for newswire texts, stories, and medical leaflets, while Postolache et al. (2006) only used texts of one genre and in fact one author (several chapters of the same fiction book).

Keeping these different parameters in mind, in order to compare our results in a fair way, we evaluated the identification of RE heads following the same rules to

[10]The COWAL word aligner is a lexical aligner which is adjusted only for Romanian-English and requires a corpus with morpho-syntactic annotations (Tufiş et al., 2006).

[11]Since Postolache et al. (2006) only reported on MUC and B-cubed scores in their evaluation, we also compare our scores to theirs using these metrics and their unweighted average.

extract minimal spans of the projected REs, and we evaluated them against manually annotated heads in the gold standard. In this setting, we obtained higher precision than in the previous setting, and in comparison to Postolache et al. (English- Romanian, avg. F1 = 80.5), our results are somewhat lower for English-German (avg. F1 = 74.1) and slightly better for English-Russian (avg. F1 = 81.3), which we attribute to the overall more difficult (and therefore more generalizable) projection scenario in our approach.

In sum, we conclude that the annotation projection method to a high degree depends on the quality of statistical word alignments for the language pair in question. Also, we see that using limited syntactic information can improve the quality of coreference projection for German, but not for Russian, given the difference in the alignment quality for the corresponding language pairs. Moreover, our qualitative error analysis showed that further problems of the projection method are due to a set of structural differences of NPs in the three languages. Comparing our results quantitatively to the most closely related work, we argue that they are already competitive in a more target-language-neutral task setting, in particular because we scored the results with all the target mentions, used three languages rather than two, and we worked on three different genres of text. Having implemented and assessed this 'light-weight approach', in the next experiments, we are interested in implementing and testing more complex annotation projection strategies in order to see how much performance can be gained by using several source annotations.

Chapter 7

Multi-source projection of coreference chains

In this chapter, we report on our experiments on projecting coreference annotations from multiple source languages. The main idea of this part of the study is that multi-source annotation projection for coreference resolution would grant a bigger pool of potential mentions to choose from, which can be beneficial for overcoming language divergences and missing referring expressions as shown in the previous chapter. Therefore, the main goals of this part of the study are: (a) to explore different strategies of the multi-source projection of coreference chains in our experimental corpus, and (b) to evaluate the projection errors and assess the prospects of this approach for multilingual coreference resolution.

In the following, we examine the possibility of using annotation projection from English-German and English-Russian for automatically obtaining coreference annotations in the target languages (Russian and German respectively). To achieve this, we implement a multi-source annotation projection algorithm and apply it on an English-German-Russian parallel corpus in order to transfer coreference chains from two sources to the target side. Again, we operate not only in a low-resource setting, but also in a more linguistically-informed one, where we use the output of syntactic parsers to improve the identification of target mention boundaries.

Similar to the previous one, this chapter is structured as follows: We first present the experimental setup and then give an overview of the results of our experiment. Thereafter, we provide a detailed error analysis and conclude with a discussion of the results of this part of the study.

Previously published material

The experiment presented in this chapter has been published as (Grishina and Stede, 2017).

	News			Stories			Total		
	EN	DE	RU	EN	DE	RU	EN	DE	RU
Sentences	229	229	229	184	184	184	413	413	413
Tokens	6033	6158	5785	2711	2595	2307	8744	8753	8092
Markables	560	586	604	466	491	471	1026	1077	1075
Chains	115	133	133	40	40	45	155	173	178

Table 7.1: Corpus statistics for English, German and Russian

7.1 Experimental setup

In this section, we describe the experimental setup for this study, which to a high degree builds upon the direct project algorithm described in the previous chapter. However, in this part, we assess several strategies to project annotations, and we adopt several settings to test our method. The novelty of this approach is that it allows for combining coreference chains transferred from more than one source, which, to our knowledge, has not yet been implemented for coreference.

7.1.1 Parallel dataset

As already briefly described in Chapter 5.2, for this experiment, we only selected parallel sentences present in all the three languages, therefore the average number of sentences per language dropped by 5% and the average number of coreference chains per language by 6% (as compared to the corpus statistics presented in 4.4.1). Corpus statistics for the reduced corpus are presented in Table 7.1.

7.1.2 Projection strategies

Combining information coming from two or more languages is a more challenging task as compared to single-source projection where one just transfers all the information from one language to the other. For coreference, this task is non-trivial (as opposed to, for instance, multi-source projection of POS information where an intuitive majority voting strategy could be chosen), since we cannot operate on the token level and not even on the mention level: We cannot implement a strategy to choose e.g., the most frequent label for a token or a sequence of tokens (coreferent/non-coreferent), since

they belong to mention clusters which are not aligned on the source sides. In other words, if mention x_a belongs to chain A in the first source language and mention y_b belongs to chain B in the second source language, and they are projected onto the same mention z_{ab} on the target side, we do not know whether both target chains A' and B' projected from A and B respectively (and both containing the mention in question) are equal or not, as we cannot rely on chain IDs that are not common across languages. Therefore, we have to operate on the chain level and first compare projected coreference chains. We treat coreference chains as clusters, measure the similarity between them, and use this information to choose between them or combine them in the projection.

Projecting coreference chains (=clusters of mentions) from more than one language, we can have the following cases:

(a) Two chains are identical (contain all the same mentions);

(b) Two chains are disjoint (contain no same mentions);

(c) Two chains overlap (contain some identical mentions).

While cases (a) and (b) are quite straightforward, case (c) is more difficult since we have to determine whether to treat these chains as being equal or not.

Following the work of Rasooli and Collins (2015) (see Chapter 3.2.3), we rely upon two strategies – concatenation and voting – to process coreference chains coming from two sources. Since we only have two sources, instead of voting we implement intersection. In the case of coreference, we can enrich annotations from one language with the annotations from the other or create a completely new set out of two projection sets. In particular, we experiment with several naive methods and evaluate their quality, and then we combine them with each other to see how much performance can be gained.

We implement the following methods:

(1) **Concatenation:** Data is obtained from each of the languages separately and then concatenated.

   (a) `add:` Disjoint chains present in only one language are added to the projected chains from the other language. Typically, we would take projected

annotations for the best-scored[1] language and enrich them with annotations from the less-scored language.

(b) `unify-concatenate (u-con)`: Overlapping chains from both languages are merged together: If chain A and chain B overlap, we concatenate the mentions from both chains that form a new chain AB. The difference to the previous method (1a) is that here we operate on mention level while in 1a we add complete coreference chains from the first source to the target annotations coming from the second source.

(2) **Intersection:** Projected annotations are obtained by intersecting projections coming from two sources.

(a) `intersect (int)`: The intersection of coreference chains that are present in both languages is chosen. In this setting, we only look at fully identical coreference chains[2]: If chain A coming from the first source language is identical to some chain B coming from the second source language, this chain will be chosen for the output.

(b) `unify-intersect (u-int)`: The intersection of the mentions for overlapping chains is chosen: If chain A and chain B overlap, we intersect the mentions from both chains that form a new chain AB. Similar to the settings described in (1), the difference between 2a and 2b is that in 2b we operate on mention level while in 2a we intersect complete coreference chains coming from the first and the second source.

The `unify-...` methods are illustrated in Fig. 7.1[3], which shows the difference in the target coreference mentions obtained by using each of the approaches. In particular, using the `unify-concatenate` method (Fig. 7.1: (1)), all mentions coming from the source languages (English and Russian) are transferred into the target language

[1]Given that we have several source languages and one target language, the *best-scored* language (as opposed to less-scored) is defined as the language, from which the projections exhibit the highest projection quality. In our case, for instance, for English-Russian and German-Russian projections, English is the best-scored language, since projecting from it we obtain better projection scores as compared to German-Russian.

[2]As already mentioned above, we consider two coreference chains as identical if they contain all the same mentions.

[3]The example presented in this figure is taken from the corpus developed as part of this work and is slightly modified for our purposes.

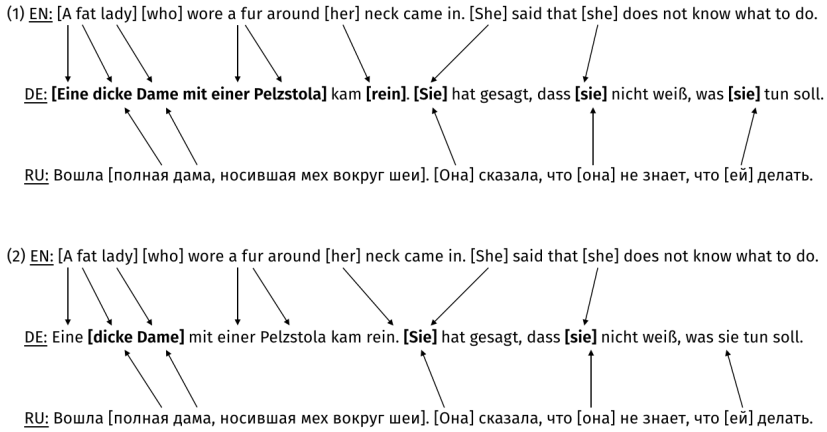


Figure 7.1: Example of automatic annotation transfer from English and Russian to German using the `unify-concatenate` (1) and the `unify-intersect` (2) methods.

(German), which results in better Recall scores. However, one can see that there are also mis-aligned mentions from one language that are mistakenly projected (*rein* 'in'), which does not happen in the `unify-intersect` method (Fig. 7.1: (2)) that only selects mentions coming from both languages simultaneously. Still, in this scenario, some correct mentions present in only one of the languages cannot be projected (*eй* 'she').

Another important aspect is to define the overlap between two coreference chains. After experimenting with several measures, we chose the Dice coefficient to estimate the overlap between two coreference chains, since it treats coreference chains from both sources equally. Specifically, we use the following formula to compute the Dice coefficient:

$$\frac{2|A \cap B|}{|A| + |B|} \tag{7.1}$$

where A and B are the mentions in coreference chains in question. We experiment with different values of overlap and choose the best one for each of the methods: We use one part of the corpus (3 texts) to determine optimal thresholds and the other one (7 texts) to obtain the results. For `u-int`, we perform intersection of mentions for all the chains with mention overlap over 0.05. For `u-con`, we select chains with 0.5 overlap value for German and 0.7 for Russian. If the overlap is less than these values, we treat these chains as disjoint.

Furthermore, each of the multi-source projection methods described above is applied in two settings: a knowledge-lean one (relying solely on the word alignments) and a more linguistically informed one (using additional linguistic resources). As described in 6.1.2, in the first setting, no additional linguistic information is used to support the projection method. In this setting, only bidirectional word alignments computed by GIZA++ (Och and Ney, 2003) are available to transfer information from one language to the other. Conversely, in the second setting, there is a mention extractor for both German and Russian available to support the recovery of the projected coreference mentions on the source side. Specifically, we relied upon the output of the MATE dependency parser[4] (Bohnet, 2010) for German and the MALT dependency parser[5] (Nivre et al., 2006) for Russian. For further details on the mention extraction procedure and the mapping of its output to the projected mentions, one should refer to Chapter 6, Section 6.1.2, where the two settings were extensively described.

7.1.3 Baselines

To assess the benefits of the multi-source approach, we re-implement several baselines for each of the settings. As baselines, we select the previously described single-source projection method: We run a direct projection algorithm for the English-German, English-Russian, German-Russian and Russian-German language pairs, since we are not interested in projecting into English, and we compute the standard mention identification and coreference scores for each of them. However, as opposed to the experiment presented in Chapter 6, we do not rely on intersective word alignments, but on all word alignments, since we are not interested in maximizing Precision at the cost of low Recall, and our goal is to obtain balanced scores to base our experiments upon.

Furthermore, we run the algorithm in the two settings described above: using only word alignments for the corresponding language pairs (setting 1) and using German and Russian mention extractors to recover the target mentions (setting 2). All the scores for the baselines in each of the settings are reported together with the results of the multi-source projection method (see Tables 7.2, 7.3, 7.4).

It should be noted that the projection results are slightly different as compared to the results reported in Chapter 6, since we did not rely on intersective word align-

[4]https://code.google.com/archive/p/mate-tools/ [accessed on 31.07.2017]

[5]http://www.maltparser.org [accessed on 31.07.2017]

ments, but used all of the alignments. In particular, we see that by using all the alignments we obtain higher Recall numbers (43.8/51.6 in setting 1, 50.0/52.4 in setting 2, for English-German (Table 7.3)/English-Russian (Table 7.4) respectively) as compared to Chapter 6 (analogously, 42.0/46.3, 57.5/55.0, see Table 6.4), but lower Precision numbers (55.3/68.0, 63.2/68.4 (Tables 7.3, 7.4) vs. 59.4/64.9, 75.6/72.8 (Table 6.4)). Therefore, the average coreference scores for English-German and English-Russian obtained in this experiment are slightly lower than those in Experiment 1 (48.7/58.5, 55.7/58.8 (Tables 7.3, 7.4) vs. 48.9/64.9, 53.7/62.3 (Table 6.4)). Also, in this experiment, using a mention extractor (setting 2) leads to an improvement for both languages as compared to Experiment 1, where it was beneficial only for the English-German language pair.

7.2 Evaluation

Similar to the previous experiment, we compute the standard coreference metrics using the latest version of the CoNLL-2012 official scorer[6], and we also compute the average scores for all the coreference metrics. For this experiment, we only evaluate strict mention matching, since we are using syntactic information in Setting 2 and are interested in comparing both settings. The results for the identification of mentions are presented in Table 7.2: on the right side for German and on the left side for Russian. The results for the baselines and the experiments are presented in Table 7.3 for German and Table 7.4 for Russian.

Furthermore, in this experiment, we focus on evaluating the projection quality for each of the methods separately and subsequently compare them to each other; therefore, we look at the projection accuracy across NP types (see Tables 7.7, 7.8) and NP length (see Fig. 7.2). In addition, we analyze the structure of coreference chains projected by different methods. Finally, we compare the overall results of the two projection settings and discuss the benefits and drawbacks of the methods in question.

[6]https://github.com/conll/reference-coreference-scorers [accessed on 01.05.2017]

	Mentions		
	P	R	F1
EN-DE	63.2	51.1	56.4
RU-DE	41.5	33.7	39.5
EN,RU→DE:			
- `add`	55.9	53.1	54.4
- `int`	87.5	3.6	6.5
- `u-con`	61.5	53.5	57.1
- `u-int`	68.2	33.5	44.6
EN→DE+*ment*	71.9	58.3	64.3
RU→DE+*ment*:	49.9	34.7	40.7
EN,RU→DE			
- `add`+*ment*	62.9	61.4	62.0
- `int`+*ment*	97.8	4.9	8.6
- `u-con`+*m*	71.1	60.5	65.3
- `u-int`+*ment*	77.5	38.0	50.5

	Mentions		
	P	R	F1
EN→RU	74.4	58.5	65.4
DE →RU:	61.6	37.4	46.3
EN,DE→RU			
- `add`	68.0	60.3	63.8
- `int`	85.0	4.8	9.0
- `u-con`	73.9	59.0	65.4
- `u-int`	79.6	36.3	49.7
EN→RU+*ment*	75.1	59.1	66.0
DE →RU+*ment*	62.5	37.9	47.0
EN,DE→RU			
- `add`+*ment*	69.1	61.5	64.9
- `int`+*ment*	85.0	3.9	7.5
- `u-con`+*ment*	74.8	59.7	66.3
- `u-int`+*ment*	81.2	36.8	50.4

Table 7.2: Results for German (on the left) and Russian (on the right): identification of mentions in a multi-source scenario

	MUC			B^3			$CEAF_m$			Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EN→DE	57.6	46.0	51.1	47.3	35.6	40.4	61.1	49.7	54.7	55.3	43.8	48.7
RU→DE	43.3	28.4	34.1	33.3	18.9	23.5	46.2	32.7	38.1	40.9	26.7	31.9
EN,RU→DE:												
- `add`	52.7	46.1	49.1	41.5	36.5	38.6	53.5	51.2	52.2	49.2	44.6	46.6
- `int`	46.7	2.5	4.5	82.3	3.1	5.6	87.5	3.6	6.5	**72.2**	3.1	5.5
- `u-con`	56.0	48.8	52.1	44.5	38.8	41.3	59.4	51.9	55.3	53.3	**46.5**	**49.6**
- `u-int`	64.7	26.1	36.7	58.6	18.7	27.3	65.7	32.4	43.1	63.0	25.7	35.7
EN→DE+*ment*	66.7	53.1	59.0	54.8	41.6	47.0	68.1	55.3	61.1	63.2	50.0	55.7
RU→DE+*ment*:	43.6	28.5	34.2	34.3	19.1	24.0	47.1	33.4	38.8	41.7	27.0	32.3
EN,RU→DE												
- `add`+*ment*	60.0	53.1	56.2	47.0	42.7	44.3	57.9	56.9	57.2	55.0	50.9	52.6
- `int`+*ment*	56.7	3.6	6.3	96.7	4.5	7.9	97.8	4.9	8.6	**83.7**	4.3	7.6
- `u-con`+*ment*	66.1	55.7	60.4	53.4	45.0	48.6	67.4	57.4	61.9	62.3	**52.7**	**57.0**
- `u-int`+*ment*	73.7	29.6	41.7	68.1	21.6	31.3	73.6	36.1	48.0	71.8	29.1	40.3

Table 7.3: Results for German: multi-source projection of coreference chains from English and Russian vs. single-source baselines

	MUC			B^3			$CEAF_m$			Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EN→RU	71.3	55.1	62.0	61.2	43.0	50.3	71.5	56.6	63.1	68.0	51.6	58.5
DE →RU:	59.1	32.0	41.3	46.8	19.6	27.3	57.3	35.1	43.3	54.4	28.9	37.3
EN,DE→RU												
- `add`	67.8	55.5	60.9	55.8	43.7	48.8	64.8	57.9	61.0	62.8	**52.4**	56.9
- `int`	87.5	3.0	5.9	85.0	4.3	8.2	85.0	4.8	9.0	**85.8**	4.0	7.7
- `u-con`	70.6	55.7	62.2	60.1	43.6	50.4	71.0	57.1	63.2	67.2	52.2	**58.6**
- `u-int`	81.6	29.3	42.9	74.8	19.5	30.6	77.6	35.5	48.6	78.0	28.1	40.7
EN→RU+*ment*	71.6	55.4	62.3	61.7	43.2	50.6	72.0	57.1	63.5	68.4	52.4	58.8
DE →RU+*ment*	59.2	32.0	41.4	47.5	19.7	27.6	57.9	35.4	43.8	54.9	29.0	37.6
EN,DE→RU												
- `add`+*ment*	68.0	55.7	61.1	56.7	44.1	49.3	65.1	58.3	61.4	63.3	**52.7**	57.3
- `int`+*ment*	87.5	2.4	4.7	85.0	3.5	6.6	85.0	3.9	7.5	**85.8**	3.3	6.3
- `u-con`+*ment*	70.9	56.0	62.4	60.9	43.9	50.8	71.5	57.5	63.6	67.7	52.5	**59.0**
- `u-int`+*ment*	82.2	29.2	42.9	76.4	19.4	30.6	78.7	35.7	49.0	79.1	28.1	40.8

Table 7.4: Results for Russian: multi-source projection of coreference chains from English and German vs. single-source baselines

7.3 Error analysis

In this section, we first analyze the projection quality for each of the projection methods (Section 7.3.1), and thereafter we perform an error analysis across the projection settings (Section 7.3.2).

7.3.1 Comparing the projection methods

We perform the error analysis by evaluating the projection quality for each of the methods described above. We first look at the common and distinct chains projected from two languages, followed by an evaluation the projection quality for different NP types and for the mentions of different length.

Common chains projected from two sources (`int`).

To analyze the common chains projected from two sources into German and Russian, we extract these chains from the target annotations and discard singletons (if any). We compute the average chain length – 2.75 and 2.13 for German and Russian respectively – and look at the types of mentions that occur in these chains. Interestingly, string match is the most frequent type, e.g., Indien - Indien, *'Афганистане'* - *'которого'* - *'Афганистане'* ('Afghanistan' - 'which' - 'Afghanistan'). Named Entities form 46% of all the markables, followed by pronouns, which are 27% of all markables. Still, the Recall numbers are too low (e.g., 3.1 and 4.0 for German and Russian in setting 1 (Tables 7.3, 7.4)) to apply this method to a small corpus.

Distinct chains added from one source to the other (`add`). We examine the chains added from the less-scored language to the best-scored, by extracting these chains separately and computing their Precision. The results for both languages exhibit low Precision: 20.0 Precision for mention extraction and 15.0 average Precision for coreference for projecting into German, and 14.0 and 7.0 for projecting into Russian. These numbers are too low to improve the projection performance in a low-resource setting.

Evaluation by NP type (`u-int`, `u-con`). In order to evaluate the projection quality for different NP types, we computed the distribution of types for the source and target annotations. For that reason, we use the POS-tagged output of TreeTagger[7] (Schmid, 1994) with the pre-trained models for German and Russian. Subsequently, we extract gold and projected markables and compare them according to their types.

[7]http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/ [accessed on 01.05.2017]

	unify-int			
	→DE #	→DE %	→RU #	→RU %
NPs	146	29.6	286	58.8
Named Entities	145	29.4	26	0.05
Pronouns			113	23.3
- Personal pronouns	82	16.6	-	-
- Possessive pronouns	35	7.1	-	-
- Demonstrative pronouns	2	0.4	-	-
- Relative pronouns	11	2.2	-	-
Total	494	100	486	100

Table 7.5: Distribution of all projected markables by type for the `u-int` for German and Russian

For German, we distinguish between the most frequent markable types: common NPs, Named Entities, personal, possessive, demonstrative and relative pronouns. For Russian, we only distinguish between the common NPs, Named Entities and pronouns, relying on the tagset available for TreeTagger[8]. Tables 7.5, 7.6 show the distribution of all markables, regardless of whether they are correct or incorrect, for both the `u-int, u-con` settings. We do not show the percentage for the markables that are not of the types described below, but count them in the total numbers.

Interestingly, the percentage of NPs + Named Entities (computed together) and pronouns for both projections and for both methods is quite comparable (e.g., 59.0 vs. 59.3 for `u-int`, 54.7 vs. 58.4 for `u-con` for NPs and NEs). However, the percentage of common NPs and Named Entities in German and Russian (computed separately) is not the same, the reason being different POS tagsets for the two languages used by TreeTagger. For Russian, a large amount of proper names were identified as common nouns, e.g., 'India', 'Mumbai', 'Hamas' etc. For German, these were identified as Named Entities.

Based on these observations, we compute the projection accuracy of each NP type as the number of correct markables of this type divided by the total number of projected markables of the same type. Tables 7.7, 7.8 show the projection accuracy

[8]http://corpus.leeds.ac.uk/mocky/ [accessed on 01.05.2017]

	u-con			
	→DE #	→DE %	→RU #	→RU %
NPs	264	28.4	450	52.2
Named Entities	245	26.3	53	6.2
Pronouns			237	27.5
- Personal pronouns	143	15.4	-	-
- Possessive pronouns	69	7.4	-	-
- Demonstrative pronouns	5	0.5	-	-
- Relative pronouns	12	1.3	-	-
Total	931	100	862	100

Table 7.6: Distribution of all projected markables by type for the `u-con` method for German and Russian

for both settings. According to these results, in the knowledge-lean approach, NPs are the less reliable projected type for German as compared to Named Entities, which is due to the fact that most of them lose their determiners at the alignment stage. For Russian, both NPs and Named Entities show similar results of over 80% with the `u-int` method. With the `u-con` method, all the scores are a bit lower due to lower Precision obtained by concatenation. As one can see from Tables 7.7 and 7.8, it is possible to significantly improve the NP identification accuracy for German by using only a mention extractor: over 17% for both methods. However, this is not the case for Russian, where NP extraction relying on word alignment does not produce that much noise: the improvement is around 0.5-2.8%.

Pronouns exhibit the best projection accuracy for both languages. For German, the highest scores are achieved by the projection of possessive (97.1), personal (95.1) and relative (81.8) pronouns. Demonstrative pronouns show the lowest score (50.0) due to their scarcity in the gold and projected data. In setting 2, we can only achieve a small improvement for different pronoun types, except for personal pronouns for German that exhibit lower accuracy.

These results explain the better projection quality when projecting to Russian compared to projecting to German, since all the projected types show fair projection accuracy. Conversely, German NPs show poorer accuracy, while constituting almost

	u-int		u-con	
	→DE %	→RU %	→DE %	→RU %
NPs	53.4	82.5	53.0	77.8
Named Entities	91.0	92.3	82.0	88.7
Pronouns		92.0		89.9
Personal pronouns	95.1	-	95.1	-
Possessive pronouns	97.1	-	94.2	-
Demonstrative pronouns	50.0	-	40.0	-
Relative pronouns	81.8	-	83.3	-

Table 7.7: Projection accuracy for the `u-int` and `u-con` methods in setting 1

one third of all the projected markables, which inevitably leads to lower Precision and Recall scores.

Evaluation by mention length (`u-int`). Finally, we compare mentions according to the number of tokens they consist of. Fig. 7.2 shows the overall amount of tokens and the number of correct tokens of this length for German (a) and Russian (b) in the `u-int` setting, in which higher Precision results were achieved. For German, the number of correct mentions gradually decreases up to the length of 5; after that, only one or no correct mentions are to be found in the target annotations. For Russian, the situation is almost the same, except for the mentions with a length of 3, which are mostly incorrect. This we attribute to the differences in the NP structure: 3-token mentions in English and German frequently contain a determiner (an article), which is erroneously projected to Russian.

7.3.2 Comparing the projection settings

Analyzing the two projection settings, one can see that both languages do not show the same improvement when applying additional syntactic information. Looking at the mention extraction scores (Table 7.2), we see the maximal improvement of 8.2 points F1 when projecting from English to German (`u-con`); however, for Russian it is only 1.1 point F1 (`add`). Similarly, looking at the scores for the coreference chains, it can be noted that the best results for German using the `u-con` method in setting 2 outperform the same method in setting 1 by 7.4 points F1, while for Russian the

	u-int+*ment*		u-con+*ment*	
	→DE %	→RU %	→DE %	→RU %
NPs	72.0	85.3	70.1	78.3
Named Entities	95.2	92.3	84.1	88.7
Pronouns		92.9		90.3
Personal pronouns	87.8	-	92.3	-
Possessive pronouns	97.2	-	98.6	-
Demonstrative pronouns	100.0	-	40.0	-
Relative pronouns	100.0	-	100.0	-

Table 7.8: Projection accuracy for the u-int and u-con methods in setting 2

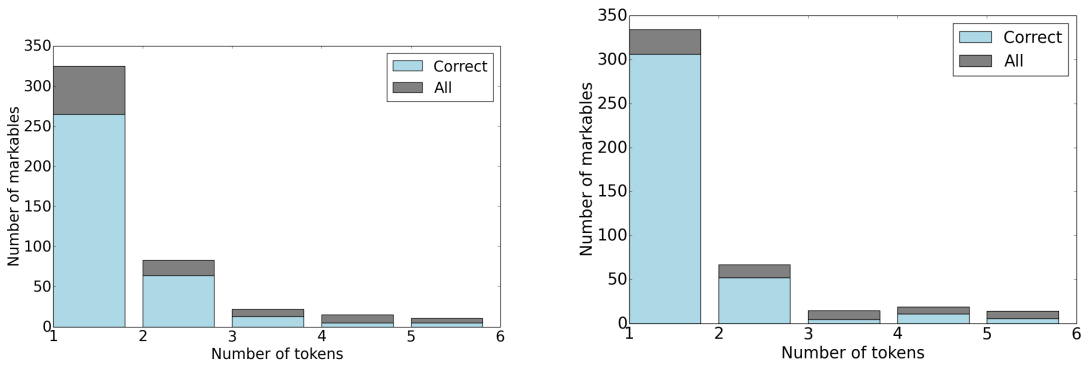
difference is only 0.4 points.

Comparing the projection accuracy in the two settings, it is clear that, for German, adding syntactic information brings the biggest improvement for the projection of nominal phrases, as seen from Tables 7.7, 7.8 (18.6% for u-int and 17.1% for u-con); the scores for Named Entities remain unchanged. Surprisingly, for personal pronouns, using mention extraction only adds noise to the markable and therefore lowers the projection performance (e.g., from 95.1% to 87.8% projection accuracy score for u-int). While the scores for possessive pronouns show small changes, the projection accuracy for demonstrative and relative pronouns also improves.

For Russian, the projection accuracy for noun phrases improves less significantly as compared to German (2.8% for u-int and 0.5% for u-con). Similarly, for the pronouns, the improvements are less than 1%.

7.4 Discussion

Analyzing the results for multi-source projection for both target languages, one can see that the scores achieved are quite comparable: the highest Precision of 83.7/85.8 for German/Russian and the highest Recall of 52.7 for both. Looking at the u-int method in setting 2, we still see that Precision is higher for Russian than for German (79.1 vs. 71.8 respectively). Overall, the best F1-scores for both languages are 57.0/59.0 German/Russian in the u-con method.



(a) German (b) Russian

Figure 7.2: Overall number of mentions and the number of correct mentions according to the number of tokens

Importantly, for both target languages and in both settings, the multi-source projection results outperform the single-source results in terms of Precision or Recall; however, still not both simultaneously. In particular, the `u-con` method exhibits higher F1 scores as compared to single-source projection (57.0 vs. 55.0 for German and 59.0 vs. 58.8 for Russian).

As for the different projection methods, the results show that the balance between Precision and Recall scores is quite stable in both settings. In particular, concatenating mentions in overlapping chains (`u-con`) resulted in the most balanced Precision and Recall scores for both German and Russian. Furthermore, Precision can be improved in two ways: by taking the intersection of chains coming from two languages and by taking the intersection of mentions in the overlapping chains in two languages. While the first scenario is more unrealistic, leading to extremely low Recall numbers, the second scenario returns much better results in terms of both Precision and Recall.

Automatic mention extraction and the mapping of target mentions to the extracted mentions to a high degree supported the identification of mentions and hence coreference scores for the English-German language pair. For Russian, in contrast, this method only helped to a small extent, the reason being already high Precision scores achieved by projecting through word alignment. The qualitative analysis has shown that incorrectly identified mentions were of wrong parts-of-speech (e.g., verbs, therefore it was not possible to map them to the automatically extracted mentions) or they were not markables in the gold annotations.

Comparing our results to the previous chapter, we can see a large improvement

in the projection quality for English-German in terms of both Precision and Recall already within the knowledge-lean setting: best Precision of 72.2 vs. 59.4, and best Recall of 46.5 vs. 42.0. In setting 2, the results are even better: 83.7 vs. 64.9 and 52.7 vs. 46.3. As for Russian, we conclude that the multi-source approach leads to a fair improvement of projection results in terms of Precision (best Precision of 85.8 for settings 1,2 vs. 75.6/72.8), but not in terms of Recall (52.4 for setting 1 and 52.7 for setting 2 vs. 57.5/55.0), which is also due to the fact that the single-source projection performed slightly worse in the absence of intersective alignments. Overall, in our experiment, we conclude that the multi-source approach outperforms the single-source approach for both language pairs; however, intersective alignments are highly beneficial for the projection into Russian (and more important than using a mention extractor).

Interestingly, the results for single-source projection also show that the different directions of projection are not equally good: Projection from English still shows the best results, while Projection from German to Russian and from Russian to German exhibit much lower F1 numbers. We attribute this to the direction of translation: Since our texts were translated from English into German and Russian, English-German and English-Russian texts are probably closer in translation than German and Russian. In our opinion, the fact that projection results with languages other than English as source are much lower has had a negative impact on the multi-source projection, since adding lower-quality annotations leads to a decrease in both Precision and Recall scores. Therefore, concatenation of the two projections with one of them being of lower quality results in a slight drop in Precision and does not improve the Recall numbers significantly. Using projections of similar quality and more languages would result in better overall scores.

In sum, our results have shown that projecting from two sources rather than one helps both to improve Precision and Recall. However, improving Precision appears to be an easier task than improving Recall. Achieving higher Recall seems to be a more difficult and expensive task as compared to eliminating noisy alignments and ensuring correct mention boundaries. If a potential target mention is absent on the source side, it can hardly be recovered in the resulting annotations.

Chapter 8

Single- and multi-source projection of automatic annotations

In this chapter, we aim at exploring the usability of annotation projection for the transfer of automatically produced coreference chains. In particular, our idea is that using several source annotations produced by different coreference resolution systems could improve the performance of the projection method in a fully automatic scenario. Our approach to the annotation projection builds upon the approach introduced in Chapter 7, but this time our goal is slightly different: We are interested in developing a fully automatic pipeline, which would support the automatic creation of annotated parallel corpora in the target language. Therefore, in contrast to the previous chapter, we use automatic source annotations produced by two state-of-the-art coreference systems, and we combine the output of our projection method for two source languages (English and German) to obtain target annotations for a third language (Russian). Our choice of the source and target languages is motivated by the availability of coreference resolution systems for English and German, which, however, were not available for Russian.

In the rest of this chapter, we will first present the experimental setup and describe each of the stages in more detail (Section 8.1). Through performing an in-depth evaluation and error analysis of the projected annotations (Sections 8.2 and 8.3), we subsequently investigate the most common projection errors assessing the benefits and drawbacks of our method, and we discuss the results (Section 8.4).

Previously published material

The experiment described in this chapter has been published as (Grishina, 2017).

8.1 Experimental setup

In this experiment, we propose a fully automatic projection setup: First, we perform coreference resolution on the source language data, and then we implement the single- and multi-source approaches to transfer the automatically produced annotations. We use our English-German-Russian unannotated corpus as the basis for our experiment and the manual annotations as the gold standard for our evaluation. The experimental setup consists of the following stages:

1. Coreference resolution on the source language data: We run source coreference resolution systems for English and German to obtain automatic annotations on the source sides of the corpus;

2. Annotation projection of automatically produced annotations: Similar to Chapter 7, we use single- and multi-source approaches and apply each of them in two settings: (1) using only word alignments and (2) using a mention extractor.

In sum, in this section, we use both single- and multi-source approaches as well as two projection settings in order to assess the applicability of our method to a new type of source data. Although using a mention extractor did not significantly improve the results for Russian (see Chapters 6 and 7), we are still interested in investigating this issue further using the new sources.

8.1.1 Coreference resolution on the source language data

Since the main goal of this experiment is to assess the quality of the projection of automatic annotations, first we need to automatically label the source language data. As we are not aware of any high-quality Russian coreference resolvers, we use Russian as our single target language and do coreference resolution on English and German in order to subsequently project annotations from each of the languages as well as from both simultaneously.

For the English side of the corpus, similar to Martins (2015), we chose the Berkeley Entity Resolution system (Durrett and Klein, 2014), which was trained on the English part of the OntoNotes corpus (Hovy et al., 2006) and is therefore compatible with our annotations. We use the joint model provided with the system which performs three core tasks: coreference resolution, Named Entity recognition and entity linking. Since our texts were already tokenized and split into sentences, we omit this step

	News		Stories		Total	
	EN	DE	EN	DE	EN	DE
Markables	486	621	429	414	915	1035
Chains	125	200	57	68	182	268

Table 8.1: Number of markables and coreference chains in the automatic annotations

during coreference resolution. According to Durrett and Klein (2014), the joint model achieves the average F1 of 61.71 on the OntoNotes dataset.

For the German side of the corpus, we use a state-of-the-art system CorZu (Tuggener, 2016) to obtain the source annotations. CorZu uses an incremental entity-mention model, which additionally makes use of filtering rules and binding theory. This system achieves the average of 61.65 F1 (without using gold mention boundaries) when tested on TüBa-D/Z dataset, according to Tuggener (2016).

After running coreference resolvers on the source language datasets, we computed corpus statistics for both languages. Corpus statistics for the English and German datasets are presented in Table 8.1. Interestingly, CorZu identified slightly more markables and coreference chains in total than Berkeley (1035 vs. 915, 268 vs. 182 respectively). In particular, the numbers of found markables and chains in English and German highly diverge for the newswire texts (486 vs. 621, 125 vs. 200 respectively), which is probably due to the fact that the newswire texts contain more complex NP types than the stories (see 4.4.1 for the distribution of NP types).

To estimate the quality of the automatically produced annotations, we evaluate the resulting dataset against our manually annotated English and German parts of the corpus. The scores for the standard coreference metrics are presented in Table 8.2. As one can see from this table, CorZu outperforms Berkeley on our dataset: the average F1 of 53.8 for German as compared to the average F1 of 40.0 for English. We attribute this drop in performance to the peculiarities of the texts in the corpus: Newswire texts contain a large number of complex nominal phrases that were not correctly identified by the systems, and short stories exhibit dialogue speech (with a large number of first and second person pronouns), which posed additional difficulties for the systems. At this point, we conclude that applying coreference models on a different type of text results in the decrease in performance, depending on the text domain (similarly, Klenner and Tuggener (2011) show that applying the English

	MUC			B^3			$CEAF_m$			Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Berkeley (En)	49.5	41.4	45.0	38.9	27.8	32.1	45.9	40.4	42.9	44.7	36.5	40.0
CorZu (De)	66.9	59.2	62.5	59.2	41.3	46.6	52.4	52.8	52.3	59.5	51.1	53.8

Table 8.2: Evaluation of the automatic source annotations vs. manual source annotations

model of CorZu in the biomedical domain leads to a significantly lower F1=30.96, as compared to F1=54.6 on the CoNLL dataset (Pradhan et al., 2011)).

Another observation regarding the difference in performance between the English and the German coreference systems is that our corpus contained a large number of Named Entities that were unknown to the coreference resolvers. Interestingly, CorZu seems to be more successful at linking them to their anaphors, which is probably due to the fact that, in German, one can rely on more morphological features than in English. As already mentioned in Chapter 6, in German, Named Entities tend to be used with an article more frequently than in English; furthermore, articles indicate their number and gender, which can be used by a coreference resolver to make a better decision, which is not possible for English Named Entities.

8.1.2 Projection of automatic annotations

As already mentioned in the beginning of this section, we perform the annotation projection of automatic annotations in two steps. Firstly, we perform a single-source annotation projection for the English-Russian and German-Russian language pairs using the algorithm described in Chapter 6, and, similar to Chapter 7, we use all the word alignments to transfer the annotations. Secondly, we use a multi-source approach to transfer annotations from both English and German into Russian. Based on the results obtained in Chapter 7, for this experiment, we select several multi-source strategies. In particular, we (a) look at disjoint chains coming from different sources and (b) use the notion of chain overlap to measure the similarity between two coreference chains that contain some identical mentions[1]. In our experiment, we apply the following strategies described in more detail in 7.1.2:

1. `add`: disjoint chains from one source language are added to all the chains projected from the other source language;

[1]Computed as Dice coefficient.

2. `unify-intersect`: the intersection of mentions for overlapping chains is selected.

3. `unify-concatenate`: chains that overlap are treated as one chain starting from a certain percentage of overlap.

As already stated before, each of the approaches is applied in two settings:

- Setting 1: relying only on word alignments produced by GIZA++ (Och and Ney, 2003);
- Setting 2: using the output of the MALT dependency parser (Nivre et al., 2006) to extract the target mentions (see 6.1.2 for details).

In sum, in this experimental setup, our main focus is to assess the applicability of the approaches developed in this study in a broader setting, given we have access to one or more coreference resolution systems on the source side.

8.2 Evaluation

Similar to Chapter 7, in the evaluation phase we compute the scores for the identification of mentions as well as standard coreference metrics for each of the approaches, and we use strict mention matching for each of the metrics. The results for the identification of mentions are presented in Table 8.3. In this table, we compute the scores for the identification of mentions in the single-source projection from English to Russian and from German to Russian, and we also compute the results for the identification of mentions using the `add` strategy, since it combines disjoint chains in the output. Importantly, we notice that using the combination of coreference chains coming from two languages can improve Recall (from 41.6 to 47.7 in setting 1 and from 41.9 to 49.1 in setting 2) as compared to the single-source methods. However, for single-source projection, in setting 2 we do not observe any considerable improvement for English-Russian as compared to setting 1 (49.1 vs. 49.5 F1), and only some improvement for German-Russian (38.7 to 40.5 F1).

Table 8.4 presents the projection results computed as standard coreference metrics. In this table, we present the results for the `unify-concatenate`, `unify-intersect` and `add` methods separately. As one can see from the table, projection from English to Russian in setting 1 outperforms projection from German to Russian by 6.5 points

	Mentions		
	P	R	F1
EN→RU	60.7	41.6	49.1
DE→RU	54.1	30.6	38.7
EN,DE→RU (add)	52.7	**47.7**	49.7
EN→RU+*ment*	61.2	41.9	49.5
DE→RU+*ment*	56.5	32.1	40.5
EN,DE→RU+*ment* (add)	54.1	**49.1**	51.2

Table 8.3: Results for the identification of mentions

F1. Moreover, while Precision numbers are quite similar, projections from English exhibit higher Recall numbers.

As for the multi-source strategies, in the knowledge-lean setting (setting 1), we were able to achieve the highest F1 of 36.2 by combining disjoint chains (`add`), which is 1.9 points higher than the best single-source projection scores and constitutes almost 62% of the quality of the projection of gold standard annotations reported in Chapter 7. We were able to achieve the highest Precision scores of 79.3 by intersecting the overlapping chains (`u-int`) and the highest Recall of 31.1 by concatenating them (`u-con`).

As for setting 2, we can see that using a mention extractor does not significantly improve the overall scores, which conforms to the results for the projection to Russian in Chapter 6. In particular, the results for English-Russian only improve from 34.3 F1 to 34.6 F1, while the results for German-Russian improve from 27.8 to 28.4 F1. Overall, for the multi-source strategies, the difference in F1 for the highest scores is only 0.4 (36.2 vs. 36.6).

	MUC			B^3			$CEAF_m$			Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EN→RU	51.7	32.6	39.8	40.6	19.6	26.0	45.7	31.3	37.0	46.0	27.8	34.3
DE→RU	55.5	23.6	32.8	42.1	13.0	19.1	43.0	25.3	31.6	46.9	20.6	27.8
EN,DE→RU												
`add`	58.5	33.6	42.5	43.9	19.8	26.9	55.7	30.3	39.1	52.7	27.9	**36.2**
`u-int`	85.2	14.9	24.7	76.8	7.8	13.8	75.8	17.1	27.6	**79.3**	13.3	22.0
`u-con`	49.4	36.1	41.5	35.9	22.1	26.7	38.3	35.2	36.5	41.2	**31.1**	34.9
EN→RU+*ment*	52.0	32.8	40.0	41.0	19.7	26.2	46.3	31.8	37.5	46.4	28.1	34.6
DE→RU+*ment*	56.2	23.7	33.0	44.2	13.4	19.8	44.4	26.0	32.5	48.3	21.0	28.4
EN,DE→RU+*ment*												
`add`	58.5	33.9	42.8	44.2	20.2	27.3	56.1	30.8	39.6	52.9	28.3	**36.6**
`u-int`	85.9	14.9	24.8	77.4	7.8	13.9	76.4	17.2	27.8	**79.9**	13.3	22.2
`u-con`	49.6	36.2	41.6	37.1	22.4	27.4	39.2	36.0	37.3	42.0	**31.6**	35.4

Table 8.4: Projection results from English and German into Russian

8.3 Error analysis

Analyzing the errors coming from each of the source languages, we first looked at the percentage of transferred mentions (Table 8.5): Using our method we were able to automatically transfer 82.7% of all the source markables from English and only 57.6% of all the source markables from German; similarly, the percentage of the transferred chains is lower for German than for English. Interestingly, while CorZu performs better on the source dataset than Berkeley, the results for the projected annotations are the opposite: Annotation projection from English to Russian performs better than from German to Russian. Our hypothesis is that the reason for the lower percentage of transferred annotations is the lower quality of word alignments for German-Russian as compared to English-Russian.

	English		German	
	#	%	#	%
Markables	757	82.7	596	57.6
Chains	182	100	227	84.7

Table 8.5: Transferred chains and markables

Similar to the evaluation presented in 6.3, we estimate the quality of the word alignments by looking at the number of unaligned tokens computed in Chapter 5.3. Not surprisingly, we see a higher percentage of unaligned words for German-Russian than for English-Russian: 17.03% vs. 14.96% respectively, which supports our hypothesis regarding the difference in the alignment quality for both language pairs. Furthermore, we compute the distribution of unaligned words: The highest percentage of unaligned tokens disregarding punctuation marks are prepositions; pronouns constitute only 3% and 5% of all unaligned words for the alignments between English-Russian and German-Russian respectively. However, these numbers do not constitute more than 5% of the overall number of pronouns in the corpus, therefore we cannot conclude that missing alignment links are responsible for the lower quality of NP alignments.

Following the error analysis strategy adopted in Chapter 7.3, we analyze the projection accuracy for common nouns ('Nc'), named entities ('Np') and pronouns ('P')

	en-ru	de-ru	en-ru	de-ru
Nc	64.5	60.7	62.4	60.2
Np	70.5	66.6	72.1	77.3
P	83.6	76.5	81.4	78.6

Table 8.6: Projection accuracy for common nouns, proper nouns and pronouns in setting 1 (on the left side) and in setting 2 (on the right side) (%)

separately[2]: Table 8.6 shows the percentage of correctly projected markables of each type out of all the projected markables of this type, with (on the right) and without (on the left) using mention extraction. Our results conform to the results of Chapter 6: For both languages, pronouns exhibit the highest projection quality, while common and proper nouns are projected slightly less accurately. Overall, for all the markables, the projection accuracy for English-Russian is around 10% better than projection accuracy for German-Russian.

Furthermore, we see that in setting 2, while using mention extraction improves the projection quality for proper names for German-Russian (at 10.6%), there is only a moderate improvement for English-Russian (at 1.6%); for the projection of common nouns, the scores decrease for both language pairs. As for pronouns, while there is also a slight improvement for German-Russian (from 76.5% to 78.6%), we see a drop in performance for English-Russian (from 83.6% to 81.4%).

Finally, we compare the projected annotations across the two genres. Interestingly, the results for the two languages vary: While the average coreference scores for English-Russian are quite comparable (news: 34.2 F1, stories: 33.3 F1), the scores for German-Russian differ considerably (news: 30.8 F1, stories: 20.8 F1). This difference we attribute to the quality of the automatic source annotations due to the lower performance of the source coreference resolvers on different genres of texts.

8.4 Discussion

In this study, our results have shown that projection from two source languages is able to reach 62% of the quality of the projection of manual annotations and improves

[2]Using the automatic POS annotations already present in the corpus and provided by TreeTagger Schmid (1994).

the projection scores by 1.9 F1 in the knowledge-lean setting and by 2.0 F1 in a more linguistically-informed setting. In general, we were able to achieve the highest F1 of 36.2 and 36.6 F1 for both settings respectively. Moreover, using the output of two completely different coreference resolution systems, we observed the similar tendencies as while projecting gold standard annotations: Projection from English to Russian achieves higher scores than projection from German to Russian, and pronouns have the highest projection accuracy. Interestingly, while the Precision scores for both language pairs are of similar range, we see a bigger difference in the Recall scores: Projection from German is around 7 points F1 worse than projection from English.

Also, we found that our multi-source strategies behave similarly when applied to automatic annotations as compared to using manual annotations. If one aims at achieving higher Precision, it is useful to use the intersection of mentions projected into the target language from both sources, while to maximize Recall it is necessary to combine the mentions coming from source languages. Still, the highest F1 score is achieved by combining complete coreference chains coming from different sources.

Furthermore, we noticed that using the output of a dependency parser for Russian to support target mention extraction does not significantly improve the results of the projection method, which is similar to the results of projecting manual annotations obtained in Chapters 6 and 7. Moreover, it even decreases the performance for some of the markable types. In our opinion, this shows that the identification of mention boundaries while projecting to Russian is not that problematic, as compared to recovering missing alignment links or dealing with noisy alignments, and was therefore not improved by a mention extractor.

Another important finding of this chapter is that using better source annotations does not necessarily result in better projection scores, which can be explained by the different quality of word alignments for both language pairs. Having investigated this issue, we conclude that alignments between German and Russian contain more unaligned units than the alignments between English and Russian, which influences the projection quality. However, using several source annotations is still beneficial for the overall performance of the projection method.

Chapter 9

Manual projection of bridging pairs

Bridging is a complex category that poses difficulties for the annotation and automatic resolution. As already described in detail in 3.1.2, the recent work on bridging typically focuses on only one language, and the inter-annotator agreement numbers are lower than those for identity coreference. Therefore, no annotation projection experiments for bridging have been performed so far, and we are not aware of any ongoing multilingual work on bridging at the point of writing this thesis.

For that reasons, in this chapter, we propose the first experiment on transferring bridging annotations from one language to the other. Since this approach has not yet been explored, in this preliminary experiment, we concentrate on the manual transfer, in order to be able to directly evaluate the advantages and shortcomings of this approach. Since bridging annotations were performed on the German side of the corpus (see Chapter 4 for details), we take these annotations as the basis and manually transfer them to the English and Russian sides of the corpus.

In Section 9.1, we report on the methodology adopted for our experiment. Subsequently, in Section 9.2 we evaluate the resulting annotations, and then, in Section 9.3, we empirically analyze the projection errors, comparing bridging relations across languages, and conclude on the applicability of our method to the bridging annotations.

Previously published material

A more compact description of this experiment, including the results presented in Section 9.2 and Section 9.3, was previously published in (Grishina, 2016).

9.1 Experimental setup

In this experiment, taking German annotations as a starting point, we annotate the English and Russian sides of our parallel corpus by manually transferring the annotations from source to target. In particular, the transfer was done by the author of this thesis, and the procedure consisted of the following steps:

1. In the source file, we identified an annotated bridging pair and the sentence the antecedent in question is located in;
2. With this information, we identified the corresponding target sentence;
3. In the target sentence, we looked for the bridging antecedent corresponding to the source bridging antecedent;
4. If the antecedent was found in target, we looked for the corresponding bridging anaphor, which can be located in the same or some other sentence;
5. If the anaphor was found, we checked if it already participated in a coreference chain;
6. If it did not participate in any coreference chain, we annotated the bridging pair in the target text and assigned it the corresponding bridging relation.

In this manner, we only annotated bridging in the target texts if both the anaphor and the antecedent were present. Furthermore, if the anaphor was already annotated as part of some other coreference chain, it was not annotated as bridging, since its definiteness was determined by the fact that it was already referring to some other entity. This also corresponds to the principle of primacy postulated in the annotation guidelines, which instructs the annotators to prefer the annotation of identity relation over bridging (see annotation principles in 4.2.4).

For an illustration, we present Fig. 9.1, which shows example source and target files. In the source file, there is an annotated bridging pair *den Schlüssel* - *ins Schloss*[1], with the former being the antecedent and the latter - the anaphor[2]. Having identified

[1]The preposition *ins* is annotated as part of the markable since it represents a contraction of the preposition *in* and the article *das*, see 4.2.2 for details.

[2]In this example, the markable *den Schlüssel* already refers to another markable *das Büro* in the previous context; however, the markable *ins Schloss* appears for the first time, therefore we consider it as a bridging anaphor.

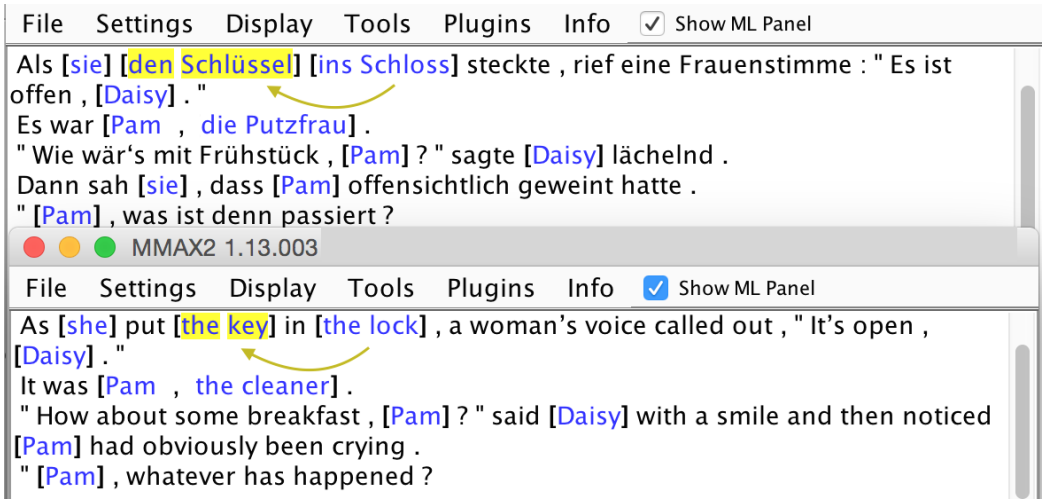


Figure 9.1: Manual transfer of bridging annotations from the German to the English side

the source pair, we first search for the corresponding bridging antecedent in the target sentence aligned to the source one. After the target antecedent (*the key*) is successfully found, we look for the corresponding bridging anaphor – *the lock* – and thereafter annotate the corresponding target pair.

9.2 Evaluation

After performing the manual transfer, we computed the number of transferred bridging pairs for each of the languages. Interestingly, the resulting number of bridging pairs for the English and Russian sides of the corpus was 188 each. As our evaluation shows, this number is lower than the original number of source pairs (432), constituting 43.5% of the total number of German bridging pairs. Moreover, we computed the percentage of transferred pairs for English and Russian out of the total number of source pairs according to the relation type (Table 9.1). As one can see from this table, for English, we transferred the highest percentage of pairs with Set-Membership and Entity-Attribute/Function relation (over 40%). Similarly, for Russian, we were able to transfer the highest percentage of Entity-Attribute/Function, Set-Membership, and Part-Whole relations. However, we notice lower numbers for the transfer of Location-Attribute, Event-Attribute and Part-Whole (for English only) relations. It should also be noted that these relations were not present in all the genres: In 4.4.1, we found that Event-Attribute and Location-Attribute relations were mostly present in

Relation	EN	RU
Part-Whole	33.3	42.1
Set-Membership	68.8	43.8
Entity-Attr/F	49.6	51.1
Event-Attr	15.6	21.9
Location-Attr	33.9	22.0

Table 9.1: Percentage of transferred bridging pairs from German to English and Russian (%)

Relation	German		English		Russian	
	#	%	#	%	#	%
Part-Whole	57	13.2	19	10.1	24	12.8
Set-Membership	16	3.7	11	5.9	7	3.7
Entity-Attr/F	268	62.0	133	70.7	137	72.9
Event-Attr	32	7.4	5	2.7	7	3.7
Location-Attr	59	13.7	20	10.6	13	6.9
Total	432	100.0	188	100.0	188	100.0

Table 9.2: Distribution of bridging relations in English, German and Russian

the newswire texts and almost absent in the other genres, posing additional difficulties for the annotation transfer.

Furthermore, we investigated the types of relations that were successfully transferred from source to target. Table 9.2 shows the distribution of different types of relations for German, English and Russian. As one can see from the table, Entity-Attribute/Function is the most frequent relation in the source annotations (constituting 62% of all relations in the corpus), followed by Location-Attribute (13.7%), and Part-Whole (13.2%). Similarly, for the target annotations, Entity-Attribute/Function is the most frequent relation in the target corpus for both English and Russian. Also, we see the lowest percentage of Event-Attribute and Location-Attribute relations, which, as already mentioned previously, were not transferred properly.

9.3 Error analysis and discussion

As we can see from the previous section, the manual projection of annotations across languages resulted in transferring less than half of the source bridging pairs to the target sides. In this section, we investigate the results of our manual annotation transfer in more detail and identify the main sources of missing links.

Having empirically investigated the resulting annotations, we found that several scenarios are responsible for the missing bridging links. Specifically, given that the antecedent and the anaphor are present on the source side, they may not be projected due to one of the following reasons:

a. The antecedent and/or the anaphor are not present on the target side. For instance, in example (76), the source bridging anaphor *das Telefon*, which refers to the antecedent *ihr Büro*, cannot be projected to the target side, since the whole source clause *dass das Telefon klingeln .. würde* is translated as a single nominal phrase *phone calls* on the target. However, the nominal premodifier *phone* cannot serve as a markable according to our guidelines.

b. The antecedent and the anaphor are present on the target side, but the anaphor already participates in some identity coreference chain, therefore, according to the principle of primacy, it cannot be annotated as bridging. For example, in (76), on the source side, there is an annotated bridging pair *ihr Büro - die Bürotür*, which, however, cannot be transferred to the target side since the corresponding English anaphor *her office door* is already linked to its antecedent *the door*.[3]

(76) (a) Jeden Morgen ging sie in [[ihr Büro]$_{B1}$]$_{B2}$ und wartete darauf, dass [das Telefon]$_{B1}$ klingeln oder ein potentieller Klient kommen würde. Eines Morgens so gegen elf klopfte jemand an [die Bürotür]$_{B2}$.
(b) Every morning she went to her office to wait for phone calls or open [the door]$_2$ to clients needing her services. One morning at about eleven o'clock someone knocked on [her office door]$_2$.[4]

[3]In this case, we notice that both sentences differ in their structure, and, upon examining the English text in more detail, one can notice that *the door* is also a bridging anaphor for which the corresponding antecedent should be found. However, this procedure is beyond annotation transfer.

[4]The examples are taken from the coreference corpus developed as a part of this work.

c. The antecedent and the anaphor are present on the target side, but the anaphor is indefinite. In this case, due to the restriction on the definiteness status of bridging markables postulated in our guidelines, the target anaphor could not be annotated. For instance:

(77) (a) Die Beziehungen zwischen den Rassen standen in [den USA]$_{B1}$ über Jahrzehnte im Zentrum der politischen Debatte. Das ging so weit, daß Rassentrennung genauso wichtig wie [das Einkommen]$_{B1}$ wurde, um politische Zuneigungen und Einstellungen zu bestimmen.
(b) Race relations in [the US] have been for decades at the center of political debate, to the point that racial cleavages are as important as *income* as determinants of political preferences and attitude.[5]

In this example, we bridge from *das Einkommen* to *den USA*, however, in the English part *income* is indefinite and thus it is not a bridging markable according to our guidelines.

Another important observation specific to the projection into Russian is that, for Russian, the lack of articles impeded the identification of bridging markables and made the decision on their definiteness much more complex. In particular, in some cases, it was not clear whether the potential anaphor is definite or indefinite. Therefore, in order to resolve controversial issues, we applied the following strategy to identify bridging markables: We used a substitution test, replacing the NP in question with the corresponding genitive NP. If the test succeeded, we considered the markable as a bridging anaphor, otherwise the markable was not annotated. For example:

(78) (a) Mary was in [the office]$_{B1}$ when somebody knocked on [the door]$_{B1}$.
(b) Мэри была в [офисе]$_{B1}$, когда кто-то постучал в [дверь]$_{B1}$.

In this example, *the door* in English is definitely unique, while in Russian we need to apply our test first: We substitute the target markable *дверь* (*door*) by a genitive form *дверь офиса* (*the door of the office*). Since it fits well into the sentence, we conclude that there is a bridging relation between the two NPs, and we annotate them as a bridging pair.

[5]The examples are taken from the coreference corpus developed as a part of this work.

Overall, the analysis of the resulting annotations has shown that our method is in general applicable to transferring bridging pairs, although it reduces the number of annotations due to the reasons described above. Furthermore, we found that the projection of different types of bridging relations is not equally good, specifically, such bridging categories as Event-Attribute and Location-Attribute exhibit the lowest numbers and require a more careful analysis. In sum, at the present stage, we conclude that bridging has to be investigated in more detail before automatic annotation projection methods could be applied.

Chapter 10

Conclusion

In this chapter, we summarize the results of this thesis and draw conclusions based on the outcomes of the previous chapters (Section 10.1). Furthermore, we provide an outlook for the future work, and we discuss the potential avenues of research that originate from this thesis (Section 10.2).

10.1 Contributions

Below, we summarize the major contributions that result from this thesis:

- We presented (a) an overview of coreference theory and (b) an extensive literature review on the annotation of various types of coreference relations (identity, near-identity, bridging) in different languages. We summarized the most prominent annotation efforts relevant to our languages.
- We provided an extensive review of the annotation projection efforts in general and those applied to coreference in particular.
- Based on our extensive study of the related work on coreference annotation, we proposed a common annotation scheme that covered the annotation of identity coreference, near-identity coreference and bridging in three languages – English, German and Russian. The annotation scheme has been made publicly available.
- As part of the annotation scheme, we developed a novel domain-independent typology of bridging relations applicable to the three languages. We have shown that our scheme achieved highly reliable inter-annotator agreement scores.

- We built the first parallel coreference corpus that consists of texts of three genres (newswire, short narratives and medicine instruction leaflets), annotated according to the developed annotation scheme. The corpus has been made publicly available.

- We designed and performed three annotation projection experiments, which adopted various approaches; furthermore, each of the experiments was carried out in two settings: a knowledge-lean one and a linguistically informed one. In the first experiment, using our manually created annotations, we considered projection from one source language only (single-source projection). Even in this scenario and using a language-independent word aligner as well as three genres of texts, we were able to outperform the most closely related projection effort of Postolache et al. (2006) in terms of coreference Precision[1] for English-Russian (average Precision of 71.4 vs. 63.04), and achieved comparable scores for English-German (63.05 vs. 63.04), although Postolache et al. (2006) used only one text genre and a language-specific word aligner. In the second experiment, we also used manual annotations, but we extended our approach to two source languages (multi-source projection), which has not been implemented for coreference before. We were the first to experiment with various combinations of coreference annotations coming from the two sources, which led to further improvement in Precision over the first experiment, by up to 18.8 points for projecting into German and by up to 10.2 points for projecting into Russian. In the third experiment, we exploited both single- and multi-source projection, but, in contrast to the first two experiments, we relied upon automatic annotations produced by two state-of-the-art coreference resolution systems, being able to achieve 93% of the precision of projecting manual annotations (P=79.9).

- We systematically compared and evaluated the projection approaches described above, and we presented both a qualitative and a quantitative error analysis to identify problematic cases. We found that noisy word alignments, translation divergences and morphological and syntactic differences between languages resulted in projection errors. Furthermore, our results showed that projection accuracy of different NP types is not equally good: Noun phrases are more challenging for the projection method than pronouns. We presented the results

[1] In terms of average MUC and B-cubed Precision, since Postolache et al. (2006) only report on these metrics.

of our evaluation across languages and across text genres, and made conclusions regarding the advantages and disadvantages of each of the approaches.

Furthermore, there are several minor contributions that result from this thesis:

- We compared our annotation scheme to the most prominent annotation schemes for English and for German and outlined the commonalities and the differences between them.
- We annotated near-identity coreference in German texts for the first time, using the already existing typology of relations.
- We analyzed the correlation between identity coreference and bridging in the annotated data. Interestingly, we found that there is a strong correlation between the length of coreference chains and the number of bridging markables attached.
- We presented a literature overview on corpus alignment approaches, and we provided an evaluation of our own corpus alignment, being able to achieve fair alignment results using a much smaller training set than in the related work.
- We demonstrated how annotation projection can be applicable to bridging by performing the first experiment with manual projection of annotations from German into English and Russian. The analysis of the resulting annotations has shown that our method is in general applicable to transferring bridging pairs, although it reduces the number of projected annotations due to several reasons (e.g., the projection anaphor already participates in target coreference chains). Therefore, we concluded that a more detailed multilingual investigation of bridging relations is required for a potential automatic transfer.

Overall, the contributions listed above helped us answer the central research questions stated in the beginning of this thesis (see Chapter 1.2). First, having explored coreference phenomena in English, German and Russian, we proposed an annotation scheme applicable to the three languages, and we also collected corpus evidence for the language divergences by projecting the annotations in our parallel texts. Second, we found that annotation projection can be successfully exploited to support and facilitate the creation of coreference corpora in new languages. In particular, it can be used as a first step in corpus creation, paired with manual assessment and correction of the resulting annotations. Furthermore, we concluded that exploiting two different

languages rather than one results in improving the projection performance for both target languages, particularly in terms of Precision.

10.2 Discussion and future directions

In this section, we summarize the directions for future research, and we present several suggestions for the described cases.

Refining the typology of bridging relations

One of the main goals of this thesis was to introduce a domain-independent typology of bridging relations, which is applicable across languages. In Chapter 4, we presented the developed bridging typology as a part of our annotation scheme, and we showed that our scheme achieves reliable inter-annotator agreement scores for anaphor and antecedent selection, and on the assignment of bridging relations. However, we found that the frequency of the appearance of these relations in the corpus varies across genres, and the inter-annotator agreement numbers are not similar for all the categories.

Therefore, in future work, we are interested in refining our typology of bridging relations by introducing a set of possible subrelations. In particular, we are interested in further exploring the Entity-Attribute/Function relation, which appeared to be the most frequent one in the corpus and for all of the genres. Also, it would be worth focusing on the relations with lower frequency (Entity-Event/Attribute, Set-Subset) and investigating their distribution on a larger amount of data.

Furthermore, as a future step, we suggest reconsidering the definition of bridging markables in our guidelines, particularly in respect to the requirement of the definite status of bridging anaphors. Following the theoretical frameworks of e.g., Asher and Lascarides (1998), and the studies of Markert et al. (2012), Hou et al. (2013) described in Chapter 3.1.2, we conclude that a possible direction would be to explore this issue in depth by taking indefinite noun phrases into consideration when annotating bridging anaphors. Having observed the differences between the definiteness of the same markables in the three languages, we consider this approach to be an interesting step for future work.

Exploring near-identity in depth

Another goal of this study was to explore the category of near-identity and its applicability to annotating a multi-genre German corpus. As stated in Chapter 4, the annotations of our corpus exhibited only a small number of near-identical markables, which was not sufficient to compute inter-annotator agreement. However, these results also conform to the results obtained by Recasens et al. (2012) for English and Catalan.

Overall, as could be seen from this study and the studies of Recasens et al. (2012), the infrequency of near-identity relations in our corpus posed major challenges for the annotators. One of the possible approaches to facilitate the annotation was introduced in the work of Recasens et al. (2012), who suggested annotating near-identity based on disagreements that arose when annotating coreference chains (see Chapter 3.1.3 for details). However, it was not possible to adopt this approach at the present time, since it requires the availability of more than two annotators. Therefore, developing an annotation strategy that would address these issues and help annotators identify near-identity in texts has to be left for future work.

Extending the projection method

Having tested and assessed several methods of annotation projection with a focus on the precision of the resulting annotations, we concluded that using two languages instead of one results in an improvement of performance. Moreover, we discovered that different language pairs behave differently in the projection (projection from English to Russian achieves higher scores than projection from English to German).

In this respect, we suggest several possible extensions:

- First, we envision future work in exploiting more than two source annotations and thus extending our approach to multiple languages. Combining coreference information coming from several sources, we can achieve higher accuracy of the incoming annotations by using e.g., majority voting.

- Second, it could be beneficial to exploit multiple coreference resolution systems for a single source language to improve the source coreference annotations.

- Furthermore, one could think of a method to align annotations on the source sides and perform the projection with this knowledge in mind.

- Also, since our study has shown an imbalance between Precision and Recall, we suggest improving Precision by using higher quality word alignments (and particularly intersective alignments). As for improving Recall, one of the possible steps could be using e.g., an automatic paraphrase detection module to recover target mentions that could not be captured by alignment links due to reformulations.

- Finally, extending our approach to other language pairs and text genres would be helpful to explore its generalizabililty for a wider range of languages.

Another important direction of future research could be exploiting annotation projection for bridging relations. As we have shown in our experiment on manual projection of bridging pairs, only 43% of the markables were successfully projected. Therefore, we suggest that a more thorough comparative investigation of bridging relations across languages is required. In addition, a more extensive study of bridging in Russian would be helpful to provide better instructions on how to determine bridging anaphors in a language without articles.

Using projection as a first step for large-scale corpus annotation and the study of language contrasts

Finally, we are interested in exploiting our approach as a first step in creating coreference corpora in new languages by providing automatically projected target coreference chains to human annotators for a subsequent validation. To reach this goal, one can exploit both manually or automatically created annotations as sources, depending on the availability of (a) a highly reliable annotation scheme as well as trained human annotators or (b) a high quality coreference resolution system for the source language.

Providing projected annotations of high Precision, we can use human annotators to correct and refine the existing annotations, which, in our opinion, can reduce the workload of the corpus annotation as well as ensure the comparability of annotations across languages. It is worth pointing out that this process can be facilitated by using recently developed tools that allow for projecting annotations in an interactive mode. For instance, the tool developed by Akbik and Vollgraf (2017) enables the user to execute annotation projection, visualize and examine its results in various settings. However, at the present stage, the projection and visualization of coreference annotations is still under development. Therefore, this step should be left as future work.

Using annotation projection to facilitate corpus creation could also help to identify language divergences, which were presented in Chapter 6.3. As a result, systematic language contrasts can be collected and described, as shown in the recent work of Lapshinova-Koltunski and Hardmeier (2017). In particular, quantifying and classifying such contrasts would be an important contribution to the field of Machine Translation, since it would minimize the information loss when translating between languages, as the authors themselves acknowledge. Therefore, a systematic investigation of these issues is an important direction for future work.

Appendix A

Parallel annotation guidelines

A.1 Introduction

These guidelines present instructions for the annotation of nominal coreference in multilingual texts. They are based on the adaptation and extension of the German part of the Potsdam Coreference Scheme (PoCoS) (Krasavina and Chiarcos, 2007), but deviate from it on a few points. Moreover, they take into account the annotation conventions of the English part of the OntoNotes coreference scheme (Hovy et al., 2006) and the RefLex guidelines (Riester and Baumann, 2017). For the time being, these guidelines address only nominal coreference; event anaphors or abstract anaphors are not being annotated. Likewise, pleonastic pronouns and pronouns with no specific antecedent are excluded from the annotation. Regarding the coreference relation, we focus on:

- *identity* (If the two nominal expressions have the same referent.);
- *near-identity* (If the two nominal expressions are partially the same in that they share most of the important characteristics, but differ in one crucial dimension.);
- *bridging*, also called 'indirect anaphora' (If the two nominal expressions refer to two objects that are related but not identical.).

For the annotation, we use the freely available MMAX-2 annotation tool[1].

In the following, Section A.2 describes in detail the types of referring expressions that are subject to the annotation. Section A.3 describes the annotation process,

[1] http://mmax2.sourceforge.net

and Section A.4 presents the classification and annotation procedure for bridging and near-identity relations. Thereafter, Section A.5 defines the attributes that have to be selected for each markable, and Section A.6 shows the annotation of a sample text in German and English.

A.2 Markables

In this section, we first discuss the various types of markables to be annotated in A.2.1, and then in A.2.2, we provide guidance on identifying their spans.

A.2.1 Types of markables

Syntactically, markables are phrases with nominal or pronominal heads. The following referring expressions are to be considered as markables[2]:

1. FULL NOMINAL PHRASES, e.g. *the big blue sky;*
2. PROPER NAMES AND TITLES, e.g. *Mr. Black;*
3. PRONOUNS:
   - PERSONAL PRONOUNS (FIRST, SECOND AND THIRD-PERSON):

     We only annotate personal pronouns if they have a specific referent in the text.

     (79) a. Hello, can [I]$_1$ help [you]$_2$? - [Daisy]$_1$ asked [the lady]$_2$.
     b. Hallo, kann [ich]$_1$ Ihnen helfen? - fragte [Daisy]$_1$ die Dame.

     (80) a. If *you* need more information about *your* medical condition, read the Package Leaflet.
     b. Wenn *Sie* weitere Informationen über *Ihre* Krankheit oder deren Behandlung benötigen , lesen *Sie* bitte die Packungsbeilage.

     In example (79), we annotate the first-person pronoun *[I]* as referring to the specific antecedent *[Daisy]* and the second-person pronoun *[you]* as referring to the specific antecedent *[lady]*. In example (80), we do not

[2]In the following, examples (79)-(87), (90), (91), (93), and (95)-(97) are taken from the corpus developed as part of this work.

annotate the personal pronouns *you* and *your*, because they do not have any specific antecedent in the text but refer to the abstract reader.

We also do not annotate first-person pronouns if they denote the author of the text (and do not have a specific antecedent in the text):

(81) *I* am sure, *our* time for standing pat, for protecting narrow interests and putting off unpleasant decisions - that time has surely passed.

- DEMONSTRATIVE PRONOUNS

(82) a. You need [a camera]$_1$ [that]$_1$ works in the dark. Hm, take [this]$_1$.
b. Sie brauchen [ein Modell]$_1$, [das]$_1$ auch nachts funktioniert. Hm, nehmen Sie [dieses]$_1$.

In example (82), the demonstrative pronoun *[this]* points back to its antecedent *[a camera]* mentioned in the previous sentence and must be annotated. Keep in mind that we do not annotate event coreference, that is why we do not consider demonstrative pronouns if they refer to a verb phrase or to a bigger discourse unit, as in the following example:

(83) The London G-20 meeting recognized that *the world's poorest countries and people should not be penalized by a crisis for which they are not responsible.* With *this* in mind, the G-20 leaders set out an ambitious agenda for an inclusive and wide-ranging response.

In example (83), *this* does not have a specific referent, but refers to the whole subordinate clause of the previous sentence and therefore should not be marked.

- RELATIVE PRONOUNS, such as *who*, *whom*, *whose*, *which*, *that* etc.

(84) a. [The Army]$_1$, [which]$_1$ recruits heavily in the Punjab, will not use [their]$_1$ force there in the way [it]$_1$ is doing in the tribal areas.
b. [Die Armee]$_1$, [die]$_1$ einen Großteil [ihrer]$_1$ Soldaten im Punjab rekrutiert, wird dort nicht mit Gewalt vorgehen, so wie [sie]$_1$ es in den Stammesgebieten tut.

Keep in mind that pronouns can be ambiguous:

(85) For both India and Pakistan, Afghanistan risks turning into a new disputed territory, like [Kashmir]$_1$, [where]$_1$ the conflict has damaged both countries for more than 50 years.

(86) Daisy managed to discover *where* Mr. Baccini's dishonest partner was now living and was anxiously expecting her cheque.

In example (85), *where* is a relative pronoun and refers to *Kashmir* (to show this, one can substitute *where* by *in which*). Conversely, in (86), *where* is not a relative pronoun and should not be annotated.

- REFLEXIVE PRONOUNS

(87) a. It's beginning to rain! - [Daisy]$_1$ exclaimed to [herself]$_1$.
b. Es fängt an zu regnen! - sagte [Daisy]$_1$ zu [sich]$_1$ [selbst]$_1$.

For German, reflexive pronouns must be annotated only if they are independent constituents but not part of a reflexive verb:

(88) Ich habe *mich* gestern gewundert. (*Mich habe ich gestern gewundert)

(89) Ich habe [mich]$_1$ gestern gesehen. (Mich habe ich gestern gesehen)

The following test should be applied: If the position of the reflexive pronoun can be changed, then the pronoun is an independent unit (89), otherwise it belongs to the verb (88). Reflexivity in German and in Russian can also be marked by other units, such as *selbst*, *selber*, *persönlich* that also must be annotated.

- PRONOMINAL ADVERBS (GERMAN)

(90) Viele Amerikaner haben Probleme mit [Rassismus]$_1$; doch wir sind [dagegen]$_1$ immun.

(91) The Army, which recruits heavily in [the Punjab]$_1$, will not use force [there]$_1$.

- HIS/HERS, HIS OR HERS are annotated as a single markable.

4. NPS WITH QUANTIFIERS

Be careful when annotating NPs with quantifiers, e.g., *all people*, *two people*, *105 Million euro* etc. If you are not sure about the definiteness of an NP, apply the following test: Try inserting a definite article or a demonstrative pronoun. If the meaning of the phrase is not changed, then the NP is definite. For example: given a markable *all people*, try to replace it with *all these people*. If it works, then the NP in question is definite.

5. NOMINAL PREMODIFIERS

In case of English nominal premodifiers, we only annotate a nominal premodifier if it can refer to a named entity (*[the [US]$_1$ politicians]$_2$*) or is an independent noun in the genitive form (*[[creditor's]$_1$ choice]$_2$*); in all other cases, nominal premodifiers are not annotated as separate markables (*bank account*).

6. GENERIC REFERENCE

Generic nouns can co-refer with definite full NPs or pronouns, but not with other generic nouns. For example:

(92) a. [Computers]$_1$ are expensive. But [they]$_1$ are really useful. *Computers* cost a lot of money.
b. [Computer]$_1$ sind teuer. Aber [sie]$_1$ sind richtig nützlich. *Computer* kosten viel Geld.

In this case, we only link the anaphoric pronoun *[they]* to its antecedent in the first sentence, *[computers]*, but we do not annotate the generic noun *computers* in the third sentence.

7. GROUPS

If all elements from a group are referred to by an anaphoric pronoun, create a group markable consisting of the set elements and then link the anaphoric pronoun to it.

(93) Did [your husband]$_1$ buy Lorna, [Mrs. Humphries]$_2$? - No, [we]$_{1+2}$ bought her together.

8. TEMPORAL EXPRESSIONS

Temporal expressions are to be annotated if they co-refer.

As mentioned previously, we do not annotate predicative forms. When a copula is used to 'equate' two nominal expressions, the predicated one is not a markable:

(94) a. [Oxford]$_1$ is *a university*. [It]$_1$ has a long history.
b. [Oxford]$_1$ ist *eine Universität*. [Sie]$_1$ hat eine lange Geschichte.

Keep in mind that in the case of change of perspective on the referent of an anaphoric expression, we should start a new chain if the already mentioned referent becomes unspecific. See the following example:

(95) So [Daisy]$_1$ tried to turn it off but pushed the wrong button and the whirring sound increased. At this point Pam's ex-husband became aware of it and turned round furiously. He realized [someone]$_2$ was watching him and swore profusely. Then he made towards [Daisy]$_2$ as though to hit her.

A.2.2 Spans of markables

Markables are always rooted in some nominal phrase (NP), and their extension is defined as follows:

- the syntactic head of the NP;
- determiners and adjectives (if any) that modify the NP;
- deverbal modifiers (participial constructions, regardless whether in pre- or postposition) that can be substituted by a subordinate clause, for example:

  (96) [Regional conflict, involving all of the region's states and increasing numbers of non-state actors]$_1$, has produced large numbers of [trained fighters, waiting for the call to glory]$_2$.

  In this case, both *[regional conflict, involving all of the region's states and increasing numbers of non-state actors]* and *[trained fighters, waiting for the call to glory]* are markables.

- dependent prepositional phrases (for example, *[Queen of England]*$_1$).
- appositions, i.e., additive material that is not syntactically integrated, are included into the markable span, but are not annotated separately:

  (97) a. [JuD, Party of Proselytizing,]$_1$ was founded in 1972.
  b. [Jud, Partei der Missionierung,]$_1$ wurde 1972 gegründet.

However, full clauses, in particular relative clauses, are not taken as parts of the markable rooted in the NP head. Therefore we annotate relative pronouns separately (see A2.1).

A.3 Annotation process

The annotation process selects only those nominal expressions that actually appear in a coreference chain (i.e., those that are mentioned at least two times in the text). When an entity is mentioned only once by some referring expression (a so-called 'singleton'), this expression is not a markable. Therefore, the annotation process involves a certain amount of 'going back and forth' in the text. Moving from left to right, when you encounter a referring expression R, check whether it anaphorically refers to an entity that has already been mentioned. If this is the case, establish a markable for R and link it to its nearest antecedent, i.e., the most recent expression A that has the same referent. If A is already a markable – that is, it already participates in a coreference chain, – this can be done right away. If, on the other hand, A is the first mention of that referent, then A also has to be annotated as a markable before the coreference link can be established. In the case of cataphoric pronouns ('Before she left, Sue locked the door') the relation is to be established in forward direction (here: from 'she' to 'Sue').

The annotation process for bridging is similar to those for identity coreference. Moving from left to right as described above, carefully examine each definite noun phrase that is not linked to any preceding antecedent. If it is definite due to the common knowledge, leave it unannotated, otherwise establish the corresponding bridging relation to its antecedent.

A.4 Bridging and near-identity

A.4.1 Bridging

Bridging relations are indirect relations that hold if a referring expression is definite because it builds a bridge to a previous expression, without being identical to it. For example, two NPs that are in a part-whole semantic relationship can refer to two objects that are related but not identical, as one being a part of the other (*room - ceiling*). Only expressions, where the referents of which are unique within a particular discourse fragment can be considered as bridging anaphors. Therefore, only definite descriptions can be linked together.

We annotate three types of referring expressions that can be bridged to: (1) sets or collections, (2) entities, and (3) events. The motivation for this is that we assume that these units are characterized by different semantic frames, and they can be bridged

to from the respective frame elements. Based on the pilot annotation rounds, we adopted the following typology of bridging relations (the lists of subcategories are only given to illustrate possible relations between the corresponding categories and are by no means complete):

1. **Physical parts - Whole**

   One NP represents a physical part of the whole expressed by the other NP. For example:

   - *the militant organization - the offices in the whole country*
   - *the telephone - the dial pad*
   - *the knee - the bone*

2. **Set - Membership**

   Sets can be represented by multiple entities or events. One can refer to a certain subset or to a single definite element of the set and bridge from this subset or element to the whole collection. We do not distinguish between sets and collections, as is done in some of the related work. Sets are homogeneous and imply that their elements are equal.

   A. SET-SUBSET

   - *the European Union - the least developed countries*
   - *the patients - the patients treated with Abraxane*

   B. SET-ELEMENT

   - *these studies - the main study*
   - *Pakistan major cities - the most populous city*

3. **Entity - Attribute/Function**

   An entity is a person or an object that has certain attributes characterizing it and certain functions it fulfills with respect to some other entity.

   A. ENTITY-ATTRIBUTE

   - *Kosovo - their current policy of rejection*

- *Mrs. Humphries - the monotonous voice*

b. Entity-function

This relation involves a bridge holding between individuals with one of the related individuals being described by his profession or function with respect to the other (Gardent et al., 2003).

- *Trends, the shop - Mr. Rangee, the owner*
- *Kosovo region - the government*

4. **Event-attribute**

   Core semantic frame elements of events are commonly time and place, while optional ones can include duration, participants, explanation, frequency etc. From these frame elements, one can bridge to the event itself.

   - *the regional conflict - the trained fighters*
   - *the surgical intervention - the operating room*

5. **Location - Attribute**

   As locations, we consider geographical entities that have permanent locations in the world. Such locations exhibit different semantic frames as compared to entities and events.

   - *Germany - in the south*
   - *Afghanistan - the population*

6. **Other**

   Other bridging relations (if any), that can not be described using the categories presented above.

If the antecedent of a bridging markable is contained inside the same NP, we mark such NPs as bridging-contained, following Riester et al. (2010). For example:

(98) [The wheel of [the bike]$_{B1}$]$_{B1}$ was completely broken.[3]

In this case, we link *the wheel of the bike* to the closest antecedent *bike*, which is a part of the same NP, and we mark it as bridging-contained. However, German compound nouns are not considered as such (*der Tisch - das Tischbein* ('the table' - 'the table leg')).

[3]The example is taken from the work of Riester et al. (2010).

A.4.2 Near-identity

Near-identity relations are seen as the middle ground between identity coreference and non-identity. They represent the so-called partial identity and are different from bridging in the way that they can not be described by any other semantic relation than identity.

If two NPs are near-identical, they share most of the important characteristics, but differ in at least one crucial dimension. This can be, for example, the change of an object through the time or the reference to different roles of the same person. In our classification, we follow the typology of near-identity relations introduced by Recasens et al. (2010)[4]. Therefore, we distinguish between the following types and subtypes of the near-identity relations[5]:

1. **Name metonymy**

   - Role
     A specific role or function performed by a person or an object is distinguished from their other facets.

     (99) "Your father was the greatest" commented an anonymous old lady while she was shaking Alessandro's hand - [[Gassman's]$_{NI1}$]$_{NI2}$ best known son. "I will miss [the actor]$_{NI1}$, but I will be lacking [my father]$_{NI2}$ especially," he said.

   - Location
     The name of a location can be used indiscriminately to describe facet(s) such as the physical place, the place associated with a (political) organization, the population living in that location, the ruling government, an affiliated organization, an event celebrated at that location, etc.

     (100) The Jordan authorities arrested, on arriving in [Iraq]$_{NI1}$, an Italian pilot who violated the air embargo to [this country]$_{NI1}$.

   - Organization
     The name of a company or other social organization can be used indiscriminately to describe facet(s) such as the legal organization itself, the

[4]The definitions and most of the relation types are taken from the corresponding work of Recasens et al. (2010a).

[5]Examples (99)-(110) are taken from the work of Recasens et al. (2010a).

facility that houses the organization or one of its branches, the company shares, a product manufactured by the company, etc.

(101) The strategy has been a popular one for [McDonalds]$_{NI1}$... It's a very wise move for them because if they would have [only just original McDonalds]$_{NI1}$, I don't think they would have done so great.

- Information realization
  A discourse entity corresponding to an informational object (e.g., story, law, review, etc.) can be split according to the format in which the information is presented or manifested (e.g., book, movie, speech, etc.). The content, however, is shared by all discourse entities.

  (102) She hasn't seen [Gone with the Wind]$_{NI1}$, but she's read [it]$_{NI1}$.

- Representation
  One NP is a representation of the other - as in a picture or a starring of a person, or a toy replica of a real object. The representation can also be of a more abstract kind, like one's mental conceptualization of an object. One NP corresponds to the thing represented; the other, to the element that represents it.

  (103) We stand staring at two paintings of [[Queen Elizabeth]$_{NI1}$]$_{NI2}$. In the one on the left, [she]$_{NI1}$ is dressed as Empress of India. In the one on the right, [she]$_{NI2}$ is dressed in an elegant blue gown.

- Other

2. **Meronymy**[67]

   - Part-whole
     One NP mentions a part to refer to the whole expressed by the other NP. The two NPs can be interpreted as referring to nearly the same discourse entity because one expresses a functionally very relevant part of the whole it belongs to. The whole is composed of different, functionally distinct,

[6] In near-identity, metonymy can take place between two NPs that could be substituted by one another in the text, while in bridging these NPs should be clearly different and could be linked only via a 'part-whole' relation.

[7] If meronymy and name metonymy co-occur, metonymy is preferred.

parts (e.g., the engine and the car), organized into some kind of patterned organization or structure.

(104) Bangladesh Prime Minister Hasina and [President Clinton]$_{NI1}$ expressed the hope that this trend will continue ... Both [the US government]$_{NI1}$ and American businesses welcomed the willingness of Bangladesh.

- ENTITY-ATTRIBUTE
  For example, one NP expresses the constituent material of the other NP. Unlike components, the stuff of which a thing is made cannot be separated from the object.

  (105) The City Council approved legislation prohibiting selling [alcoholic drinks]$_{NI1}$ during night hours... Bars not officially categorized as bars will not be allowed to sell [alcohol]$_{NI1}$.

- SET-SET
  The two NPs denote two largely overlapping sets. Since each is not clearly bounded, the reader intuitively interprets the two sets as near-identical even though they might not correspond to exactly the same collection of individuals. The collection consists of repeated, similar members, and the members are not required to perform a particular function distinct from one another. This is the preferred type whenever a strict identity relation between the two plural NPs is dubious (no total overlapping).

  (106) Last night in Tel Aviv, [Jews]$_{NI1}$ attacked a restaurant that employs Palestinians, "we want the war", [the crowd]$_{NI1}$ chanted.

3. **Spatio-temporal function**
   The discourse entity is split based on different values for its spatial or temporal characteristics: it is the 'same' entity or event but realized in another location or time.

   - PLACE
     The same discourse entity is instantiated in different physical locations, each time resulting in a different discourse entity due to the change in the spatial feature. It is possible for the entities to coexist but not in the same place.

(107) a. [New York's New Year's Eve]$_{NI1}$ is one of the most widely attended parties in the world... Celebrating [it]$_{NI1}$ in the Southern Hemisphere is always memorable, especially for those of us in the Northern Hemisphere.

- TIME
  Similar to *Place* but the split into different discourse entities is due to a change of the temporal value. It sees a physical object as a function from time to a portion of space, a slice of the object's history. Thus, it is not possible for the temporally-different discourse entities to coexist.

  (108) On homecoming night [Postville]$_{NI1}$ feels like Hometown, USA, but a look around this town of 2,000 shows it's become a miniature Ellis Island... For those who prefer [the old Postville]$_{NI1}$, Mayor John Hyman has a simple answer.

- NUMERICAL FUNCTION
  The two NPs refer to the same function (e.g., price, age, rate, etc.) but have different numerical value due to a change in time or a change in space. Although *Place* or *Time* might apply, *Numerical function* is more specific.

  (109) At 8, [the temperature]$_{NI1}$ rose to 99° F. This morning [it]$_{NI1}$ was 85° F.

- ROLE FUNCTION
  The two NPs refer to the same role (e.g., president, director, etc.) but is filled by a different person due to a change in time or space. Although *Place* or *Time* relations might apply, *Role function* is more specific.

  (110) In France, [the president]$_{NI1}$ is elected for a term of seven years, while in the United States [he]$_{NI1}$ is elected for a term of four years.

A.4.3 General approach

Bridging and near-identity relations are generally directed from right to left and are annotated separately from identity relations. Each markable could have only one outgoing relation, but multiple incoming relations are allowed (e.g., a markable can not be an anaphor of both identical and near-identical antecedent, but it can be the antecedent for a bridging and an identical anaphor). The aforementioned principles

regarding the types and the size of markables hold for bridging and near-identity markables as well.

Cataphoric bridging and near-identity relations (directed from left to right) are allowed if the cataphoric antecedent is semantically closer to the anaphor than the possible anaphoric antecedent. For example:

(111) Ich kam [ins Büro] und nahm [den Hörer]$_{B1}$ ab. [Das Telefon]$_{B1}$ hat nicht funktioniert.
I came into [the office] and took up [the receiver]$_{B1}$. [The telephone]$_{B1}$ did not work.

In this example, the possible anaphoric antecedent *ins Büro* ('into the office') is less semantically related to the markable *den Hörer* ('the receiver') than the cataphoric antecedent, so we establish the relation from left to right and mark *das Telefon* ('the telephone') as its antecedent. Cataphoric antecedents can only be found in the same or in the next sentence, but not further in the text. Consider the following example:

(112) Ich kam [ins Büro]$_{B1}$ und nahm [den Hörer]$_{B1}$ ab. Meine Kollegin hat angerufen.
< ... > [Das Telefon] hat nicht funktioniert.
I came into [the office]$_{B1}$ and took up [the receiver]$_{B1}$. My colleague called.
< ... > [The telephone] did not work.

In this case, we may want to link *den Hörer* ('the receiver') und *ins Büro* ('into the office'), but we do not link any of them to *das Telefon* ('the telephone').

Furthermore, we postulate several principles to resolve controversial issues:

A. The principle of semantic relatedness: In the case of multiple candidates, one has to pick the candidate that is semantically most closely related to the anaphoric (or cataphoric) markable.

(113) [[Das Telefon]$_{B1}$]$_{B2}$ klingelte. Ich kam [ins Büro]$_{B2}$ und nahm [den Hörer]$_{B1}$ ab.
[[The telephone]$_{B1}$]$_{B2}$ rang. I came into [the office]$_{B2}$ and took up [the receiver]$_{B1}$.

In this case, we link *das Telefon* ('the telephone') to *ins Büro* ('into the office') and *den Hörer* ('the receiver') to *das Telefon* ('the telephone').

B. The principle of primacy: In case of multiple possible relations, one has to prefer identity over near-identity and near-identity over bridging. When in doubt, establish an identity relation rather than near-identity, and a near-identity relation rather than bridging. In other words, the hierarchy is as follows:
Identity ← Near-Identity ← Bridging

(114) Last night in Tel Aviv, [Jews]$_1$ attacked a restaurant that employs Palestinians, "we want the war", [the crowd]$_1$ chanted[8].

In this example, both near-identity and bridging are possible; however, according to the principle of primacy, near-identity is to be chosen.

C. The proximity principle: One always has to link a markable to its closest antecedent rather than establish a new relation, in accordance with the principle of primacy. For example[9]:

(115) Today [the right knee]$_{B1}$ is markedly swollen and there is a deformity overlying [[the patella]$_1$]$_{B1}$. [The patella]$_1$ appears to be high riding at this time.

In the first sentence, the markable *the patella* is bridged to *the right knee*. However, in the second case, we do not have to bridge *the patella* to *the right knee* again, for there is a preceding markable that *the patella* can be linked to with the identity relation according to the proximity principle.

A.5 Attributes

A.5.1 Attributes for all markables

A.5.1.1 referentiality

1. not_specified - during the annotation, no decision was taken
2. discourse_cataphor - newly mentioned underspecified discourse entity for which meaning is denoted later by a more specific antecedent (except for bridging-contained)
3. referring - discourse entity that can interpreted on the basis of the previous context
4. discourse_new - the first mention of a discourse entity + all bridging markables, including bridging-contained, except for cataphoric (to be marked as such)
5. other - impossible to decide between (a)-(d)

[8]This example is taken from the work of Recasens et al. (2010a).

[9]The example is taken from the corpus developed as part of this work.

A.5.1.2 dir_speech

1. text_level - the markable does not appear in direct or indirect speech
2. direct_speech - the markable appears in the direct speech
3. indirect_speech - the markable appears in the indirect speech

A.5.1.3 phrase_type

1. np - nominal phrase (in general, markables are nominal phrases)
2. pp - prepositional noun phrase (used only in the case of contractions, when we have to annotate the preposition as a part of the markable)
3. other - if (a) and (b) are not applicable

A.5.1.4 np_form

1. NE - named entity
2. defNP - definite NP
3. indefNP - indefinite NP
4. ppers - personal pronoun
5. ppos - possessive pronoun
6. padv - pronominal adverb
7. pds - demonstrative pronoun
8. rel - relative pronoun
9. refl - reflexive pronoun

A.5.1.5 ambiguity

1. not_ambig - there is no ambiguity present
2. ambig_ante - the antecedent is ambiguous, therefore the anaphor is linked with several possible candidates

3. ambig_rel - the relation is ambiguous (i.e., it is not clear whether the relation is marked correctly)

4. ambig_rel_ante - both the antecedent and the relation are ambiguous

A.5.1.6 complex_np

Complex NPs are those containing embedded NPs, deverbal modifiers or appositions (as defined in A.2.2).

1. not_specified - during the annotation, no decision was taken

2. yes

3. no

A.5.1.7 grammatical role

1. not_specified - during the annotation, no decision was taken

2. sbj - subject

3. dir_obj - direct object

4. indir_obj - indirect object

5. other - none of the above applicable

A.5.1.8 comment

The comment field is to be filled in case of any uncertainty regarding the aforementioned attributes.

A.5.2 Attributes: identity

A.5.2.1 identical_antecedent

The value of this field represents the antecedent of the markable and is filled automatically once the relation is established.

A.5.3 Attributes: near-identity

A.5.3.1 nident_antecedent

The value of this field represents the near-identical antecedent of the markable and is filled automatically once the relation is established.

A.5.3.2 nident_type

The type values correspond to the types of near-identity explained in detail in A.3.2.

1. name metonymy
    (a) role
    (b) location
    (c) organization
    (d) information realization
    (e) representation
    (f) other
2. meronymy
    (a) part-whole
    (b) stuff-object
    (c) set-set
3. spatio-temporal relation
    (a) place
    (b) time
    (c) numerical function
    (d) role function
4. other
    (a) other_comment

A.5.4 Attributes: bridging

A.5.4.1 bridging_antecedent

The value of this field represents the antecedent of the bridging markable and is filled automatically once the relation is established.

A.5.4.2 bridging_type

The type values correspond to the types of bridging relations explained in detail in A.3.1.

1. part-whole
2. set-membership
3. entity-att_func
   (a) entity_attribute
   (b) entity_function
4. event-attribute
5. location-attribute
6. other
   (a) other_comment

A.5.4.3 bridging_contained

Mark if a markable is bridging-contained or not (see A.3.1 for details).

1. yes
2. no

A.6 Sample annotation

In the following, we demonstrate an analysis of a sample text in English and German. We annotate identity, bridging and near-identity in both texts, and we provide clarifications to some of the decisions, highlighting the differences between languages.

CAMBRIDGE - [Last month's terrorist assault in Mumbai]$_1$ targeted not only [India's]$_7$ economy and sense of security. [[Its]$_1$]$_{B1}$ broader goal was to smash [the India-Pakistan détente]$_2$ [that]$_2$ has been taking shape since 2004. [[The attackers]$_3$]$_{B1}$ did not hide [their]$_3$ faces or blow [themselves]$_3$ up with suicide jackets. Anonymity was not [their]$_3$ goal. [They]$_3$ wanted to be identified as defenders of [a cause]$_4$. Unless [this cause]$_4$ is fully understood, and [its]$_4$ roots revealed across [the region]$_5$, [this attack]$_1$ may prove to be the beginning of the unmaking of [South Asia]$_5$. [Regional conflict]$_6$, involving all of the [region's]$_5$ states and increasing numbers of non-state actors, has produced large numbers of trained fighters, waiting for the call to glory. Within both [India]$_7$ and [Pakistan]$_{11}$, economic disparities and a sense of social injustice have created fertile ground for conflict. The use and abuse of religious fervor, whether jihadi or Hindu fundamentalist, are striking at the roots of communal harmony across [South Asia]$_5$.

Much of the current trouble can be traced to [Afghanistan]$_8$, [whose]$_8$ tragedy could never have remained confined within [[[its]$_8$]$_{B2}$ designated borders]$_{B2}$. The dynamics of [the region]$_5$ changed when the Afghan freedom fighters of the 1980's were converted into mujahidin through [a criminal enterprise]$_9$ in [[which]$_9$]$_{NI1}$ both [the West]$_{10}$ and the Muslim world happily participated. [Pakistan]$_{11}$, always insecure about [India]$_7$, became the hub of this transformation. [The West]$_{10}$ thought [it]$_{11}$ had moved on after the fall of the Soviet empire, but [the region]$_5$ - and increasingly the global community - continues to pay a heavy price for [this unholy project]$_{NI1}$.

CAMBRIDGE - [[Die Terroranschläge in Mumbai im letzten Monat]$_1$]$_{B1}$ sollten nicht nur die Wirtschaft und das Sicherheitsgefühl [Indiens]$_8$ treffen. Das weiter gefasste Ziel bestand in der Zerschlagung [der Entspannungspolitik]$_2$, [die]$_2$ seit dem Jahr 2004 Gestalt angenommen hatte. [[Die Täter]$_3$]$_{B1}$ haben weder [ihre]$_3$ Gesichter verhüllt noch [sich]$_3$ [selbst]$_3$ in der Manier von Selbstmordattentätern in die Luft gesprengt. Anonymität lag nicht in [ihrer]$_3$ Absicht. Vielmehr wollten [sie]$_3$ als Kämpfer für [eine Sache]$_4$ erkannt werden. Wenn es nicht gelingt, [diese Sache]$_4$

vollständig zu verstehen und [ihre]$_4$ Ursachen in [der gesamten Region]$_5$ offen zu legen, könnte [dieser Terroranschlag]$_1$ den Beginn der Zerstörung [Südasiens]$_5$ markieren. [Der regionale Konflikt]$_6$, in [den]$_6$ alle Staaten [der Region]$_5$ und eine steigende Zahl nicht-staatlicher Akteure verwickelt sind, brachte [zahllose ausgebildete Kämpfer]$_7$ hervor, [die]$_7$ darauf warten, zu den Waffen gerufen zu werden. Sowohl in [Indien]$_8$ als auch in [Pakistan]$_{12}$ haben wirtschaftliche Ungleichheiten und ein Gefühl der sozialen Ungerechtigkeit den Boden für Konflikte aufbereitet. Die Nutzung und der Missbrauch religiösen Eifers - ob als Dschihadi oder Hindu-Fundamentalist - erschüttern die Wurzeln der regionalen Harmonie in [ganz Südasien]$_5$.

[Ein großer Teil der aktuellen Probleme]$_9$ hat [seinen]$_9$ Ursprung in [Afghanistan]$_{10}$, [dessen]$_{10}$ Tragödie niemals innerhalb [[[seiner]$_{10}$]$_{B2}$ ausgewiesenen Grenzen]$_{B2}$ bleiben konnte. Die Dynamik in [der Region]$_5$ veränderte sich, als die afghanischen Freiheitskämpfer der 1980er Jahre im Zuge [eines kriminellen Unterfangens]$_{11}$, an [[dem]$_{11}$]$_{NI1}$ sich sowohl [der Westen]$_{13}$ als auch die muslimische Welt eifrig beteiligten, zu Mudschaheddin gemacht wurden. Das gegenüber [Indien]$_8$ immer unsichere [Pakistan]$_{12}$ wurde zur Drehscheibe dieser Veränderungen. Im [Westen]$_{13}$ dachte man, [das Land]$_{12}$ hätte sich seit dem Zusammenbruch des Sowjetreiches weiterentwickelt, aber [die Region]$_5$ - und zunehmend auch die internationale Gemeinschaft - zahlen weiterhin einen hohen Preis für [dieses unheilige Projekt]$_{NI1}$.

(line 2) Here, we encounter a markable expressed by a possessive pronoun – *its*. However, on the German side, we encounter a reformulation (*das weiter gefasste Ziel*), which does not contain the corresponding pronoun.

(line 2) Although the nominal phrase *the India-Pakistan détente* is definite, we conclude that it is definite due to the common knowledge (since the conflict between India and Pakistan is well-known), therefore we do not establish any bridging links.

(line 3) Here, we encounter the first bridging anaphor *the attackers*, which refers to the aforementioned event *last month's terrorist assault in Mumbai*. Therefore, we establish the link to the closest referring expression of this chain, which is *its*. Since the anaphor describes the participants of an event, the corresponding type of relation is Event/Attribute. In German, this pair is also present, although due to the reformulation described above, the antecedent of the bridging anaphor is the full noun phrase *Die Terroranschläge in Mumbai im letzten Monat*.

(line 15) We annotate *its designated borders* as bridging-contained, since *the borders* refer to the respective country – Afghanistan. The corresponding relation is Location/Attribute. This case is marked similarly in German (see line 40).

(line 21) We see *this unholy project* as near-identical to *a criminal enterprise*, and we establish the link to its closest antecedent – *which*. As for the relation, we see *a criminal enterprise* as an entity (representing a criminal activity) that might have a project (plan of action) as one of its attributes, therefore we choose Meronymy, Entity/Attribute. Alternatively, one can also mark this relation as Set-Set, treating the two NPs as overlapping sets (the enterprise includes projects, and the project implies business activity). In German, we annotate this case in a similar way (see line 48).

(line 27) Here, we encounter two markables expressed by reflexive pronouns – *sich* und *selbst*. However, it is worth noticing that in English only one markable is present – *themselves* (see line 4), which is an example of language divergence.

(line 34) Although the nominal phrase *zu den Waffen* is definite, it is part of an idiomatic expression *zu den Waffen gerufen werden*, therefore it should not be annotated as bridging.

Bibliography

Agić, Ž., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing*.

Agić, Ž., Johannsen, A., Plank, B., Martínez, H. A., Schluter, N., and Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Akbik, A., Chiticariu, L., Danilevsky, M., Li, Y., Vaithyanathan, S., and Zhu, H. (2015). Generating high quality proposition banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*, pages 397–407.

Akbik, A. and Vollgraf, R. (2017). The projector: An interactive annotation projection visualization tool. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 43–48.

Ariel, M. (1985). The discourse functions of given information. *Theoretical Linguistics*, 12(s1):99–114.

Ariel, M. (1988). Referring and accessibility. *Journal of linguistics*, 24(1):65–87.

Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8:29–87.

Asher, N. and Lascarides, A. (1998). Bridging. *Journal of Semantics*, 15(1):83–113.

Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th International Conference on*

Computational linguistics-Volume 1, pages 79–85. Association for Computational Linguistics.

BBN-Technologies (2006). *Coreference Guidelines for English OntoNotes—Version 6.0*. Linguistic Data Consortium. BBN Pronoun Coreference and Entity Type Corpus.

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., and Žabokrtskỳ, Z. (2012). Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling)*, pages 231–246.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd international conference on computational linguistics (Coling)*, pages 89–97. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. volume 19, pages 263–311. MIT Press.

Calhoun, S., Nissim, M., Steedman, M., and Brenier, J. (2005). A framework for annotating information structure in discourse. In *Proceedings of the workshop on frontiers in corpus annotations ii: pie in the sky*, pages 45–52. Association for Computational Linguistics.

Čermák, F. and Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

Chiarcos, C., Stede, M., and Warzecha, S. (2016). Nominale Koreferenz. In Stede, M., editor, *Handbook Textannotation*, chapter 6, pages 71–85. Universitätsverlag Potsdam, Potsdam.

Chinchor, N. and Hirschman, L. (1997). MUC-7 Coreference Task Definition. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Clark, H. H. (1975). Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174. Association for Computational Linguistics.

Clark, K. and Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Denis, P. and Baldridge, J. (2008). Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 660–669. Association for Computational Linguistics.

Durrett, G. and Klein, D. (2014). A joint model for entity analysis: Coreference, typing, and linking. In *Transactions of the Association for Computational Linguistics*.

Ehrmann, M., Turchi, M., and Steinberger, R. (2011). Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the Recent Advances in Natural Language Processing conference*, pages 118–124.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Ganchev, K., Gillenwater, J., and Taskar, B. (2009). Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 369–377. Association for Computational Linguistics.

Gardent, C., Manuélian, H., and Kow, E. (2003). Which bridges for bridging definite descriptions. In *Proceedings of the Workshop on Linguistically Interpreted Corpora*, pages 69–76. Association for Computational Linguistics.

Gärtner, M., Björkelund, A., Thiele, G., Seeker, W., and Kuhn, J. (2014). Visualization, search, and error analysis for coreference annotations. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12. Association for Computational Linguistics.

Grishina, Y. (2016). Experiments on bridging across languages and genres. In *Proceedings of the Coreference Resolution Beyond OntoNotes (CORBON) Workshop*. Association for Computational Linguistics.

Grishina, Y. (2017). Combining the output of two coreference resolution systems for two source languages to improve annotation projection. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*. Association for Computational Linguistics.

Grishina, Y. and Stede, M. (2015). Knowledge-lean projection of coreference chains across languages. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora (BUCC)*, page 14. Association for Computational Linguistics.

Grishina, Y. and Stede, M. (2016). *Parallel coreference annotation guidelines*.

Grishina, Y. and Stede, M. (2017). Multi-source projection of coreference chains: assessing strategies and testing opportunities. In *Proceedings of the 2nd Coreference Resolution Beyond OntoNotes (CORBON) Workshop*, Valencia, Spain. Association for Computational Linguistics.

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014). ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *9th International Conference on Language Resources and Evaluation (LREC)*, pages 3191–3198. European Language Resources Association.

Haghighi, A. and Klein, D. (2009). Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, pages 1152–1161. Association for Computational Linguistics.

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

Harabagiu, S. M. and Maiorano, S. J. (2000). Multilingual coreference resolution. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 142–149. Association for Computational Linguistics.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). 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 (DiscoMT)*, pages 1–16. Association for Computational Linguistics.

Hinrichs, E., Kübler, S., Naumann, K., Telljohann, H., and Trushkina, J. (2004). Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, pages 51–62.

Hinrichs, E. W., Kübler, S., and Naumann, K. (2005). A unified representation for morphological, syntactic, semantic, and referential annotations. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 13–20. Association for Computational Linguistics.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44(4):311–338.

Hoste, V. and De Pauw, G. (2006). Knack-2002: a richly annotated corpus of Dutch written text. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*.

Hou, Y., Markert, K., and Strube, M. (2013). Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917.

Hou, Y., Markert, K., and Strube, M. (2014). A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2082–2093.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90% solution. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(03):311–325.

Johannsen, A., Agić, Ž., and Søgaard, A. (2016). Joint part-of-speech and dependency projection from multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Klenner, M. and Tuggener, D. (2011). An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In *In Proceedings of the Recent Advances in Natural Language Processing Conference*, pages 178–185.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Kolhatkar, V., Roussel, A., Dipper, S., and Zinsmeister, H. (2018). Survey: Anaphora with non-nominal antecedents in Computational Linguistics: a survey. In *Computational Linguistics*, volume 44, pages 547–612.

Krasavina, O. and Chiarcos, C. (2007). PoCoS: Potsdam coreference scheme. In *Proceedings of the Linguistic Annotation Workshop*, pages 156–163. Association for Computational Linguistics.

Kunz, K., Lapshinova-Koltunski, E., and Martınez, J. M. (2016). Beyond identity coreference: Contrasting indicators of textual coherence in English and German. *Proceedings of the Coreference Resolution Beyond OntoNotes (CORBON) Workshop*, page 23.

Kunz, K. A. (2010). *Variation in English and German nominal coreference: a study of political essays*, volume 21. Peter Lang.

Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. In *Computational Linguistics*, volume 20, pages 535–561.

Lapshinova-Koltunski, E. and Hardmeier, C. (2017). Discovery of discourse-related language contrasts through alignment discrepancies in English-German translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81. Association for Computational Linguistics.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). 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 (CoNLL): Shared task, pages 28–34. Association for Computational Linguistics.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Li, Z. and Zhou, M. (2010). Use semantic meaning of coreference to improve classification text representation. In *Proceedings of the 2nd IEEE International Conference on Information Management and Engineering (ICIME)*, pages 416–420.

Loáiciga, S., Stymne, S., Nakov, P., Hardmeier, C., Tiedemann, J., Cettolo, M., and Versley, Y. (2017). Findings of the 2017 DiscoMT shared task on cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 1–16.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics.

Markert, K., Hou, Y., and Strube, M. (2012). Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 795–804. Association for Computational Linguistics.

Martins, A. F. (2015). Transferring coreference resolvers with posterior regularization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1427–1437.

McCarthy, J. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. In *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*.

McEnery, T. and Xiao, R. (2007). Parallel and comparable corpora: What is happening. *Incorporating Corpora. The Linguist and the Translator*, pages 18–31.

Melamed, I. D. (1998). Annotation style guide for the blinker project. *arXiv preprint cmp-lg/9805004*.

Mitkov, R. (2002). *Anaphora Resolution*. Longman.

Mitkov, R. and Barbu, C. (2002). Using bilingual corpora to improve pronoun resolution. *Languages in contrast*, 4(2):201–211.

Mitkov, R., Evans, R., Orăsan, C., Dornescu, I., and Rios, M. (2012). Coreference resolution: To what extent does it help NLP applications? In *Text, Speech and Dialogue*, pages 16–27. Springer.

Moosavi, N. S. and Strube, M. (2016). Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Müller, C. and Strube, M. (2001). MMAX: A tool for the annotation of multimodal corpora. In *In Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

Naumann, K. and Möller, V. (2006). Manual for the annotation of in-document referential relations. *University of Tübingen*.

Nedoluzhko, A., Mírovskỳ, J., Ocelák, R., and Pergler, J. (2009). Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, pages 1–16.

Neumann, A. (2015). Discoursegraphs: A graph-based merging tool and converter for multilayer annotated corpora. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA)*, number 109, pages 309–312. Linköping University Electronic Press.

Nicolov, N., Salvetti, F., and Ivanova, S. (2008). Sentiment analysis: Does coreference matter. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 37.

Nissim, M., Dingare, S., Carletta, J., and Steedman, M. (2004). An annotation scheme for information status in dialogue. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Language Resources and Evaluation Conference*, volume 6, pages 2216–2219.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29, pages 19–51.

Ogrodniczuk, M. (2013). Translation-and projection-based unsupervised coreference resolution for Polish. In *Language Processing and Intelligent Information Systems*, pages 125–130. Springer.

Ogrodniczuk, M., Głowinska, K., Kopec, M., Savary, A., and Zawisławska, M. (2013). Polish coreference corpus. *Journalism*, 3(7,078):19–53.

Ogrodniczuk, M., Kopec, M., and Savary, A. (2014). Polish coreference corpus in numbers. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3234–3238.

Ozdowska, S. (2006). Projecting POS tags and syntactic dependencies from English and French to Polish in aligned corpora. In *Proceedings of the International Workshop on Cross-Language Knowledge Induction*, pages 53–60. Association for Computational Linguistics.

Padó, S. (2007). *Cross-lingual annotation projection models for role-semantic information*. PhD thesis, Saarland University.

Padó, S. and Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

Poesio, M. (2000). Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Poesio, M. (2004). The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162.

Poesio, M. and Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Poesio, M., Bruneseaux, F., and Romary, L. (1999). The MATE meta-scheme for coreference in dialogues in multiple languages. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 65–74.

Poesio, M., Delmonte, R., Bristot, A., Chiran, L., and Tonelli, S. (2004a). The VENEX corpus of anaphora and deixis in spoken and written Italian. *University of Essex*.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004b). Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 143. Association for Computational Linguistics.

Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Postolache, O., Cristea, D., and Orasan, C. (2006). Transferring coreference chains through word alignment. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Pradhan, S., Luo, X., Recasens, M., Hovy, E. H., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, pages 30–35.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

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 the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics.

Pradhan, S. S., Ramshaw, L. A., Weischedel, R. M., MacBride, J., and Micciulla, L. (2007). Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the International Conference on Semantic Computing (ICSC)*, pages 446–453. IEEE.

Rahman, A. and Ng, V. (2012). Translation-based projection for multilingual coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730. Association for Computational Linguistics.

Rasooli, M. S. and Collins, M. (2015). Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338.

Recasens, M. and Hovy, E. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Recasens, M., Hovy, E. H., and Martí, M. A. (2010a). A typology of near-identity relations for coreference (NIDENT). In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010b). SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

Recasens, M. and Martí, M. A. (2010). Ancora-co: Coreferentially annotated corpora for Spanish and Catalan. *Language resources and evaluation*, 44(4):315–345.

Recasens, M., Martí, M. A., and Orasan, C. (2012). Annotating near-identity from coreference disagreements. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 165–172.

Reinhart, T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, 6(1):47–88.

Riester, A. and Baumann, S. (2017). The RefLex scheme-annotation guidelines. Stuttgart: Universität Stuttgart, SFB.

Riester, A., Lorenz, D., and Seemann, N. (2010). A recursive annotation scheme for referential information status. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Ritz, J., Dipper, S., and Götze, M. (2008). Annotation of information structure: an evaluation across different types of texts. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Rodrıguez, K. J., Delogu, F., Versley, Y., Stemle, E. W., and Poesio, M. (2010). Anaphoric annotation of Wikipedia and blogs in the Live Memories corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 157–163.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Sharoff, S. and Nivre, J. (2011). The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Proceedings of the Russian Conference on Computational Linguistics Dialogue*.

Sidarenka, U. (2016). PotTS: The Potsdam Twitter Sentiment Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Souza, J. G. C. and Orăsan, C. (2011). Can projected chains in parallel corpora help coreference resolution? In *Discourse Anaphora and Anaphor Resolution Colloquium*, pages 59–69. Springer.

Spreyer, K. (2011). *Does it have to be trees?: Data-driven dependency parsing with incomplete and noisy training data*. PhD thesis, University of Potsdam.

Spreyer, K. and Kuhn, J. (2009). Data-driven dependency parsing of new languages using incomplete and noisy training data. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 12–20. Association for Computational Linguistics.

Stede, M. (2011). Discourse processing. *Synthesis Lectures on Human Language Technologies*, 4(3):1–165.

Stede, M. (2016). *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*, volume 8. Universitätsverlag Potsdam.

Stede, M. and Neumann, A. (2014). Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 925–929.

Stoyanov, V., Cardie, C., Gilbert, N., Riloff, E., Buttler, D., and Hysom, D. (2010). Coreference resolution with Reconcile. In *Proceedings of the ACL Conference: Short Papers*, pages 156–161. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, volume 2012, pages 2214–2218.

Toldova, S., Azerkovich, I., Grishina, Y., Ladygina, A., Lyashevkaya, O., Roytberg, A., Sim, G., and Vasilieva, M. (2015). Pre-experiments on annotation of Russian Coreference Corpus. *Higher School of Economics Research Paper No. WP BRP*, 35.

Tufiş, D., Ion, R., Ceauşu, A., and Ştefănescu, D. (2006). Improved lexical alignment by combining multiple reified alignments. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Tuggener, D. (2016). *Incremental Coreference Resolution for German*. PhD thesis, University of Zurich.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). Parallel corpora for medium density languages. *Amsterdam studies in the theory and history of linguistic science series 4*, 292:247.

Vieira, R. and Teufel, S. (1997). Towards resolution of bridging descriptions. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 522–524. Association for Computational Linguistics.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC)*, pages 45–52. Association for Computational Linguistics.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics.

Wiseman, S., Rush, A. M., and Shieber, S. M. (2016). Learning global features for coreference resolution. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004.

Wu, H. and Wang, H. (2007). Comparative study of word alignment heuristics and phrase-based SMT. *Proceedings of the MT Summit XI*.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus V. 1.0. In *Proceedings of the Language Resources and Evaluation Conference*.

Zikánová, Š., Hajicová, E., Hladká, B., Jínová, P., Mírovskỳ, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M., and Václ, J. (2015). Discourse and coherence. *From the Sentence Structure to Relations in Text. Institute of Formal and Applied Linguistics*.

Zitouni, I. and Florian, R. (2008). Mention detection crossing the language barrier. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 600–609. Association for Computational Linguistics.