Joint Multilingual Learning for Coreference Resolution

by

Andreea Bodnari

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Author . . . .
Department of Electrical Engineering and Computer Science
May 14th, 2014

Certified by . . . .
Peter Szolovits
Professor of Computer Science and Engineering- MIT CSAIL
Thesis Supervisor

Certified by . . . .
Pierre Zweigenbaum
Senior Researcher, CNRS
Thesis Supervisor

Certified by . . . .
Özlem Uzuner
Associate Professor of Information Studies- SUNY, Albany
Thesis Supervisor

Accepted by . . . .
Professor Leslie A. Kolodziejski
Chair of the Department Committee on Graduate Students

Joint Multilingual Learning for Coreference Resolution

by

Andreea Bodnari

Submitted to the Department of Electrical Engineering and Computer Science
on May 14th, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Natural language is a pervasive human skill not yet fully achievable by automated computing systems. The main challenge is understanding how to computationally model both the depth and the breadth of natural languages. In this thesis, I present two probabilistic models that systematically model both the depth and the breadth of natural languages for two different linguistic tasks: syntactic parsing and joint learning of named entity recognition and coreference resolution.

The syntactic parsing model outperforms current state-of-the-art models by discovering linguistic information shared across languages at the granular level of a sentence. The coreference resolution system is one of the first attempts at joint multilingual modeling of named entity recognition and coreference resolution with limited linguistic resources. It performs second best on three out of four languages when compared to state-of-the-art systems built with rich linguistic resources. I show that we can simultaneously model both the depth and the breadth of natural languages using the underlying linguistic structure shared across languages.

Thesis Supervisor: Peter Szolovits
Title: Professor of Computer Science and Engineering- MIT CSAIL

Thesis Supervisor: Pierre Zweigenbaum
Title: Senior Researcher, CNRS

Thesis Supervisor: Özlem Uzuner
Title: Associate Professor of Information Studies- SUNY, Albany

Acknowledgments

This thesis would not be possible without my advisors: prof. Peter Szolovits, prof. Özlem Uzuner, and prof. Pierre Zweigenbaum. Their continuous support and guidance has helped me discover new horizons. I would also like to acknowledge the feedback received from the rest of my thesis committee members, prof. Regina Barzilay and prof. Patrick Winston.

The research reported here has been supported by a Chateaubriand Fellowship to study at LIMSI in France, and by Research Assistantships at MIT supported by Grant U54 LM008748 (Informatics for Integrating Biology and the Bedside) from the National Library of Medicine and ONC #10510949 (SHARPn: Secondary Use of EHR Data) from the Office of the National Coordinator for Health Information Technology.

Part of the work presented in this thesis came to reality due to the collective efforts of the three annotators Julia Arnous, Aerin Commins, and Cornelia Bodnari, and with the help of Cosmin Gheorghe who helped analyze the annotation results. For their insightful feedback and discussions, I would like to thank Tristan Naumann, Victor Costan, and Rohit Joshi.

I was very fortunate to be surrounded by brilliant lab-mates and extraordinary friends: Amber, Fern, Marzyeh, Rohit, Tristan, Ying, and Yuan. Thank you for making our lab feel like home.

Last but not least, I would like to thank my family and friends for their caring and support. I dedicate this thesis to my parents and my sister who have unconditionally loved and believed in me.

Meae familiae, ignoscite mihi relinquenti.

Omne opus meum dedicatum est vobis, magno cum amore, ex toto corde meo.

Per donum Dei, credo.

Fortitudo mea de caelis est. Deo adiuvante, non timendum.

Contents

Chapter 1

Introduction

Natural language is the fundamental mechanism of communication among humans. Even though use of language is a fundamental human skill, we do not yet thoroughly understand language and are unable to fully implement it in automated computing systems. The inherent complexity of natural language makes it difficult for natural language processing (NLP) systems to master both the depth (i.e., comprehensive understanding of a natural language) and breadth (i.e., common level of understanding across natural languages) of language.

Natural language processing systems approach the depth of language understanding incrementally. Each level of understanding is associated with a formal NLP task. NLP systems build up from shallow-level to deep-level tasks in order to generate a complete language understanding system. The more shallow depths include tasks like sentence and token identification, while deeper depths of understanding are concerned with tasks like semantics and pragmatics disambiguation. Despite exhibiting high accuracy for the shallow tasks, NLP systems are challenged by the complex tasks that require a deep understanding of language (e.g., text summarization, word sense disambiguation, coreference resolution).

Regardless of the task, NLP systems do not display a common level of language understanding across natural languages. Some natural languages are more commonly investigated (i.e., English, Japanese, German, French, Russian, and Mandarin Chinese) and consequently have a rich set of linguistic and computational resources available,[81] but most natural languages have had few linguistic resources developed for them. Linguistic resources range from corpora annotated for information like parts of speech, syntactic or dependency structure, or meaning, to automated systems specialized for specific NLP tasks. From the estimated 6000-7000 spoken languages of the world, at least 1000 are present in written form on the Internet,[21] but to my knowledge NLP systems have been developed on 17 languages at most.[59]

The more complex an NLP task is in regards to the depth of language understanding, the harder it is to obtain high accuracy on natural languages, and in particular on resource-poor languages. One solution to overcome the lack of resources and difficulty in obtaining these for resource-poor languages is to transfer linguistic knowledge and NLP systems from resource-rich languages to resource-poor languages. Even though the research community has scaled its efforts in solving complex NLP tasks across multiple languages, the performance of such systems that learn from multiple languages simultaneously (i.e., multilingual learning) lags far behind systems that learn from a single language (i.e., monolingual learning).[73, 62, 16] This is due to both a lack of available annotated resources in multiple languages and to the lack of an infrastructure for automatically porting the relevant linguistic phenomena from resource-rich source languages to resource-poor target languages.

This thesis tackles the problem of creating natural language processing systems that can handle both the depth and breadth of language understanding. Throughout this thesis, the depth of language understanding is reflected by two deep-level NLP tasks (i.e., sentence parsing and coreference resolution), and the breadth of language understanding is represented by scaling the two mentioned NLP tasks to resource-poor languages.

1.1 Motivation

Understanding a written text in a computational setting requires several steps, regardless of the language of interest. Those steps can be broadly differentiated into shallow-level analysis and deep-level analysis. Shallow-level analysis is concerned with pre-processing steps like token and sentence identification, text normalization, and part-of-speech tagging. Given a text in the form of a single sentence s = *She ran the marathon yesterday*, an NLP system could perform a combination of pre-processing steps. One possible combination would be to first identify the set of tokens {*she, ran, the, marathon, yesterday*} and then tag each token with a relevant part-of-speech tag, i.e., {*she:PRP, ran:VBD, the:DT, marathon:NN, yesterday:NN*}.

The shallow-level analysis results become substrate to infer deeper understandings of the text. Depending on the type of deep-level analysis, different approaches may be taken. For example, mentions of pre-defined categories (e.g., people, organizations, locations) can be identified together with the hierarchical relations that might exist between them. The identification of relations depends on the accuracy of mention identification, and is a difficult task, as it requires both a syntactic and semantic understanding of the written text.[27] For the sentence s above, many of the hierarchical relations can be extracted from a syntactic hierarchical representation. Figure 1.1 presents a sample hierarchical representation in the form of a parse tree for the sentence s. Given the parse tree, one can infer syntactic dependencies between the subject of the sentence (i.e., *she*), the action being performed (i.e., *ran*), the immediate object of the action (i.e., *marathon*), as well as temporal relations (i.e., *yesterday*).

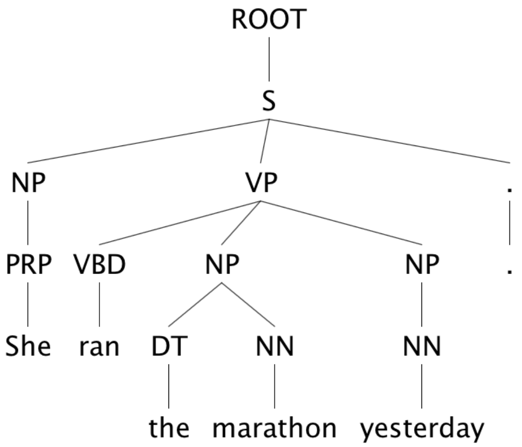

Figure 1.1: Sample parse tree for the sentence “She ran the marathon yesterday”.

The problem of relation identification becomes more complex when relations span sentences. In these cases, the syntactic dependency formalism is usually designed to tackle relations between the tokens within a sentence and not across sentences. NLP systems that perform a comprehensive analysis are required in order to perform parsing based on semantics and pragmatics understanding. In our case, the subject of the above sentence is a generic person referred to by a pronoun *She*, and one would expect a reference to the actual person name to be included earlier in the text. Additional intricacies arise from the ambiguous nature of natural language. For example, in the sentence *I met the girl running on the hill*, the phrase *running on the hill* could refer to either the *girl* or to the location where the subject met the girl. Such ambiguities cannot be explained by syntax alone.

One of the most common relation identification tasks in NLP research is coreference resolution. Resolving coreference is an important step in language understanding[31] and facilitates other NLP tasks like information retrieval, question answering, and text summarization.[6] Coreference resolution determines whether two expressions, often referred to as named entities in NLP literature are coreferent, that is, linked by an *identity* or *equivalence* relation. For example, given the sentences "*I met my friend Diana. She ran the marathon yesterday.*", the named entities *Diana* and *she* are equivalent because they refer to the same person. Coreference resolution systems build on shallow-level and syntactic parsing systems. Resolving coreference in a computational setting involves a two-step process that initially requires the identification of the named entities of interest. The identified named entities are then linked based on equivalence relations. Having a sound coreference resolution system impacts many areas of research that are dependent on automated text processing. For example in the medical field, patient information is largely stored in the form of free text, and coreference resolution can help extract information about disease development, patient progress, or medication effectiveness. The judicial system is another example, as coreference resolution systems may help automate the process of offender profiling from written police reports.

Despite ongoing research efforts in coreference resolution, the current state-of-the-art multilingual NLP solutions still have important challenges to overcome.[56, 39]

Past competitions on coreference resolution acknowledged the need for both better external knowledge integration and better techniques for coreference resolution.[73, 16, 87] The 2010 SemEval[73] and the 2012 CoNLL[16] shared tasks on multilingual coreference resolution also acknowledged inferior performance of coreference resolution systems on non-English languages. The most extensively researched language is English. Proposed solutions range from rule-based to machine learning and probabilistic methods, and from supervised to semi-supervised and unsupervised learning settings. Even still, a recent systematic evaluation of state-of-the-art approaches in coreference resolution reported a top performance of 0.77 F-measure on English.[1] The performance levels are low mainly due to the inability of automated systems to account for contextual cues, perform disambiguations, or generalize from training instances in the supervised and semi-supervised settings.[44, 4] Coreference resolution is a difficult task as it also requires a good understanding of syntax, semantics, and pragmatics.[30] These three foundation layers required for coreference resolution are not equally well understood and documented across all languages.

Recent research showed that supervised and unsupervised NLP systems trained in a multilingual setting perform better than systems developed in a monolingual setting when evaluated on individual languages.[**?**, 79] This finding was explained by the fact that some languages express ambiguities at levels where other languages are very clear. Using a multilingual context should help better clarify those ambiguities.[79] Harabagiu and Maiorano[32] show that a supervised coreference resolution system trained on a bilingual English-Romanian corpus outperforms monolingual English and Romanian coreference resolution systems. Their system learns independently on each language but makes use of aligned coreference relations in the other language in order to improve on coreference resolution performance. Snyder et al.[79] experimented with unsupervised multilingual grammar induction. They showed a significant improvement of a bilingual system that uses bilingual cues to improve over a monolingual baseline. These initiatives and ideas support the development of NLP systems that can benefit from the multilingual context and solve NLP tasks across multiple languages with increased accuracy.

Community efforts helped the development of NLP tools in new languages, but in general these tools were monolingual systems requiring expertise in the target language and availability of human annotations. The NLP community has acknowledged the need for language processing resources that solve more complex language phenomena within a larger number of the world's languages. Actions taken in this respect range from the development of machine-translation systems,[37] preparation of annotated resources in multiple languages,[9, 64] task-specific NLP systems targeting multiple languages,[32, 41] and the organization of shared tasks that address multilingual NLP tasks.[9, 64, 70] More recent initiatives propose to take advantage of the plethora of text available in different languages and develop NLP systems that can generalize linguistic properties from source languages to a target resource-poor language.[59, 22]

Based on previous research efforts and the current challenges in coreference resolution and multilingual processing, an immediate step needs to be taken in bridging the gap on multilingual deep-level language processing.

1.2 Contributions

The work presented in this thesis advances NLP research in multilingual settings by focusing on the design of NLP systems that can generate a deep language understanding across a broad range of natural languages. In order to reach this goal, I first show how multilingual syntactic parsers can be trained to selectively learn deep-level syntactic phenomena from several source languages, and successfully apply those phenomena to resource-poor target languages. I then create a corpus with deep-level annotations for multiple languages (i.e., named entities and coreference resolution). I use the syntactic information learned from the multilingual syntactic parsers to build an end-to-end coreference resolution system that performs joint named-entity recognition and coreference resolution in a multilingual setting. And finally, I evaluate the end-to-end coreference resolution system on the annotated corpus.

1.2.1 Multilingual structure detection and analysis

I design and implement a linguistic model that captures the universals of syntactic structure shared across languages. My model uses the language universals to learn dependency parsing models for languages for which linguistic resources are sparse. I show that the proposed model can adapt itself based on the language it targets, and that its best performance occurs when the model transfers information at the sentence-level. My model performs better than or as well as state-of-the-art parsing systems.

This work is grounded on the hypothesis that languages share structural characteristics at a higher level, but differ in the way information is conveyed at a lower level. Linguist Noam Chomsky advocates that the human brain contains a basic set of rules for handling language, and those rules are universal and formalized within an universal grammar.[17] One postulate of the universal grammar theory is that human languages are built on a common fundamental structure. This abstract common structure and the levels of divergence between languages are illustrated by the Vauquois triangle (see Figure 1.2).[89] At the bottom level of the triangle languages differ in regards to their words. As we move upward the triangle, language properties become more universal and get expressed at the syntactic and semantic levels. The syntactic dependencies are more universal as they consider grammatical relations, which hold more often across languages. The semantic dependencies are also transferable as the meaning of a sentence will usually be carried across any language, just in a different form. At the highest level of the triangle we have the *interlingua* abstractization, the common structure shared by all languages. The Vauquois triangle is used as an inspiration in the design of my multilingual structure detection algorithm.

I adopt the idea that most natural languages are comprehended through a fundamental structure and aim to predict a syntactic structure for a resource-poor target language by using the shared language structure. The identified structure form can (*a*) facilitate the transfer of NLP techniques across languages, (*b*) help clarify ambiguities across languages, and (*c*) ultimately help the advancement of NLP research

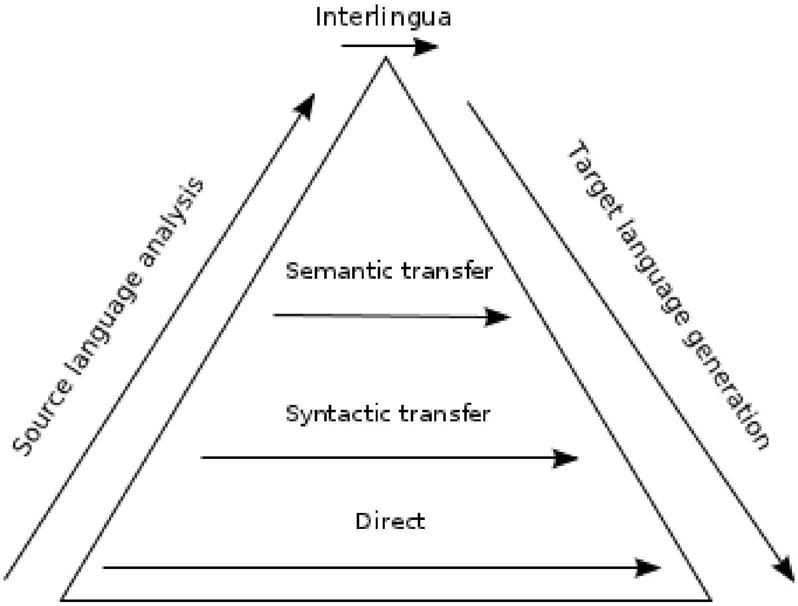

Figure 1.2: The Vauquois triangle

across and within multiple languages. I perform analysis on the learned structure and try to explain commonalities and dissimilarities across languages. Namely, how the computationally induced structure compares to documented linguistic universals.

1.2.2 Multilingual parallel coreference resolution corpora

Motivated by the lack of available annotations in multilingual settings for semantically-equivalent text, I develop a new corpus of multilingual annotations on top of parallel corpora. My goal is to annotate semantically equivalent (i.e., parallel) text in different languages in order to have a better informed analysis over the system errors and the difficulty of the NLP task on each language. Such a corpus contains reference annotations that can guide automated systems to gain a deep-level understanding of language. Because the corpus consists of documents with semantically equivalent content across the different languages, one can analyze the degree of variations and idiosyncrasies present between languages. This is the first attempt at generating multilingual annotations for coreference resolution on parallel multilingual documents.

I choose a set of three named-entity categories commonly used in NLP literature

- person, organization, and location - and annotate parallel text in the form of parliamentary proceedings of the European Union[84] for both named-entity recognition and coreference resolution. I present inter-annotator agreement results together with a discussion on the difficulty of the human annotation task.

1.2.3 Multilingual end-to-end coreference resolution

I develop a multilingual system for joint named-entity recognition and coreference resolution, that builds on the multilingual syntactic structure mentioned above. My system is the first attempt at jointly modeling named-entity recognition and coreference resolution in a multilingual setting.

In order to overcome the challenges faced by current coreference resolution systems in semantic understanding, I design a system that integrates multilingual syntactic structure learning with named-entity recognition and coreference resolution. Unlike most state-of-the-art systems, my model jointly learns the named entities and solves coreference on the learned entities, without relying on gold standard annotations for the target language. This complex system benefits from the integration of soft linguistic constraints that characterize syntactic and grammatical structures. I experiment with manually specified shallow-level constraints, and show that the induced constraints help guide the model learning.

1.3 Thesis Overview

In **Chapter 2** I open with a discussion on multilingual syntactic parsing. I also present my multilingual parsing system and the relevant experiment results together with an analysis on system performance across languages. I continue in **Chapter 3** by presenting my work on creating a multilingual-annotated corpus for end-to-end coreference resolution. In **Chapter 4**, I give a literature review on multilingual coreference resolution, I describe my implemented solution for the task, and discuss its performance. I conclude in **Chapter 5** with some final thoughts and directions for future work.

Chapter 2

Multilingual syntactic parsing

2.1 Chapter overview

I present a linguistic model that can capture the cross-lingual common properties of syntactic structure, represented using dependency graphs. I hypothesize that within each language and across languages there exists a wide degree of syntactic flexibility, while certain syntactic properties are maintained across languages. Even though some languages are related and grouped into language families, the best predictor source language for a target language does not necessarily come from its family (see discussion in 2.7). The model discussed below can identify the best predictor source language or the set of predictor source languages, and customize the predictor selection given a target language or given a sentence in a language. The model is not tied to a certain configuration of the source or target language, and can transfer syntactic structure across languages from different language families.

The next two sections give an overview on multilingual syntactic parsing (section 2.2) and outline related work in the field of syntactic parsing (section 2.3). Next, I introduce the multilingual syntactic parsing model (section 2.4). I then present the experimental setup and model results (section 2.5), and I discuss the linguistic phenomena discovered by my model (section 2.6). Finally, I conclude with the main contributions brought by my model (section 2.8).

2.2 Introduction

Language has long been studied by scholars, initially from a theoretical standpoint in the field of linguistics, and later from a computational standpoint in the field of computational linguistics. Linguistics attempts to systematically define formal rules that characterize three aspects of the human language: language form, language meaning, and pragmatics. Similarly, computational linguistics models the same language aspects through statistical or rule-based computer algorithms.

To understand how language is generated, it is first important to understand its structure. Linguists have defined language structure by closely examining the rules followed by native speakers of a language. The formalization of this set of rules creates the language grammar, with specific subfields of morphology, syntax, and phonology. Language structure has been studied across a set of multiple languages, with the goal of defining similarities and dissimilarities between languages based on structural properties.[33] Language typology focuses on the structural properties specific to individual languages (e.g., different types of word order used in forming clauses), while language universals focuses on what common properties are shared by languages (e.g., all languages have nouns and verbs). Language universals are common to all languages, regardless of their typological classification.

Typological properties of languages have been summarized in the World Atlas of Language Structures (WALS),[33] a large online database of the structural properties of languages manually gathered by a team of 55 scholars. WALS covers over 2676 languages in 208 language families, using 144 linguistic features in 11 different categories. It presents the distribution of linguistic features for features categories that span over all areas of language structure: phonology, morphology, nominal categories, nominal syntax, verbal categories, word order, simple clauses, complex sentences, lexicon, sign languages, and others. The first eight categories are concerned with structural properties of grammar, while the last three categories are more varied.

The set of WALS linguistic categories and the number of features included by each category are presented in Table 2.1. The feature count per feature category varies from 28 in the nominal categories to two in the sign language category.

Feature Category	Counts	Sample Features
Phonology	19	Syllable Structure,
		Consonant-Vowel Ratio
		Consonant types (e.g., uvular, glottalized, lateral)
Morphology	10	Zero Marking of A and P Arguments
		Prefixing vs. Sufixing in Inflectional Morphology
Nominal Categories	28	Numeral Classifiers
		Number of Genders
		Definite Articles
Nominal Syntax	7	Noun Phrase Conjunction
		Adjectives Without Nouns
		Possessive Classification
Verbal Categories	16	The Optative
		The Past Tense
		The Prohibitive
Word Order	17	Order of Object and Verb
		Prenominal Relative Clauses
Simple Clauses	24	Passive Constructions
		Negative Morphemes
		Predicative Possession
Complex Sentences	7	"When Clauses"
		"Want" Complement Subjects
Lexicon	10	Finger and Hand
		Red and Yellow
Sign Languages	2	Question Particles in Sign Languages
Other	2	Para-Linguistic Usages of Clicks

Table 2.1: Description of the WALS database: feature category names with associated feature counts and sample features.

Typological features, and specifically the WALS database, have been used for investigating language universals. Georgi et al.[25] compared languages grouped based on similar diachronic patterns (i.e., phylogenetic groups) to language groups created by automated clustering methods built on the WALS features. The typologically-based language groups looked different from the phylogenetic groups, but performed better when used to predict characteristics of member languages. Typological features were also used to induce better language models across languages by Naseem et al.[59] These studies support further investigation of typological properties shared across languages for multilingual modeling, and specifically finding a methodology for automatically identifying the shared typological features for a set of languages.

2.2.1 Language structure representation

Natural language processing systems assume the existence of a correct representation for language structure. The most commonly used structural representation in computational models is the bag-of-words model,[76] but this model does not capture relationships between words or discourse segments. Alternative structural representations are more complex and harder to generate. They usually evolve as an attempt to explain aspects of the human language. Such alternative representations include:

Sequence prediction:[46, 36] The written form of the natural language has an inherent sequential structure, as conveyed by its representation as a sequence of characters or tokens. This sequential form underlies the language model: a probability distribution over the next symbol given by the symbols that precede it. This language structure representation is the underlying foundation of many NLP systems.

Sequence segmentation:[55] Sequences of written text can be broken down into contiguous parts called segments. Two common choices for segment types are words and sentences. Identifying the words of a piece of text (i.e., tokenization) or sentences of a discourse are basic pre-processing steps assumed by most NLP systems.

Sequence labeling: Because data are sparse, attempts are made to identify smaller classes into which words can be grouped. Such methods that create refined word classes are stemming, lemmatization, and word chunking[85] which map each word or group of words to a simpler vocabulary. Another level of classes that abstract away from the raw words are syntactic classes, or parts of speech (POS).[91] Mapping words to POS involves assigning each word a POS tag depending on the context it belongs to. Another type of sequence labeling problem is that of named entity recognition, where each token gets labeled based on its membership in a named entity category.[58]

Syntax:[69, 28] The previous steps output stand-alone linguistic information that

> must be further manipulated to obtain relationships between various sentence and discourse segments. Syntax rules help label segments that are related to each other via a process called parsing. This is a salient representation of language structure as it helps build a higher order representation of natural language: semantic representation or discourse representation.

Syntax emerges as the more complex and complete representation model of language structure. Even though syntax cannot incorporate all the contextual information relevant for discourse understanding, it is the closest language model with a discrete structure that can be defined in mathematical terms and manipulated in polynomial time by a computer program. Throughout this work I define language structure as the syntactic dependency structure and aim to characterize and identify language syntactic universals.

2.3 Related work

The field of syntax has been documented by grammarians since antiquity.[75] Syntax focuses on formalizing the rules by which phrases and sentences are formed. Syntax understanding is a building block in NLP research, as more complex NLP tasks like machine translation rely on the availability of a syntax analyzer (i.e., parser). Linguists have taken two approaches to formalizing syntax, and defined the constituency[2] and dependency formalisms.[34] In the constituency formalism, a grammar is defined by syntax rules expressed at the level of syntactic phrases (i.e., segments which function as single units in the syntax of a sentence). This formalism works well for languages like English that follow stricter word-order rules, but for languages with more free word-order rules like Russian, constituency grammars are not easily applicable.[18] On the other hand, the dependency formalism defines grammars based on binary asymmetric relations between two words, where one word represents the head and the second word represents a dependent.[63]

The dependency formalism generates a one-to-one correspondence between the words in a sentence and the structural representation of the sentence. Specifically, for each word in the sentence there is only one node in the sentence structure that corresponds to it. In contrast, the constituency formalism generates a one-to-one or one-to-many correspondence between words and nodes in the structural representation of a sentence, and consequently it requires a larger grammar size. Early approaches to dependency parsing were hand-crafted grammars,[83] but machine learning approaches quickly replaced the traditional methods. Machine learning methods developed for dependency parsing were preferred over the constituency parser methods, since the constituency parsers were computationally more demanding. For comparison, given the sentence *She ran the marathon yesterday* with the constituency parse tree depicted in Figure 1.1, one can see that the dependency parse tree presented in Figure 2.1 has a more compact form, specifically in terms of the maximum path length from the parent to its children.

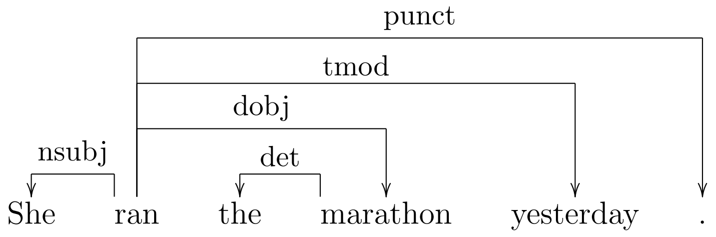

Figure 2.1: Dependency parse tree for the sentence *She ran the marathon yesterday.*

For each word in a given sentence (i.e., *child*), the dependency parsing model generates $(child, head, DEPREL)$ triplets, where $DEPREL$ characterizes the type of dependency relation between the *child* and one other word of the sentence, i.e., the *head.* In Figure 2.1, one such triplet would be *(the, marathon, det).* A word can be included in a triplet as a *child* one time only, but it can take the role of *head* for any number of *child* words in the sentence. The $DEPREL$ labels are formally specified in the annotation guidelines of a dependency treebank, and consequently vary based on the treebank and language of interest. Sample $DEPREL$ labels from the Penn Treebank[50] (a collection of English documents annotated for syntactic relations) and from the UniDep Treebank[52] (a multilingual collection of documents annotated with language universal dependencies) are included in Table 2.2. In this work, I focus on representing syntactic parse trees using the dependency formalism.

Penn Treebank	
$Modifier$	Description
vmod	Reduced non-finite verbal modifier
nmod	Noun modifier
cc	Coordination

UniDep Treebank	
$Modifier$	Description
adp	Adposition analyzed as dependent of noun (case marker).
aux	Auxiliary verb (dependent on main verb), including innitive marker.
cc	Coordinating conjunction (dependent on conjunct).
vmld	Verbal modier (underspecied label used only in content-head version).

Table 2.2: Sample modifier $DEPREL$ values from the Penn and UniDep Treebanks.

Approaches to generating dependency parse trees are grouped into two categories: graph-based and transition-based. The two approaches learn probabilistic models for scoring possible dependency trees for a given sentence. The difference between the two approaches comes in the way in which they decompose the possible dependency tree during scoring. The graph-based parsers decompose the dependency tree into either individual arcs scored separately (i.e., arc-factored models) or higher order factors in which several arcs are treated as a unit, and scoring is performed at the unit level.[24, 96] Higher order parsers have better accuracy but also involve higher computational cost. The research community has devoted time to developing approximation methods that can reduce the computational cost while maintaining a minimum decrease in parser accuracy (e.g., structured prediction cascades,[92] cube-pruning,[13] dual-decomposition[43]). The transition-based parsers build the dependency tree incrementally, through the application of a small set of parser actions. A pre-trained classifier dictates each parser action. The most commonly used transition method is the one proposed by Covington,[19] but additional methods were proposed by Yamada and Matsumoto,[94] and Nivre and Nillson.[65] Properties of transition and graph-based parsers were combined into a single parser and it was shown that the combination of the two is beneficial, as each parser type makes different errors.[51]

2.3.1 Multilingual syntactic parsing

Language parsers were initially developed for English,[15] and subsequently for other languages like Japanese,[45] Turkish,[67] German,[23] Spanish,[20] and French.[3] The 2006 and 2007 CoNLL Shared Task proposed the development of a standard dependency corpus, evaluation scheme, and state-of-the-art analysis for syntactic dependency parsing in multiple languages. The 2006 CoNLL Shared Task used 13 dependency treebanks (i.e., Arabic, Bulgarian, Chinese, Czech, Danish, Dutch, German, Japanese, Portuguese, Slovene, Spanish, Swedish, Turkish), while the 2007 CoNLL Shared Task included 10 treebanks (i.e., Arabic, Basque, Catalan, Chinese, Czech, English, Greek, Hungarian, Italian, Turkish). The system performance in the supervised training setting for the two shared tasks ranged from 80.95 label attachment score[1] on English to 68.07 on Basque for the 2007 Shared Task, and 85.90 on Japanese to 56.00 on Turkish for the 2006 Shared Task. The 2006 and 2007 Shared Tasks showed that even in the supervised setting, automated parsing systems cannot feasibly solve the syntax parsing task for some of the languages.

In the setting of the two shared tasks, the availability of annotations and linguistic resources has contributed to the performance of the best dependency parsing systems. But developing the same level of resources for all natural languages is too time consuming (and to some extent impossible for extinct languages). Thus, the research community has started investigating methods for transferring knowledge from source languages with more available resources to target languages without available resources. One such method of knowledge transfer is grammar induction from one or more source languages to a target language with the aid of parallel corpora. Wu[93] presents the concept of bilingual language modeling for sentence-pair parsing on Chinese and English sentence-pairs. In his model, the parser takes as input a sentence pair (s_A, s_B) rather than a sentence, where s_A and s_B come from different languages, and generates dependency predictions using a specialized form of a grammar that models sentence pairs simultaneously (i.e., the transduction grammar[47]). Burkett and Klein use a similar idea of parser modeling on parallel text to show that

[1] See Section 2.3.2 for a detailed explanation of evaluation metrics for dependency parsers.

parallel information substantially improves parse quality on both languages.[10] The authors employ a maximum entropy parsing model that operates on English-Chinese sentence-pair inputs. Their model maximizes the marginal likelihood of training tree pairs, with tree alignments treated as latent variables. Snyder et al.[79] show that even in unsupervised settings the presence of parallel bilingual information helps increase the performance of a syntax parser on both languages. The authors propose a generative Bayesian model that uses observed parallel data together with a combination of bilingual and monolingual parameters to infer parse trees for the languages of interest.

Recent work has shown the possibility of transferring syntactic information across languages in the absence of parallel data. Such transfer is possible due to the dependency structure universals present in languages. Bern and Klein[7] investigate multilingual grammar induction by extracting language similarities from patterns of language evolution. The authors consider six Indo-European languages: English, Danish, Portuguese, Slovene, Spanish, and Swedish. Their model couples together different languages in a way that resembles knowledge about how languages evolved. In general, models that couple more languages have better performance results than models that only consider pairs of languages. When their model encodes properties of language evolution within the same language family as well as across language families, it obtains the best error reduction in dependency generation.

Naseem et al.[60] encode the language-structure universals in the form of handwritten rules. Those rules guide grammar induction on the target language using language-structure information available on a set of source languages. The model proposed by the authors takes as input a corpus annotated with universal part-of-speech tags and encodes a declarative set of dependency rules defined over the universal part-of-speech tags. The dependency rules are universal across languages and can help disambiguate syntactic ambiguities that are difficult to learn from data alone. In addition, the model incorporates the intuition that languages present their own idiosyncratic sets of dependencies by requiring the universal rules to hold in expectation rather than absolutely. Naseem et al. test their model on six Indo-European languages from three language families: English, Danish, Portuguese, Slovene, Spanish, and Swedish. The evaluation results on those languages show that universal rules help improve the accuracy of dependency parsing across all languages. The model outperforms other unsupervised grammar induction methods, and specifically the model of Berg and Klein,[7] but has the downside of relying on a manually specified set of universal dependency rules.

Cohen et al.[14] define a method for capturing the contribution of a set of source languages to the target language. The authors use the annotated data from source languages to generate dependency parsers for resource-poor target languages. Their approach is to initialize a parsing model for the target language using supervised maximum likelihood estimates from parsing models of the source languages. The authors use maximum likelihood over interpolations of the source language parsing models in order to learn a parsing model for the target language using unannotated data. The model is evaluated over a set of four source languages – Czech, English, German, and Italian – and ten target languages: Bulgarian, Danish, Dutch, Greek, Japanese, Portuguese, Slovene, Spanish, Swedish, and Turkish. Yet, the authors only report model evaluation results on sentences of length at most ten, and cannot be easily compared to other state-of-the-art models.

Søogard[80] use the idea of source language weighting to represent the similarity of data available in a source language to data available in a target language. The authors rank the labeled source data from most similar to least similar to target data and run experiments only with a portion of the source data that is similar to the target data. The similarity is computed using perplexity per word as metric. The source data most similar to the target data is used to then train a dependency parser for the target language. The model is evaluated on Arabic, Bulgarian, Danish, and Portuguese. It generates results comparable to more complex projection-based cross-language adaption algorithms for dependency parsing.

Naseem et al.[59] propose a more generic method that can generate dependency parsers for target languages by learning relevant syntactic information from a diverse set of source languages. Their model learns which syntactic properties of the source languages are relevant for the target language. It first generates the distribution of dependents for each part-of-speech tag using information available in the source languages. The distribution of dependents is generated independent of the order the dependents have inside the sentence. Then, it orders the dependents to match the word-order of the respective target language. The dependency generation step is universal across languages, and is learned in a supervised fashion across all source languages. The ordering of the syntactic dependencies is only learned from languages with word-order properties similar to the target language. The authors report results on 17 languages from the CoNLL 2006/2007 corpora. The largest improvements compared to state-of-the-art systems emerge on non Indo-European target languages.

Täckström et al.[82] study multi-source transfer of dependency parsers for resource-poor target languages. The authors adapt feature-rich supervised parsers discriminatively trained on source languages to a specific target language. Their target language model incorporates parameters from source language models based on typological traits. This transfer of parameters from source languages is possible due to the decomposition of parser features into language-specific and language-generic sets. Their resulting model outperforms the model of Naseem et al.[59] on 15 out of 16 languages.

Another approach for multi-source transfer of unlexicalized dependency parsers for resource-poor target languages comes from McDonald et al.[54] The authors propose a simple method for transferring unlexicalized dependency parsers from source languages with annotated data available to target languages without annotated data. Their contribution is to show that unlexicalized parsers transferred from source languages can perform better than their unsupervised counterparts on target languages. In addition, the authors use constraints extracted from parallel corpora to project a final parser on the target language. The model is evaluated only on Indo-European languages, with the set of source languages containing English, Danish, Dutch, German, Greek, Italian, Portuguese, Spanish, and Swedish and the set of target languages containing Danish, Dutch, German, Greek, Italian, Portuguese, Spanish, and Swedish. The final model achieves state-of-the-art performance on eight Indo-European target languages.

My goal is to take advantage of multilingual learning in order to improve performance on dependency parsing and provide new resources for less investigated languages. I improve on the current state-of-the-art systems by proposing a new methodology for transferring syntactic knowledge from source to target languages at the more granular level of a sentence, as opposed to applying the knowledge transfer at the language level. Similar to previous approaches, my model is delexicalized. It does not assume the existence of parallel text between source languages and the target language. My model automatically learns which source language or set of source languages is most similar to the sentences of the target language. My approach resembles the approach of Søogard[80]. I innovate by leveraging the properties shared between the source and target languages at a sentence level. Instead of specializing the parsing model to the target language, I identify the source parsing model that is expert at performing parsing on a structure similar to the one available in the target language. Details of this approach are outlined in section 2.4.

2.3.2 Evaluation metrics

The performance of dependency parsers is evaluated using three standard metrics: labeled attachment score (LAS), unlabeled attachment score (UAS), and label accuracy.

Given a set of triplets $(child, head, DEPREL)$ predicted by a dependency parser, LAS computes the fraction of *child* words that are assigned to the correct *head* and labelled with the correct DEPREL. UAS considers the fraction of *child* words that are assigned to the correct *head*, and the label accuracy represents the percentage of words with correct DEPREL tags.

$$LAS = \frac{\text{\# of words with correct } head \text{ and DEPREL}}{\text{\# of words}} \tag{2.1}$$

$$UAS = \frac{\text{\# of words with correct } head}{\text{\# of words}} \tag{2.2}$$

$$\text{Label Accuracy} = \frac{\text{\# of words with correct DEPREL}}{\text{\# of words}} \tag{2.3}$$

2.4 Model overview

Motivated by recent advances in multilingual parsing, I develop multilingual parsing models that make use of syntactic universals. Specifically, I employ a multilingual learning setup on dependency parsing and generate parsing models for languages for which linguistic resources are sparse. The main focus is to learn linguistic structural properties from a set of source languages constrained by the target language of interest, and to use the linguistic structural properties to induce dependency parsers on the target language. As shown by recent research, genetically related languages might not be typologically related,[25] thus using language family as a criteria for relatedness would not suffice and a more informed model for syntactic relatedness is required.

In general, the transfer of information across languages relies on linguistic layers that can be mapped across languages. The word-layer cannot be easily mapped as it requires machine translation and semantic understanding. Yet, we can relatively easily map word categories (i.e., parts of speech) between languages.[68, 97] My model assumes the existence of language-specific POS gold standard annotations and uses the method proposed by Naseem et al.[60] to map from language-specific POS tags to universal POS tags. The resulting model is unlexicalized as it does not consider the word-level information available for each language.

The model takes as input a set of *source* languages for which gold standard dependency parse and POS annotations are available, as well as a *target* language for which a dependency parser needs to be learned. The target language has associated gold standard POS annotations. The model builds a parser for the *target* language by selectively learning from each *source* language the syntactic universals applicable to the *target* language.

I design and experiment with several multilingual parsing algorithms, and try to identify the right granularity at which to transfer the syntactic knowledge. I investigate:

1. *Language-level expert voting*: In line with linguistic theories, I hypothesize that some linguistic phenomena are shared across language subsets. In such a setup, knowledge is transferred from a set of source languages $S = \{L_1, L_2, ..., L_n\}$ to a target language T. The gold standard of the source languages is used to collectively train a dependency parsing model, which will then perform dependency parsing on the target language by leveraging the shared syntactic properties.

2. *Sentence-level expert voting*: I hypothesize that each unit of communication (i.e., sentence) has its own set of linguistic phenomena that it respects, and linguistic knowledge could be better transferred at the level of a sentence. Consequently, the language-level transfer of linguistic properties is not sufficiently granular for capturing the linguistic diversity inherent in natural languages. The intuition is that each sentence within a target language might have a different source language that it most resembles, and selecting a single set of languages to perform dependency parsing on the target sentences could overlook the sentence-level structural diversity. For languages that have multiple linguistic influences, such as English, I expect that individual sentences might derive from language structure in either another language from the same language family (e.g., German, Dutch) or from a language from a different language family. For languages that do not present close neighbors within the language set of the working corpus, I expect less diversity in the set of source languages that are selected by sentence-level expert voting.

My model builds on an existing set of source languages $S = \{L_1, L_2, .., L_n\}, n \geq 1$ for which part-of-speech $POS = \{pos_{L_1}, pos_{L_2}, .., pos_{L_n}\}$ and dependency parsing gold-standard annotations $GS = \{gs_{L_1}, gs_{L_2}, .., gs_{L_n}\}$ are available. The goal is to identify a method for selectively-sharing syntactic universals between $source = \{L_1, L_2, .., L_k\} \subseteq S, k \geq 1$ and a target language $T \notin S$. The target language has available gold standard pos_T annotations, but no dependency parse information.

2.4.1 Language-level expert voting

My goal is to transfer syntactic knowledge learned on a subset of source languages to a target language. In order to transfer syntactic knowledge most relevant to the target language, I hypothesize that the model first needs to identify the languages which are syntactically closer to the target language. Syntactically close languages should exhibit similar linguistic phenomena that can be automatically transferred to the target language. Discovering the degree of syntactic relatedness is done in a stepwise fashion:

- **Step 1** For each language L_k in the set of source languages S, I first train a dependency parser on gs_{L_k} and generate a language specific parsing model m_k. The language-specific parser is generated using the MST Parser,[53] a state-of-the-art dependency parser proved to generalize well to other languages in addition to English in a monolingual setting (i.e., when working on individual languages).

- **Step 2** I use the universal syntactic properties incorporated in each language model m_k in order to parse the target language of interest. Specifically, I apply each source dependency parsing model $m_k, k = 1..n$ to the test set of the target language T. I investigate the degree to which language models can be transferred across languages without additional adaptation.

- **Step 3** I hypothesize that syntactically close languages could share more similarities, and aim to design a linguistic similarity metric to guide the transfer of parsing models from the source language to the target language. I develop a pairwise similarity metric $LS(L_k, T)$ to characterize the degree of syntactic similarity between a source language L_k and the target language T. The LS metric provides a ranking r_T of the source languages in respect to the target language. I employ r_T to decide which source language L_k can positively contribute to the parsing model m_T of the target language T.

A graphical description of the system design is included in figure 2.2.

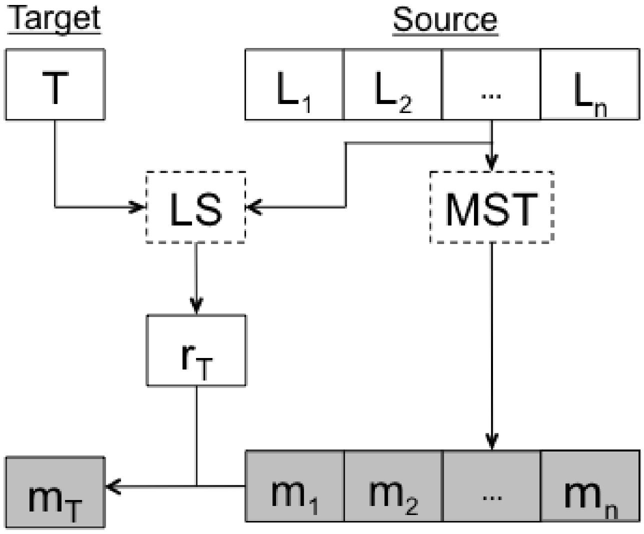

Figure 2.2: Language-level expert voting model design

Language pairwise syntactic similarity

I investigate two syntactic similarity functions computed over language pairs. Both similarity functions rely on external linguistic expert knowledge and output the degree of similarity between two input languages.

1. *Best predictor*: I use the WALS *word order* features to decide the degree of similarity between two languages. The word order features are equally weighted. For each (L_{source}, T) language pair, I compute the degree of syntactic relatedness ls_{WALS} as the percentage of common WALS word order feature values. I select the source language with the largest ls_{WALS} value as the syntactically closest source language candidate and refer to it by $L_{best\_predictor}$. I hypothesize that the parsing model of $L_{best\_predictor}$ should contain syntactic properties that are also present in the target language, and could be directly applied to the target language without further adaptation. Finally, I use $m_{best\_predictor}$, the parsing model of $L_{best\_predictor}$, to parse the test set of the target language.

   In the rest of this chapter I refer to the best predictor model $m_{best\_predictor}$ as the $Predictor_{WALS\_BEST}$ model.

2. *Weighted predictors*: I investigate whether the combined parsing predictions of several source languages could outperform the strict selection of a closest source language. I order the source languages in terms of their ls_{WALS} degree of syntactic similarity to the target language, and select the ω top languages.

   Given the target language test set $test_T = \{s_1, s_2, .., s_T\}$, where each s_k is a sentence in the test set, I apply the parsing model of the selected s languages to $test_T$ and record the predicted dependency trees on each sentence. For each sentence $s_k = \{word_1, word_2, ..., word_{n_{s_k}}\}$ I compute the set of child-parent attachment probabilities $\{p(word_i, word_j)\}, i \leq n_{s_k}, j \leq n_{s_k}, i \neq j$ based on the dependency predictions made by ω source languages on the sentence s_k. The dependency probability $p(word_i, word_j)$ represents the percentage of the ω source languages that generated the $word_i \rightarrow word_j$ dependency. I employ the attachment probabilities in a top-down parsing algorithm and generate a valid parse tree for the sentence s_k.

   I investigate what is an optimal choice for the number of top languages ω such that the contributions added by each language positively contributes to the final dependency parser performance of the target language T.

   In the rest of this chapter I refer to the weighted predictor model as the $Predictor_{WALS\_VOTED}$ model.

2.4.2 Sentence-level expert voting

I observe that the best predicting source language varies based on the target sentence on which predictions are made. Thus, I develop an algorithm that can identify the best predictor source language given the target language T and the target sentence s_T. In order to identify whether a source dependency parser could correctly predict the target sentence, I aim to identify the degree of similarity between the source sentences and the target sentence. This similarity metric should capture the probability of having the structure of the target sentence generated by the source language.

I define similarity of a sentence pair (s_{S_k}, s_T) by looking at its POS-bigram language model. Specifically, I compute the sentence-pair similarity over the universal POS transition probability matrixes of the two sentences s_{S_k} and s_T. I experiment with both KL divergence and cosine similarity as metrics for sentence-pair similarity. I choose the KL divergence as a final sentence-pair similarity measure due to better performance results obtained when integrating the KL divergence metric in the final dependency parser. For a given source language S_k and a target language T, I compute:

$$C_{S_k}(s_T) = \{(s_{S_k}, s_T), s_{S_k} \in S_k | KL(s_{S_k}, s_T) \geq t\} \tag{2.4}$$

where t is a set threshold.

The sentence pair similarity metric allows for ranking source languages based on their similarity to the target sentence, instead of their similarity to the target language alone. I select the parsing model m_{S_k} of the source language S_k with the $max(|C_{S_k}|)$ as the best source candidate for parsing the target sentence s_T.

I investigate the performance of the sentence-level parsing model when the ω_{KL} top KL-similar source languages are used to generate a final target parsing model. The voting model does not outperform the best source parsing model discussed in the paragraph above. I hypothesize that this behavior is due to the large percentage of sentences that are predicted mainly by a single sentence (see discussion in Section 2.6.5). Consequently, I only present the results for the best source parsing model.

In the rest of this chapter I refer to the sentence-level parsing model as the $Predictor_{KL\_BEST}$ model.

2.5 Experiments

2.5.1 Corpus and Annotations

The models I present are developed on top of two dependency treebanks. The first treebank is annotated with language-specific annotation guidelines,[9, 64] while the second is annotated with language universal annotation guidelines.[52] In order to compare my models against results presented in the literature, I evaluate them against all-length sentences and the subset of at most 10-length sentences from each corpus.

Language specific dependency treebank

I work with the dependency corpora released by the 2006 and 2007 CoNLL Shared Tasks.[9, 64] There are 19 available languages from 10 language families: Semitic (viz., Arabic), Sino-Tibetan (viz., Chinese), Japonic (viz., Japanese), Slavic (viz., Bulgarian, Czech, Slovene), Germanic (viz., Danish, Dutch, English, German, Swedish), Ural-Altaic (viz., Turkish), Romance (viz., Catalan, Italian, Portuguese, Spanish), Isolate (viz., Basque), Finno-Ugric (viz., Hungarian), and Greek (viz., Greek). The experiments presented in this chapter are run on all available languages.

The corpora contain literary, newspaper and newswire texts in sentence-delimited format, together with manually-annotated POS tags and dependency trees. For languages included in both the 2006 and 2007 CoNLL Shared Tasks, I use only the 2007 CoNLL version of the associated corpus as it contains additional text and annotation fixes.

The sentence count and the average sentence length for the 2006 CoNLL data are presented in Table 2.3, while Table 2.4 presents a description of the 2007 CoNLL data. The largest average sentence length is observed for Spanish and Portuguese in the 2006 CoNLL data, and for Arabic and Catalan in the 2007 CoNLL data. The smallest average sentence length comes from Japanese for the 2006 CoNLL and Chinese for the 2007 CoNLL data. Both the Chinese and Japanese sentence length metrics are computed over the gold standard tokens provided with the corpora.

Universal dependency treebank

I also evaluate my models on the universal dependency (UniDep) treebank developed by McDonald et al.[52] This corpus contains newswire, blogs, and consumer reviews documents. Unlike the 2006 and 2007 CoNLL corpora, the universal dependency corpus is developed using a standardized annotation guideline for all languages. I expect linguistic phenomena to be more clearly transferred when universal guidelines are

	Bulgarian	Danish	Dutch	German	Japanese	Portuguese	Slovene	Spanish	Swedish
Family	Sla.	Ger.	Ger.	Ger.	Jap.	Rom.	Sla.	Rom.	Ger.
Training data									
Sentences	12823	5189	13349	39216	17044	9071	1534	3306	11042
T/S	14.83	18.18	14.6	17.84	8.88	22.8	18.7	27.02	17.34
Test data									
Sentences	398	322	386	357	709	288	402	206	389
Avg. T/S	14.91	18.17	14.46	15.95	8.05	20.37	15.89	27.64	14.54

Table 2.3: Description of the 2006 CoNLL Shared Task data used in this thesis. Language families include Semitic, Sino-Tibetan, Slavic, Germanic, Japonic, Romance, Ural-Altaic. T/S represents the average count of tokens per sentence.

	Arabic	Basque	Catalan	Chinese	Czech	English	Greek	Hungarian	Italian	Turkish
Family	Sem.	Isol.	Rom.	Sin.	Sla.	Ger.	Hel.	F.-U	Rom.	Tur.
Training data										
Sentences	2912	3190	14958	56957	72703	18577	2705	6034	3110	4997
T/S	38.35	15.84	28.80	5.92	17.18	24	24.18	24.84	22.89	11.5
Test data										
Sentences	131	334	167	867	365	214	197	390	249	623
Avg. T/S	39.1	16.1	30.0	5.78	16.03	23.38	24.38	18.83	20.46	12.11

Table 2.4: Description of the 2007 CoNLL Shared Task data used in this thesis. Language families include Semitic, Isolate, Romance, Sino-Tibetan, Slavic, Germanic, Hellenic, Finno-Ugric, and Turkic. T/S represents the average count of tokens per sentence.

used as a starting point for annotation generation. The 2006/2007 CoNLL language annotation guidelines present different annotation schemes for the same underlying linguistic phenomena (e.g., different annotation schema for children of parent-token “and”, when “and” is used as a coordinating conjunction). The universal dependency annotations are available for 10 languages, but for consistency purposes I exclude the Japanese corpus as it is differently tokenized from the CoNLL Japanese corpus. I thus focus on English, French, German, Indonesian, Italian, Korean, Brazilian-Portuguese, Spanish, and Swedish.

In Table 2.5, I describe the sentence count and average sentence length per each language of the universal treebank. The shortest average sentence length is observed for Korean (average of 8.8 tokens per sentence in the test set and an average of 11.15 tokens per sentence in the training set). The largest average sentence length is observed for Spanish (average of 27.65 tokens per sentence in the test set and an average of 26.54 tokens per sentence in the training set).

	English	French	German	Indonesian	Italian	Korean	Portuguese	Spanish	Swedish
Family	Ger.	Rom.	Ger.	MP.	Rom.	Kor.	Rom.	Rom.	Ger.
Training data									
Sentences	39832	14511	14118	4477	6389	5437	9600	14138	4447
Avg. T/S	23.85	24.2	18.76	21.78	23.34	11.15	24.9	26.54	14.98
Test data									
Sentences	2416	300	1000	557	400	299	1198	300	1219
Avg. T/S	23.46	23.16	16.33	21.15	22.98	8.8	24.57	27.65	16.71

Table 2.5: Description of the universal dependency treebank. Language families include Germanic, Romance, Uralic, Korean, Malayo-Polynesian, Japonic. T/S represents the average count of tokens per sentence.

Universal POS tagset

I use the fine-to-coarse tagset mapping proposed by Naseem et al.[61] and map the language-specific POS tags to universal POS tags. The list of coarse POS tags is included in Table 2.6.

Id	Name	Abbreviation
1	noun	NOUN
2	verb	VERB
3	adverb	ADV
4	adjective	ADJ
5	pronoun	PRON
6	conjunction	CONJ
7	adposition	ADP
8	numeral	NUM
9	determiner	DET
10	punctuation sign	.
11	other	X

Table 2.6: Universal POS tagset.

2.5.2 Experiment setup

I design a set of experiments to evaluate the performance of the developed models and to analyze their specific contributions. In the voting scenarios, the source languages first generate dependencies for the target sentences. Then, the dependencies generated by the majority of the source languages are run through a top-down parsing algorithm to generate a valid dependency parse tree for each target sentence.

Setting 1: One-to-one parser transfer: In this setup the parsing model $m_{L_k}$of each source language S_k is evaluated on the remaining target languages L_{T_k}. The goal of this experiment is to investigate which source language is the best expert on a target language, and what is the performance of languages from the same language family when evaluated on target languages from the same language family.

Setting 2: All source language voting: I allow the parsing models of all source languages to vote towards the parsing model for a given target language T. The goal of this experiment is to identify whether there is a need for weighting the source languages when generating a target parsing model, or whether all languages equally contributing can bring a good parser performance on the target language.

Setting 3: Language family-based parser voting: For a given target language, I weight the contribution of the source languages towards the target parser based on the language family membership of the source and target languages. Specifically, if the source language and the target language are from the same language family, then the source language is allowed to vote. Otherwise, the source language is not included in the voting scheme. The goal of this experiment is to evaluate whether using a voting scheme based on language family membership is sufficient, or whether a more elaborate voting scheme is required.

Setting 4: Language-level expert voting: I compare the results of my models $Predictor_{WALS\_BEST}$ and $Predictor_{WALS\_VOTING}$ against an $Oracle_{language\_level}$ model. The $Oracle_{language\_level}$ system selects the parsing model m_{L_k} stemming from language L_k that performs best when directly applied to the target language. The m_{L_k} parser selection is done once only for a given language. In order to determine the source language that performs best when applied to the target language, the $Oracle_{language\_level}$ requires access to the gold standard of the target language. Thus, the $Oracle_{language\_level}$ differs from all the other parsing systems presented above by assuming the gold standard of the target language is available. Under the current setting, the $Oracle_{language\_level}$ model represents an upper bound for dependency parsing performance on the target language given the information available in the source languages.

Setting 5: Sentence-level expert voting: I compare the results of my $Predictor_{KL\_BEST}$ model against an $Oracle_{sentence\_level}$ model. The $Oracle_{sentence\_level}$ system selects a set of models $m^j_{L_k}, j \leq |T|$, where each $m^j_{L_k}$ performs best on parsing sentence s_j of the target language T. The $m^j_{L_k}$ parser selection is done at the level of sentence s_j. In order to determine the source language that performs best when applied to a sentence of the target language, the $Oracle_{sentence\_level}$ requires access to the gold standard of the target language. Thus, the $Oracle_{sentence\_level}$ differs from all the other parsing systems presented above by assuming the gold standard of the target language is available. Under the current setting, the $Oracle_{sentence\_level}$ model represents an upper bound for dependency parsing performance on the target language given the information available in the source languages.

Setting 6: State of the art comparison I compare the $Predictor_{KL\_BEST}$ model to the state-of-the-art multilingual dependency parsing system of Naseem et al.[59] Using typological features, the model of Naseem et al. learns which aspects of the source languages are relevant to the target language, and ties model parameters accordingly. I choose the *Best Pair* setting of the model proposed by Naseem et al. as a comparison point, in which target model parameters are borrowed from the best source language based on accuracy on the target language. The *Best Pair* model setting gives the best results on the target languages, and is chosen as a reference baseline as it respects similar experimental settings as my $Oracle_{language\_level}$ model.

I also compare the $Predictor_{KL\_BEST}$ model to the target language adaptation model of Täckström et al.,[82] specifically the *Similar* model setup and, to the multi-source transfer model of McDonald et al.[54], specifically the *multi-proj* multi-source projected parser.

The *BestPair*, *Similar*, and *multi-source* systems are evaluated on language subsets of the CoNLL 2006/2007 language-specific corpus.

The performance of the three proposed models and of $Oracle_{language\_level}$ and $Oracle_{sentence\_level}$ is presented on both all-length sentences and the subset of at most 10-length sentences of each corpus.

2.6 Results

The multilingual dependency parsing systems developed in this thesis do not model the $DEPREL$ labels, but only the $(child, head)$ dependency relations. This behavior is due to the lack of universal dependency labels for the CoNLL corpora that could be transferred across languages. Universal dependency labels are available for the UniDep corpus, but to perform a standard comparison across corpora, the $DEPREL$ were not incorporated in the model. Thus, the results presented in this chapter are reported in terms of UAS and not LAS. Following common practices in dependency parse evaluation, I exclude punctuation signs when computing the UAS evaluation scores.

2.6.1 Setting 1: One-to-one parser transfer

Tables 2.7 and 2.8 present the parsing results when the parsing model of a source language is used to predict the dependency trees for every other target language, including itself. The results are reported in terms of UAS over all-length sentences from the language-specific dependency treebank (i.e., Table 2.7) and the universal dependency treebank (i.e., Table 2.8). Each row label represents the selected source language, and the column labels represent the target languages. The table main diagonal contains the results of the parsing model evaluated on the same language it was trained on (i.e., trained on the training data of the same source and tested on the test data of the same target).

In general, for both the language-specific treebank and for the universal treebank, the performance on the target language is optimal when the source is identical to the target language (see main diagonal results). The only exception is Czech, for which the optimal performance is report by Slovene as source language. Looking at the performance of the source language when the source and the target are from the same language family, it can be seen that there is some variability in the UAS results. Furthermore, the source languages from the same language family do not give interchangeable results when evaluated on a target language from the same language family. For example, consider the Romance language family from the language-specific treebank with available languages Catalan, Italian, Portuguese, and Spanish (see first four columns in Table 2.8). The performance on target language Catalan is 73.72 UAS when the source is Italian, 76.92 UAS when the source is Portuguese, and 68.72 UAS when the source is Spanish. For the Germanic language family from the language-specific treebank with available languages German, English, Dutch, Swedish, and Danish, consider the setup where the target language is English (see column En in Table 2.8). The model performance ranges from 45.53 UAS when the source language is German to 57.80 UAS when the source language is Dutch. Similar observations hold for the universal dependency treebank (see Table 2.8) with the Romance (French, Italian, Portuguese, and Spanish) and Germanic (English, German, Swedish) language families.

It is important to notice that when the source language is not identical to the target language, the best performing source language can come from a different language family than the language family of the target language. In addition, the choice of the best performing source language for a target language is not symmetric. Thus, selecting a best source language L_S for target language T does not guarantee that in return T will be the best source language for L_S. For example, the best performing source language on the German target language for the language-specific treebank is Bulgarian with 57.11 UAS, while the best performing source language for the Bulgarian target language is Dutch with 61.28 UAS. Similarly, for the universal dependency treebank, the best performing source language for the English target language is Italian with 72.13 UAS, while the best performing source language for the Swedish target language is also Italian with 68.33 UAS.

These results show that there exists a certain degree of variability in the parser performance when the models are directly transferred from a source to a target language. In addition, there is no consistent best source language across corpora to select for a given target language. An intuitive approach would be to merge the predictions made by each source languages in an optimal way, in order to maximize the benefits brought by each individual language.

S → T	Ca	It	Pt	Es	De	En	Nl	Sv	Da	Bg	Cs	Sl	Ar	Eu	Zh	El	Hu	Ja	Tr
Ca	**85.37**	81.82*	74.23	76.60	52.01	60.29	61.10	63.18	45.98	50.16	58.8	51.78	39.19	37.03	41.02	67.51	35.42	14.54	31.35
It	73.72*	**79.72**	68.92	65.39	54.72	57.96	59.95	59.72	47.8	53.62	53.21	53.06	45.05	36.15	42.41	67.29	36.95	18.35	29.31
Pt	76.92	76.98*	**84.83**	69.59	52.90	67.84	65.50	71.27	53.48	59.05	59.75	39.79	43.22	45.98	52.05	65.86	42.32	16.61	30.78
Es	68.72*	67.02	64.76	**77.74**	45.92	54.48	52.39	55.80	42.20	45.98	51.03	39.95	37.97	36.91	38.25	58.14	33.48	14.89	28.95
De	59.00*	54.76	56.42	47.39	**83.66**	45.53	57.46	56.25	46.02	56.62	46.57	40.61	33.22	33.24	46.47	52.38	41.84	20.10	33.12
En	39.99	43.65	47.75	33.27	36.37	**82.70**	46.01	59.46*	44.59	37.24	37.58	40.13	27.75	45.28	51.51	46.86	43.72	27.15	32.49
Nl	48.88	52.42	52.74	39.46	42.18	57.80*	**71.29**	53.56	44.00	46.98	46.96	44.56	36.83	45.40	47.38	59.22	39.02	22.97	29.65
Sv	56.02	57.76	60.70	48.19	55.64	54.86	55.42	**84.61**	47.52	61.28*	48.63	36.73	41.44	39.46	47.97	57.07	43.75	22.68	28.47
Da	47.68	46.81	53.45*	43.75	44.69	47.86	46.39	47.37	**82.22**	50.14	38.68	38.13	47.49	32.81	37.49	43.55	33.67	15.95	16.47
Bg	60.79	60.95	69.99*	48.41	57.11	57.31	59.84	62.62	60.37	**85.49**	55.37	51.88	45.85	39.73	46.80	55.01	38.08	20.73	22.95
Cs	44.69	41.57	42.91	29.07	41.68	36.10	43.89	37.31	42.53	46.57	**70.45**	72.04*	27.79	31.26	35.00	45.94	39.24	23.66	29.55
Sl	62.28*	59.43	57.46	49.47	48.72	36.02	48.58	46.51	38.12	53.84	49.61	**72.04**	48.40	27.02	41.66	60.88	45.83	29.46	33.34
Ar	49.30*	49.30*	47.27	49.32	37.65	32.68	44.75	26.87	44.72	45.13	46.01	32.43	**73.69**	28.06	16.36	51.22	14.57	5.81	6.17
Eu	18.98	22.52	31.26	19.66	22.94	41.88*	33.50	31.52	34.07	32.61	27.69	31.91	29.75	**62.89**	39.30	24.01	38.51	30.92	27.69
Zh	39.07	39.53	39.88	34.96	40.84	54.89*	43.46	53.49	37.87	38.57	41.50	35.63	29.13	42.20	**83.62**	46.05	58.50	39.78	44.27
El	58.83	60.46	58.28	43.67	46.95	62.11*	60.13	57.16	40.33	45.76	54.12	57.01	27.70	40.28	47.43	**76.83**	45.93	26.91	37.87
Hu	47.72	48.64	48.95	35.94	50.95	49.41	51.00*	59.16	31.63	46.88	44.58	31.31	16.31	51.18	57.37	50.84	**77.24**	37.80	54.38
Ja	30.18	27.24	26.74	25.46	27.82	31.50	31.10	34.56	26.08	29.62	27.64	26.5	21.07	45.55	27.84	33.22	53.25	**83.37**	64.20*
Tr	30.25	30.65	21.08	24.14	31.58	33.15	32.99	31.40	18.42	26.57	25.62	31.47	10.08	38.91	32.49	34.88	41.18	54.58*	**71.94**

Table 2.7: One-to-one language parsing UAS results for all-length sentences for the language-specific dependency treebank. Languages are represented by their ISO language-name abbreviation. The row value represents the selected source language and the column label represents the selected target language. Note: Bolded results represent the same source-same target UAS; starred results represent the target language on which the source language performs best; gray-filled cells represent the best source predictor for each target language, when the source is different from the target language. Double horizontal lines separate languages that belong to the same language family.

S → T	Fr	It	Pt	Es	En	De	Sv	Id	Ko
Fr	**79.84**	76.98	75.94	77.53*	70.60	59.95	67.33	66.99	30.78
It	77.18	**81.43**	77.08	77.69*	72.13	58.82	68.33	65.90	26.23
Pt	75.71	75.46	**80.36**	77.33*	70.15	60.70	66.85	66.88	25.65
Es	74.61	74.66	75.73*	**78.50**	68.01	56.75	67.46	63.69	29.68
En	62.89	62.63	62.68	61.99	**84.76**	51.16	63.61*	44.05	35.81
De	57.76	56.62	57.32	56.52	58.00	**78.26**	61.40*	55.81	35.12
Sv	68.16	68.41	67.76	68.14	69.16*	62.63	**79.23**	58.85	30.64
Id	48.25	52.77*	51.77	50.99	42.07	39.59	41.73	**80.56**	15.92
Ko	33.22	34.38	34.18	36.81	41.10	40.29	41.56*	23.45	**73.39**

Table 2.8: One-to-one language parsing UAS results for all-length sentences for the universal dependency treebank. Languages are represented by the first two letters of the language name. The row value represents the selected source language and the column label represents the selected target language. Note: Bolded results represent the same source-same target UAS; starred results represent the target language on which the source language performs best; gray-filled cells represent the best source predictor for each target language, when the source is different from the target language. Double horizontal lines separate languages that belong to the same language family.

2.6.2 Setting 2: All source language voting

I evaluate the performance of a parsing model created by merging the syntactic knowledge from all source languages. Table 2.11 presents the UAS results for the language specific treebank, while Table 2.12 presents the UAS results for the universal dependency treebank. In general, the performance on the target languages drops when compared to the performance of the same-source same-target setup presented in Tables 2.7 and 2.8. Target languages that have little similarity to the source languages are more positively impacted by this voting, as the performance reported for these languages is much lower (see Japanese with 30.84 UAS in Table 2.11 and Korean with 38.90 UAS in Table 2.7). The all-source language voting scenario manages to outperform the Setting 1 scenario on Portuguese, German, Bulgarian, and Arabic for the language-specific treebank, and on French, Italian, Slovene, and Korean for the language universal treebank. These results show that in order to obtain a good overall parsing performance on the target languages, the source languages should contribute in a more informed manner to the parsing process.

Ca	It	Pt	Es	De	En	Nl	Sv	Da	Bg	Cs	Sl	Ar	Eu	Zh	El	Hu	Ja	Tr
71.68	68.9	78.38	62.93	60.72	51.56	58.64	63.62	54.13	71.24	50.49	46.10	53.05	36.64	47.61	61.76	58.14	30.84	30.52

Table 2.9: All-source language voting UAS results on all-length sentences for the language specific treebank. Gray-filled cells represent target languages with performance results better than the best source-predictor in Setting 1.

Fr	It	Pt	Es	En	De	Sv	Id	Ko
77.66	78.34	76.97	75.69	63.74	60.27	71.73	50.66	38.90

Table 2.10: All-source language voting UAS results on all-length sentences for the language universal treebank. Gray-filled cells represent target languages with performance results better than the best source-predictor in Setting 1.

2.6.3 Setting 3: Language family-based parser voting

I evaluate a simple voting scheme based on source and target language membership to a language family. The main idea of this experiment is to validate the need for a more complex voting methodology. Results are presented in Table 2.11 and Table 2.12, where no results are included for languages that do not have another language member from the same language family present in the corpus. The reported results are larger than the results presented in Setting 2 only for three out of the 19 languages in the language-specific treebank and for three out of nine languages in the language-universal treebank. In addition, I cannot compute results for languages that do not have another language member from the same language family present in the treebank. Choosing such a strict voting scheme would reduce the applicability of the parsing model to languages for which one knows a priori the language family they belong to, and for which one also has linguistic resources available for the associated language family.

Ca	It	Pt	Es	De	En	Nl	Sv	Da	Bg	Cs	Sl	Ar	Eu	Zh	El	Hu	Ja	Tr
81.55	73.44	77.12	68.34	55.56	48.83	52.01	60.01	50.90	56.28	46.9	40.09	-	-	-	-	-	-	-

Table 2.11: Language family-based parser voting UAS results on all-length sentences for the language specific treebank. Gray-filled cells represent target languages with performance results better than the results reported in Setting 2.

The results discussed so far show that:

Fr	It	Pt	Es	En	De	Sv	Id	Ko
78.71	77.40	76.52	54.59	60.48	65.37	64.08	-	-

Table 2.12: Language family-based parser voting UAS results on all-length sentences for language universal treebank. Gray-filled cells represent target languages with performance results better than the results reported in Setting 2.

- source languages need to be weighed in order to contribute relevant syntactic information to the target language

- the weighting scheme has to be more complex and customizable to the composition of the set of source languages and the input target language

2.6.4 Setting 4: Language-level expert voting results

The language-level expert voting results are included in Table 2.13 and Table 2.14 for all-length sentences and at most 10-length sentence of the language specific treebank, and in Table 2.15 for the universal dependency treebank. The tables contain the experiment results for $Oracle_{language\_level}$, $Predictor_{WALS\_BEST}$, and $Predictor_{WALS\_VOTING}$.

Column three of Table 2.13 shows that the $Oracle_{language\_level}$ is not consistently selected from the same language family with the target language. Specifically, for the romance languages, $Oracle_{language\_level}$ is always from the same language family, but for the Germanic and Slavic languages, $Oracle_{language\_level}$ is mainly represented by source languages outside the target language family (e.g., best predictor language for German is Catalan, for Slovene is Greek). In contrast, $Predictor_{WALS\_BEST}$ is from the same language family as the target language for the Romance and Germanic languages. $Predictor_{WALS\_BEST}$ overlaps with $Oracle_{language\_level}$ in terms of the best source language selection only for five of the target languages (Catalan, Chinese, Hungarian, Turkish, and Japanese). The average performance of the $Predictor_{WALS\_BEST}$ model on all target languages is approximatively 8% lower than the average performance of the $Oracle_{language\_level}$ model on all target languages.

When I combine the top ω source languages in the $Predictor_{WALS\_VOTING}$ model, I obtain better results than when using the $Predictor_{WALS\_BEST}$ model. In order to find

	$Oracle_{language\_level}$		$Predictor_{WALS\_BEST}$		$Predictor_{WALS\_VOTING}$	
Target Language	UAS	Source Language	UAS	Source Language	$\omega = 3$	$\omega = 6$(best)
Catalan	81.82	Italian	81.82	Italian*	81.04	77.56
Italian	73.72	Catalan	65.39	Spanish	68.14	72.16
Portuguese	78.00	Catalan	76.98	Italian	75.98	**78.14**
Spanish	70.23	Catalan	67.02	Italian	62.19	67.26
German	59.72	Catalan	57.46	Dutch	55.68	57.36
English	57.21	Swedish	37.24	Bulgarian	40.7	49.68
Dutch	57.70	Greek	42.18	German	48.36	57.46
Swedish	61.82	Portuguese	47.52	Danish	59.15	**64.36**
Danish	52.55	Basque	47.37	Swedish	49.92	**52.91**
Bulgarian	66.95	Portuguese	57.31	English	57.27	63.83
Czech	50.82	Slovene	36.10	English	45.52	48.01
Slovene	55.94	Greek	40.13	English	44.82	48.4
Arabic	52.71	Italian	51.22	Greek	51.77	**53.29**
Basque	39.94	English	30.92	Japanese	34.47	**40.63**
Chinese	59.32	Hungarian	58.50	Hungarian*	57.02	50.14
Greek	60.99	Italian	45.76	Bulgarian	54.63	60.04
Hungarian	58.24	Chinese	58.37	Chinese*	56.65	56.8
Japanese	64.2	Turkish	64.20	Turkish*	64.1	47.79
Turkish	54.58	Japanese	54.58	Japanese*	42.51	41.12
Average	60.86	-	53.64	-	54.54	57.20

Table 2.13: Language-level expert voting UAS results reported for all-length sentences of the language specific dependency treebank. Row labels represent the target language; the first two columns represent the UAS and the best predictor source language as generated by the $Oracle_{language\_level}$ model; the $Oracle_{language\_level}$ model represents an upper bound for dependency parsing performance on the target language, given the information available in the source languages. Columns 3 and 4 represent the UAS and best predictor source language as generated by the $Predictor_{WALS\_BEST}$ model. The last two columns represent the UAS results for $\omega = 3$ source predictors and $\omega = 6$ source predictors in the $Predictor_{WALS\_VOTING}$ model. Note: Starred language names are the best predictor source languages selected by the $Predictor_{WALS\_BEST}$ model that overlap with the best predictor source language selected by the $Oracle_{language\_level}$ model. Double horizontal lines separate languages that belong to the same language family.

the optimal ω that gives the highest average performance across all languages I run the $Predictor_{WALS_VOTING}$model with ω taking values from $1 \rightarrow 19$. The optimal ω is 6 with an average performance of 58.77 UAS across all languages, only 3% lower than the performance obtained by $Oracle_{language\_level}$ across all languages. The optimal ω involves a high number of source languages, which implies that the syntactic diversity cannot be captured by a small number of source languages alone. Similarly, adding too many source languages adds more noise to the model, so the language ranking has to score the most optimal source languages to consider for a target language.

	$Oracle_{language\_level}$		$Predictor_{WALS\_BEST}$		$Predictor_{WALS\_VOTING}$	
Target Language	UAS	Source Language	UAS	Source Language	$\omega = 3$	$\omega = 6$(best)
Catalan	100	French	100	French*	100	100
Italian	81.15	French	71.84	Spanish	76.85	79.24
Portuguese	81.91	Italian	79.46	Spanish	81.66	83.13
Spanish	77.73	Catalan	75.36	Italian	65.4	75.83
German	75.21	Dutch	75.21	Dutch*	70.99	69.31
English	77.61	Swedish	56.72	Bulgarian	60.45	64.93
Dutch	64.26	English	50.47	German	53.29	59.56
Swedish	78.25	Bulgarian	61.52	danish	74.68	79.69
Danish	61.15	Portuguese	55.41	Swedish	55.63	57.62
Bulgarian	78.89	Portuguese	72.19	English	68.65	71.43
Czech	57.78	Bulgarian	48.69	English	58.55	57.16
Slovene	66.62	French	44.99	English	49.64	54.98
Arabic	68.52	Italian	57.41	Greek	62.04	60.19
Basque	51.23	English	38.13	Korean	40.26	49.26
Chinese	62.86	Hungarian	62.86	Hungarian*	59.49	56.76
Greek	70.69	Portuguese	61.49	Bulgarian	68.39	71.84
Hungarian	72.8	Chinese	72.8	Chinese*	71.6	72.8
Japanese	76.91	Korean/Turkish	76.91	Turkish*	79.51	74.07
Turkish	67.21	Korean	55.87	Japanese	62.57	62.3
Average	71.84	-	64.05	-	66.26	68.36

Table 2.14: Language-level expert voting UAS results reported for at most 10-length sentences of the language-specific dependency treebank. Row labels represent the target language. The first two columns represent the UAS and the best predictor source language as generated by the $Oracle_{language\_level}$ model; the $Oracle_{language\_level}$ model represents an upper bound for dependency parsing performance on the target language, given the information available in the source languages. Columns 3 and 4 represent the UAS and best predictor source language as generated by the $Predictor_{WALS\_BEST}$ model. The last two columns represent the UAS results for $\omega = 3$ source predictors and $\omega = 6$ source predictors in the $Predictor_{WALS\_VOTING}$ model. Note: Starred language names are the best predictor source languages selected by the $Predictor_{WALS\_BEST}$ model that overlap with the best predictor source language selected by the $Oracle_{language\_level}$ model. Double horizontal lines separate languages that belong to the same language family.

When I evaluate my system only on sentences of length 10 or less, I observe higher UAS performance (see Table 2.14 and Table 2.15). The $Oracle_{language\_level}$ model as well as $Predictor_{WALS\_BEST}$ model experience an approximative 10% increase in overall UAS performance compared to the results over all-length sentences.

All-length sentences						
	$Oracle_{language\_level}$		$Predictor_{WALS\_BEST}$		$Predictor_{WALS\_VOTING}$	
Target Language	UAS	Source Language	UAS	Source Language	$\omega = 3$	$\omega = 6$(best)
French	77.53	Spanish	77.53	Spanish*	76.83	77.81
Italian	77.69	Spanish	77.69	Italian*	76	78.6
Portuguese	77.33	Spanish	66.88	Portuguese	76.21	77.15
Spanish	75.73	Portuguese	74.66	Spanish	72.65	76.16
English	63.61	Swedish	61.99	English	59.14	63.37
German	61.4	Swedish	61.4	Swedish*	60.94	60.08
Swedish	69.16	English	69.16	English*	70.65	72.13
Indonesian	52.77	Italian	52.77	Indonesian	52.27	51.14
Korean	41.56	Swedish	40.29	German	40.13	41.48
Average	66.31	-	64.71		64.98	66.44

At most 10-length sentences						
	$Oracle_{language\_level}$		$Predictor_{WALS\_BEST}$		$Predictor_{WALS\_VOTING}$	
Target Language	UAS	Source Language	UAS	Source Language	$\omega = 3$	$\omega = 6$(best)
French	84.23	Portuguese	83.22	Spanish	83.22	85.87
Italian	82.93	French	79.47	Spanish	81.1	82.52
Portuguese	83.28	Spanish	74.69	Indonesian	84.17	84.84
Spanish	81.51	Italian	81.51	Italian*	76.23	78.87
English	76.94	Swedish	75.74	Spanish	75.15	77.01
German	75.4	Swedish	75.4	Swedish*	74.04	73.59
Swedish	81.16	German	80.67	English	81.21	83.3
Indonesian	62.01	Spanish	60.78	Italian	60.78	59.36
Korean	45.12	German	45.12	German*	44.13	44.25
Average	74.73		72.96		73.34	74.40

Table 2.15: Language-level expert voting UAS results reported for all- and at most 10-length sentences of the universal dependency treebank. Row labels represent the target language. The first two columns represent the UAS and the best predictor source language as generated by the $Oracle_{language\_level}$ model; the $Oracle_{language\_level}$ model represents an upper bound for dependency parsing performance on the target language, given the information available in the source languages. Columns 3 and 4 represent the UAS and best predictor source language as generated by the $Predictor_{WALS\_BEST}$ model. The last two columns represent the UAS results for $\omega = 3$ source predictors and $\omega = 6$ source predictors in the $Predictor_{WALS\_VOTING}$ model. Note: Starred language names are the best predictor source languages selected by the $Predictor_{WALS\_BEST}$ model that overlap with the best predictor source language selected by the $Oracle_{language\_level}$ model. Double horizontal lines separate languages that belong to the same language family.

The same number of the $Predictor_{WALS\_BEST}$ languages identified by my system overlap with the $Oracle_{language\_level}$ languages (5 $Predictor_{WALS\_BEST}$ overlap with the $Oracle_{language\_level}$). The same performance difference is observed between the $Predictor_{WALS\_BEST}$

and the $\mathit{Predictor}_{\mathit{WALS\_VOTING}}$ when compared to the performance difference of the two systems on the language-specific treebank. For the Germanic language family, only one $\mathit{Predictor}_{\mathit{WALS\_BEST}}$ model is not selected from the same language family with the target language family (i.e.: Bulgarian as predictor for English), meanwhile for the Slavic language family all target languages have the same $\mathit{Predictor}_{\mathit{WALS\_BEST}}$ model, specifically the English model.

I evaluate my system on the universal dependency treebank, and observe an improved performance compared to the performance on the language specific CoNLL corpus. On the target languages for which the source language set contains at least one other language from the same language family, the $\mathit{Predictor}_{\mathit{WALS\_BEST}}$ model is always selected from the same language family. The choice of the $\mathit{Predictor}_{\mathit{WALS\_BEST}}$ is important even when it comes from the same language family, as different source languages from the same language family will report different results on the target language. My system performs better on the universal dependency treebank compared to its performance on the language specific dependency treebank. It more often selected the $\mathit{Predictor}_{\mathit{WALS\_BEST}}$ language that overlapped the $\mathit{Oracle}_{\mathit{language\_level}}$. In general, the $\mathit{Predictor}_{\mathit{WALS\_VOTING}}$ model performs as well as or better than the reference $\mathit{Oracle}_{\mathit{language\_level}}$ model on all languages except for German, Indonesian, and Korean.

I further evaluate the impact of ω on the system performance (see Figure 2.3). I notice a difference in performance based on the language family, but in general all languages are best predicted by a number of $3 \rightarrow 6$ source languages. The average performance on all of the languages drops systematically once the number of voting languages ω is greater than 6.

2.6.5 Setting 5: Sentence-level expert voting results

Tables 2.16 and 2.17 present the results for sentence-level parsing. The $\mathit{Oracle}_{\mathit{sentence\_level}}$ model outperforms both the $\mathit{Oracle}_{\mathit{language\_level}}$ and my $\mathit{Predictor}_{\mathit{WALS\_BEST}}$ and $\mathit{Predictor}_{\mathit{WALS\_VOTING}}$ models. For some of the languages, the $\mathit{Oracle}_{\mathit{sentence\_level}}$ performs better than or as well as the m_k model, the MST Parser trained and evaluated on the same language.

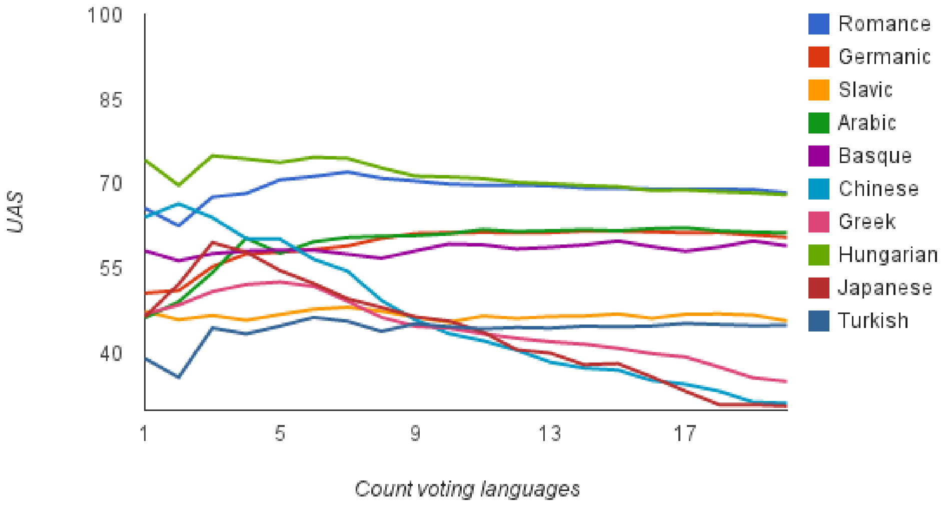

Figure 2.3: Multilingual dependency parsing performance with increase in count of voting languages ω across language families

For the language specific treebank see Italian with 80.11 UAS by $Oracle_{sentence\_level}$ vs. 79.72 UAS by MST Parser. For the universal treebank see French with 83.84 UAS by $Oracle_{sentence\_level}$ vs. 81.43 UAS by the MST Parser, Korean with 74.27 by UAS $Oracle_{sentence\_level}$ vs. 73.39 UAS by the MST Parser. The $Predictor_{KL\_BEST}$ model outperforms the $Predictor_{WALS\_VOTING}$ model, for both the language specific and the universal treebank for all languages of the corpus except for Basque, Chinese, Japanese, and Turkish.

Table 2.18 shows the distribution of languages that contribute to parsing the target language in the $Oracle_{sentence\_level}$ model. The columns represent the top five contributing language models, based on the percentage of sentences they predict better than any other source language model. The percentage of best predicted sentences is included in brackets following the language name. For the language-specific treebank, the most common top language contributor is Catalan (top language contributor for 13 out of the 19 languages). The second best language contributor is more varied across the set of target languages. In some of the cases, the top language contributor is selected from language families completely unrelated to the target language: see Chinese as the second best language contributor for English, Hungarian as the third

Target Language	$Oracle_{sentence\_level}$	$Predictor_{KL\_BEST}$
Catalan	85.28	83.59
Italian	80.11	74.73
Portuguese	84.35	81.22
Spanish	72.75	69
German	69.89	61.24
English	66.36	53.63
Dutch	69.93	58.46
Swedish	74.4	66.67
Danish	63.71	55.58
Bulgarian	79.59	69.35
Czech	62.62	54.63
Slovene	70.22	56.5
Arabic	58.01	57.25
Basque	54.2	41.51
Chinese	75.82	63.15
Greek	73.11	66.53
Hungarian	69.56	63.37
Japanese	67.6	51.54
Turkish	62.1	50.04
Average	70.50	62.07

Table 2.16: Sentence-level expert voting UAS results reported for all-length sentences of the language specific dependency treebank. Row labels represent the target language. The first column represents the UAS results generated by the $Oracle_{sentence\_level}$ model; the $Oracle_{sentence\_level}$ model represents an upper bound for dependency parsing performance on the target language, given the information available in the source languages The second column represents the UAS results generated by the $Predictor_{KL\_BEST}$ model. Double horizontal lines separate languages that belong to the same language family.

best language contributor for Japanese. Some of the top language contributors are languages for which most of the existing parsing models (including the ones presented in this thesis) have difficulties generating a high-performance parser (see Basque as the third best language contributor for Turkish). The top language contributors on the universal treebank are more consistent across the language families, although for the Germanic languages, most often Romance languages (French, Italian) rank high.

Target Language	$Oracle_{sentence\_level}$	$Predictor_{KL\_BEST}$
French	83.84	80.03
Italian	83.92	79.88
Portuguese	83.08	79.58
Spanish	81.03	77.15
English	72.74	63.89
German	71.29	63.77
Swedish	79.15	74.41
Indonesian	60.21	54.59
Korean	52.53	45.49
Average	74.27	68.75

Table 2.17: Sentence-level expert voting UAS results reported for all-length sentences of the universal dependency treebank. Row labels represent the target language. The first column represents the UAS results generated by the $Oracle_{sentence\_level}$ model; the $Oracle_{sentence\_level}$ model represents an upper bound for dependency parsing performance on the target language, given the information available in the source languages. The second column represents the UAS results generated by the $Predictor_{KL\_BEST}$ model. Double horizontal lines separate languages that belong to the same language family.

2.6.6 Setting 6: State of the art comparison

Table 2.20 presents the comparison between the $Oracle_{sentence\_level}$ and $Predictor_{KL\_BEST}$ models and the three state-of-the-art models - *Best Pair*, *Similar*, and *multi-source*. As an optimal model, the $Oracle_{sentence\_level}$ model outperforms the Best Pair baseline model across all languages. The $Predictor_{KL\_BEST}$ model manages to outperform the Best Pair model on 12 out of 17 languages. It lacks in performance on Basque, Chinese, Japanese, Arabic, and Turkish, languages that are not syntactically similar to many of the source languages.

The *Similar* model presents performance results better than the *Best Pair* model. Yet, my $Oracle_{sentence\_level}$ model outperforms the *Similar* model across the 16 target languages for which *Similar* has reported performance results. The $Predictor_{KL\_BEST}$ model manages to perform better than the *Similar* model only on 12 of the 16 target languages. It is outperformed on Basque, Hungarian, Japanese, and Turkish. The $Predictor_{KL\_BEST}$ and the *Similar* models obtain the same performance on Spanish. The *multi-source* model performs better than the $Predictor_{KL\_BEST}$ model on Dutch

Target	Percentage of target sentences best predicted by source language				
Catalan	Italian (58.68)	Spanish (16.76)	Portuguese (16.76)	English (1.79)	Slovene (1.19)
Italian	Catalan (52.4)	Portuguese (15.2)	Spanish (8.4)	Greek (7.2)	Slovene (3.2)
Portuguese	Catalan (44.36)	Italian (23.18)	Swedish (8.65)	English (6.92)	Spanish (5.53)
Spanish	Catalan (54.10)	Italian (18.35)	Portuguese (11.11)	English (2.89)	Greek (2.89)
German	Catalan (39.38)	Dutch (15.36)	Portuguese (9.21)	Italian (8.10)	Bulgarian (6.14)
English	Swedish (34.41)	Chinese (13.02)	Portuguese (9.76)	Dutch (8.83)	Greek (6.97)
Dutch	Catalan (25.06)	English (19.89)	Greek (11.88)	Italian (9.30)	Swedish (7.49)
Swedish	Catalan (26.15)	Italian (12.56)	Portuguese (10.76)	Bulgarian (10.25)	Dutch (8.71)
Danish	Catalan (21.36)	Portuguese (14.55)	English (10.52)	Italian (10.21)	Bulgarian (8.66)
Bulgarian	Catalan (27.81)	Portuguese (21.80)	Italian (9.02)	Dutch (6.26)	Danish (5.76)
Czech	Catalan (19.67)	Slovene (18.57)	Dutch (10.10)	Danish (10.10)	Italian (9.83)
Slovene	Catalan (25.55)	Italian (17.61)	Czech (13.89)	Greek (13.15)	Dutch (6.20)
Arabic	Catalan (16.03)	Italian (15.26)	Greek (12.97)	Dutch (10.68)	Spanish (9.99)
Basque	English (18.80)	Dutch (11.04)	Hungarian (10.75)	Portuguese (9.25)	Catalan (8.65)
Chinese	Catalan (22.81)	English (15.09)	Hungarian (11.06)	Dutch (8.75)	Italian (8.52)
Greek	English (21.71)	Catalan (19.69)	Dutch (14.64)	Slovene (12.12)	Italian (11.61)
Hungarian	Swedish (17.90)	Catalan (13.81)	Chinese (10.99)	Turkish (8.95)	Dutch (7.67)
Japanese	Catalan (51.54)	Turkish (23.94)	Hungarian (8.88)	Basque (4.92)	Italian (2.11)
Turkish	Japanese (34.13)	Catalan (16.66)	Basque (8.97)	English (6.57)	Hungarian (5.92)

Table 2.18: Percentage of target sentences best predicted by source languages, ordered by highest source contribution over the language-specific treebank. Contribution of source languages to parsing the target language is computed from the $Oracle_{sentence\_level}$ model. The numbers in brackets represent the percentage of sentences the language model predicts better than any other source language model. Double horizontal lines separate languages that belong to the same language family.

Target	Percentage of target sentences best predicted by source language				
French	Italian (42.53)	Portuguese (19.93)	Spanish (15.94)	Indonesian (8.63)	English (6.64)
Italian	French (43.64)	Portuguese (22.19)	Spanish (15.71)	English (11.47)	Indonesian (3.99)
Portuguese	French (39.53)	Italian (20.85)	Spanish (20.35)	English 6.92	Indonesian (5.08)
Spanish	French (40.53)	Italian (24.25)	Portuguese (21.59)	English (5.31)	Indonesian (4.98)
English	French (31.98)	Italian (19.81)	Swedish (18.49)	Portuguese (15.64)	Spanish (0.88)
German	French (29.47)	Swedish (16.58)	Italian (13.18)	English (12.38)	Indonesian (10.09)
Swedish	French (35.32)	Italian (17.62)	English (12.62)	Portuguese (12.21)	German (10.57)
Indonesian	Italian (36.73)	French (23.11)	Portuguese (18.10)	Spanish (11.29)	German (4.66)
Korean	German (20)	English (18)	French (17.66)	Swedish (13.33)	Italian (12.33)

Table 2.19: Percentage of target sentences best predicted by source languages, ordered by highest source contribution over the language-universal treebank. Contribution of source languages to parsing the target language is computed from the $Oracle_{sentence\_level}$ model. The numbers in brackets represent the percentage of sentences the language model predicts better than any other source language model. Double horizontal lines separate languages that belong to the same language family.

and Slovene, but it is outperformed on the remaining six languages for which it has reported performance results.

Target Language	$Oracle_{sentence\_level}$	$Predictor_{KL\_BEST}$	State-of-the-art models		
			Best Pair	*Similar*	*multi-source*
Catalan	85.28	83.59*	74.8	**80.2**	-
Italian	80.11	74.73*	68.3	**74.6**	65.0
Portuguese	84.35	81.22*	76.4	**78.4**	75.6
Spanish	72.75	69	63.4	**69**	64.5
German	69.89	61.24*	54.8	**58.1**	56.6
English	66.36	53.63*	**44.4**	-	-
Dutch	69.93	58.46	57.8	51.8	**65.7**
Swedish	74.4	66.67*	**63.5**	48.8	-
Danish	63.71	55.58*	-	-	**49.5**
Bulgarian	79.59	69.35*	**66.1**	62.4	-
Czech	62.62	54.63*	**47.5**	45.3	-
Slovene	70.22	56.5	-	-	**68**
Arabic	58.01	57.25	**57.6**	52.7	-
Basque	54.2	41.51	42.0	**46.8**	-
Chinese	75.82	63.15	**65.4**	54.8	-
Greek	73.11	66.53*	60.6	59.9	**65.1**
Hungarian	69.56	63.37	57.0	**64.5**	-
Japanese	67.6	51.54	54.8	**64.6**	-
Turkish	62.1	50.04	56.9	**59.5**	-
Average	70.50	62.07	-	-	-

Table 2.20: Sentence-level expert voting UAS results reported for all-length sentences of the language specific dependency treebank. Row labels represent the target language. The first column represents the UAS results generated by the $Oracle_{sentence\_level}$ model; the $Oracle_{sentence\_level}$ model represents an upper bound for dependency parsing performance on the target language, given the information available in the source languages. The second column represents the UAS results generated by the $Predictor_{KL\_BEST}$ model. The last three columns represent the UAS results of the *Best Pair* model, the *Similar* model, and the *multi-source* model, respectively. Double horizontal lines separate languages that belong to the same language family. Starred results are languages for which the $Predictor_{KL\_BEST}$ model performs better than the state-of-the-art models. Bolded results represent the best results per target language obtained by the state-of-the-art models.

In general, the $Oracle_{sentence\_level}$ model represents a upper-bound on the performance of a parsing model built from source languages at a sentence-level. Thus, it manages to outperform the three state-of-the-art models. On the other side, the $Predictor_{KL\_BEST}$ performs better than state-of-the-art models mainly on target languages for which a larger set of similar source languages are available. For example, when compared to the *Similar* model, the $Predictor_{KL\_BEST}$ model is outperformed on Basque, Hungarian, Japanese, and Turkish, languages that are the only representatives from their respective language families. The *multi-source* model manages to outperform the $Predictor_{KL\_BEST}$ model on Dutch and Slovene, even though for those languages there exists a larger set of source languages from the same language family. The improvements brought by the *multi-source* model can be explained by the constraint driven algorithm that borrows syntactic knowledge from parallel corpora.

My $Predictor_{KL\_BEST}$ model has the advantage of precisely selecting which source languages should parse each target sentence, instead of selecting a source or a set of source languages to perform parsing over the entire set of target sentences, or generating a target parser using selective sharing of model parameters from source languages. This advantage is more evident for the romance languages, where it achieves better performance results compared to the *Best Pair*, *Similar* and *multi-source* models. The largest performance improvement on the Romance languages is on Portuguese, where my model obtains 81.22 UAS compared to 78.4 UAS the best performance of the state-of-the-art systems. Based on the best source language selection made by the $Oracle_{sentence\_level}$ model, a relatively large percentage of target sentences are predicted by source languages that are not typologically close to the target language. On Portuguese in particular, 8.65% of the target sentences are best predicted by Swedish and 6.92% by English. One possible explanation why my model could achieve better performance is because it ranks source languages based on the KL divergence on the distributions of POS transitions for a specific sentence, instead of only ranking languages that are typologically similar.

2.7 Discussion

In general, the systems perform better when evaluated over shorter sentences, regardless of the implementation methodology. In addition, using the voting scheme performs better than automatically selecting the $Predictor_{WALS\_BEST}$ language. Also, the $Predictor_{WALS\_VOTING}$ model tends to favor languages from the same language family with the target language, in contrast to the $Oracle_{language\_level}$ model which can be selected from totally unrelated language families (see Greek as a source predictor for Dutch and Slovene).

The universal dependency treebank allows some interesting conclusions to surface. First, I observe that the dependency parsing results are in general better than the ones obtained on the language specific CoNLL treebank. Secondly, Germanic languages are predicted at a higher accuracy when using the universal dependency annotations. I conclude that the universality together with the consistency of the annotations allows for parsing models to correctly select and transfer the language phenomena that are consistent across languages. These universals are also correctly evaluated as they have the same schema across all languages. When using the universal treebank I notice that the $Oracle_{sentence\_level}$ language is from the same language family with the target language. This follows the linguistic intuition and also matches the automated predictions made by my system $Predictor_{WALS\_BEST}$. My system does not manage to greatly outperform the $Oracle_{language\_level}$ model predictions when using the universal treebank, but instead it manages to match the performance of the best predictor languages $Oracle_{language\_level}$ by learning linguistic phenomena from the available data. Thus, my $Predictor_{WALS\_BEST}$ model is able to identify which language can best parse a target sentence via available linguistic knowledge.

2.8 Conclusions

I conclude that sentence-level knowledge transfer is more appropriate in the multilingual setting when compared to the language level. At this level one can more finely identify syntactic rules and select the language from which to import the appropriate rules. I show that, even though source languages are available from the same language family, the best parser performance on a target language is not always given by a source language from the same language family. I attribute this to both a diversity in treebank annotations across languages and to the degree of diversity inherent in the natural language generation process.

Chapter 3

Corpus Creation

3.1 Chapter overview

I present the process of creating a multilingual corpus annotated for named entities and coreference resolution. I give an overview of existing corpora spanning across multiple languages, and I present the novelty introduced by my corpus. I discuss the annotation process and the inter-annotator agreement on the tasks of named-entity recognition and coreference annotation.

3.2 Introduction

The goal of NLP systems is to emulate a human-like understanding of natural language. In order to evaluate how accurate the designed system is, one needs to compare it against the expert in the domain, in this case the human. Such evaluations are carried out against decisions made by humans on specific documents, where the decisions are dictated by the NLP task of interest. The process of making decisions on documents is defined as annotating specific portions of the document (i.e., tokens, sentences, or even paragraphs) with a finite set of given tags. Such tags are $\{verb, noun, adjective, ...\}$ for the task of part of speech identification, or $\{$*beginning mention*, *inside mention*, *not a mention*$\}$ for the task of mention identification.

In general, natural language is an ambiguous medium of communication. Even human experts exhibit disagreement over how ambiguous language should be interpreted. In order to conclude the gold standard, the annotations made by k different annotators are presented to a human expert arbitrator who has to reconcile disagreements.

3.3 Related work

Multilingual annotation efforts for natural language were carried out mainly on the newswire genre, where different NLP tasks were investigated.[70, 62, 77] In the multilingual newswire domain, the SemEval and the CoNLL Shared Tasks made multilingual corpora for several NLP tasks available to the research community. For example, the SemEval shared task prepared multilingual corpora for semantic textual similarity in English and Spanish[77], for multilingual word sense disambiguation in English, French, German, and Spanish[62], and for coreference resolution in Catalan, Dutch, English, German, Italian, and Spanish [73]. Similarly, the 2012 CoNLL corpus[70] generated several layers of annotations including named entities and coreference resolution in Arabic, Chinese, and English. The corpora prepared by both SemEval and CoNLL contain different documents for each of the languages, and there is no semantic equivalence between the texts of the documents.

The main concern with multilingual annotations is that the texts available for each language could have different writing styles or belong to different genres. Consequently, the task of annotation might be more ambiguous and implicitly more difficult due to the different text that has to be annotated in each language. In order to overcome the issue of unequal comparison points between multilingual corpora, some authors have proposed working with parallel corpora, i.e., corpora where the same text is available in the different languages of interest. In most settings, parallel corpora are composed of bilingual corpora. Exceptions are the multilingual corpora prepared through the OPUS initiative,[84] that include corpora spanning different genres. The EuroParl corpus is an OPUS member and contains a collection of proceedings of the

European Parliament in 11 European languages. To my knowledge, this corpus has not been previously annotated for named entities or coreference resolution.

3.4 Corpus description

The EuroParl corpus is part of the OPUS initiative of the European Union and contains approximatively 40 million words per each of 11 European languages.[84] I select a subset of European languages (i.e., English, French, and German) and annotate them for named entities and coreference resolution. The named entities belong to standard named-entity categories frequently used in the literature: person, organization, and location. The annotation guidelines used for this task are included in Appendix A. In the rest of this thesis I refer to the annotated parallel corpus as EuroParl$_{parallel}$.

The EuroParl$_{parallel}$ corpus contains the written proceedings from two meetings of the European Parliament in the form of two distinct documents. After annotation, I split the EuroParl$_{parallel}$ corpus into a training and test sub-corpus by allocating a proceedings document to the training corpus and one to the test corpus. No split was made over the paragraphs or sentences of the large corpus, as the two proceedings documents are stand-alone documents.

I select one native speaker of English, German, and French, respectively, with previous experience in annotating documents for named entities and coreference resolution on English documents. Each annotator is trained on using the annotation software (i.e., NotableApp)[38] and the given annotation guidelines using an online training process. The annotators are then required to annotate documents in their native language. Annotation reconciliation is performed by a fourth annotator (i.e., arbitrator) fluent in the three languages.

The annotators first identify the mentions of the three semantic categories within the text. If a mention is semantically identical to a previous mention then the two are linked. The linked mentions create a chain, usually of length 2 or more. The mentions that are not linked to any previous mention in the document are referred to as singletons.

Several statistics on the training and test corpus are presented in Table 3.1. Each language has 171 total paragraphs in the training corpus, and 72 in the test corpus. Even though all languages have the same number of paragraphs, there is a slight variation in the number of sentences: there are 397 sentences for English and French in the training corpus compared to 413 sentences for German in the training corpus, and 145 sentences for English, 147 sentences for French, and 146 sentences for German in the test corpus. For both the training and the test corpus, the average number of words per sentence varies across the set of languages. German has the smallest average number of words per sentence (i.e., 20.43 average sentence length on the training corpus, and 21.85 average sentence length on the test corpus). French has the largest average number of words per sentence (i.e., 26.87 average sentence length on the training corpus, and 26.89 average sentence length on the test corpus). The total number of words per corpus is highest for French (i.e., 10670 words in the training corpus and 3954 in the test corpus) and smallest for German (i.e., 8439 words in the training corpus an 3201 in the test corpus).

Language	# Paragraphs	# Sentences	Average Sentence length	# Words
Training				
English	171	397	22.77	9042
French	171	397	26.87	10670
German	171	413	20.43	8439
Test				
English	72	145	23.04	3342
French	72	147	26.89	3954
German	72	146	21.85	3201

Table 3.1: EuroParl$_{parallel}$ description: sentence and paragraph count, and average sentence length, i.e., average number of words per sentence.

Tables 3.2 and 3.3 present the number of mentions and coreference chains for each of the languages in the training and test corpus, respectively. The total number of mentions is 703 for English, 719 for French, and 701 for German in the training corpus, and 293 for English, 294 for French, and 289 for German in the test corpus.

French has the largest number of mentions as well as the largest number of chains (97 chains in the training corpus and 48 chains in the test corpus). Even though French has the largest number of mentions, the largest average chain size comes from German, with an average of 6.36 mentions per chain in the training corpus and 5.63 average chain length in the test corpus. In the training corpus, the largest number of mentions and singletons come from the person category, while in the test corpus the person category has the highest number of mentions but the location category is the most numerous in number of singletons.

Language	# Mentions	# Chains	Average Chain Size	# Singletons
Person				
English	339	60	4.56	65
French	350	62	4.53	69
German	352	62	4.62	65
Location				
English	42	10	3.1	11
French	51	11	3.27	15
German	45	8	3.87	14
Organization				
English	322	24	12.16	30
French	318	24	11.91	32
German	304	22	12.18	36
Overall				
English	703	94	6.35	106
French	719	97	6.21	116
German	701	92	6.36	115

Table 3.2: EuroParl$_{parallel}$ training corpus description: the number of mentions, chains, the average chain length, and the number of singletons. Singletons are excluded when computing the statistics over the chains.

In the following section I discuss the inter-annotator agreement process and analyze the complexity of performing annotations across multiple languages. The work presented in the following chapter was carried out together with Cosmin Gheorghe, as part of his MIT Undergraduate Advanced Project requirement.

Language	# Mentions	# Chains	Average Chain Size	# Singletons
Person				
English	109	23	4.56	4
French	113	25	4.36	4
German	109	23	4.52	5
Location				
English	93	8	5.66	21
French	99	10	7.9	20
German	98	9	8.55	21
Organization				
English	91	14	6.85	17
French	82	13	5.13	15
German	82	12	5.58	15
Overall				
English	293	45	5.57	42
French	294	48	5.31	39
German	289	44	5.63	41

Table 3.3: EuroParl$_{parallel}$ test corpus description: the number of mentions, chains, the average chain length, and the number of singletons. Singletons are excluded when computing the statistics over the chains.

3.5 Inter-annotator agreement

Traditionally, inter-annotator agreement is computed for annotators working on the same task and document. In my setup, each annotator was given a document in a different language, but all annotators worked on the same task. Thus, I cannot compute language-specific inter-annotator agreement, and I only present inter-annotator agreement results for the cross-lingual annotations.

Inter-annotator agreement is computed by running two comparisons:

- **Comparison of annotator decision against the final gold standard**: I analyze how often the annotator decisions agree with the reconciled annotations. I compare the agreement of each annotator with the resolved annotations in terms of mention recognition and coreference resolution.

  I compute annotator agreement to the gold standard using the Precision, Recall, and F-measure metrics on named entities. For coreference resolution, I use the MUC,[90] B^3,[5] and CEAF metrics.[49] See Section 4.3.3 for a detailed description of those metrics.

The annotator agreement results on coreference resolution are reported for the test section of the corpus only. Because of an error generated by the annotation software, the annotator training files had offset chain numbers that broke parts of the coreference chains. This problem is fixed in the gold standard files, and did not occur for the annotator test files.

- **Comparison of inter-annotator decisions**: I perform a pairwise comparison on the annotator decisions to evaluate the agreement between annotations on different languages. Because each individual annotator worked on an independent language, this evaluation involves finding an alignment between the languages of interest.

  Given two annotators A and B with associated languages L_A and L_B, I first perform language alignment between the sentences in languages L_A and L_B using the Giza++ software.[66] The alignment process takes as input a pair of manually aligned sentences (s_A, s_B) from languages L_A and L_B respectively, and outputs the word alignment on those sentences. If a sentence s_A is aligned to $\{s_B^1, s_B^2...\}$, than the set of sentences $\{s_B^1, s_B^2...\}$ are concatenated into a single sentence. The sentence alignment is manually generated by the author based on gold standard paragraph alignment available with the raw text of the EuroParl$_{\text{parallel}}$ corpus.

  The output of the word-based alignment process is a set of word pairs $(w_k^{L_A}, w_j^{L_B})$, where either $w_k^{L_A}$ or $w_j^{L_B}$ could be $NULL$, which means no alignment was found for the specific word. I assume that if an aligned word pair is annotated with the same named-entity label, then the two words belong to the same mention in the two different languages. The words that are not aligned by the alignment algorithm are discarded when computing the IAA scores.

  I compute inter-annotator agreement (IAA) on mention recognition using the Cohen's *kappa* metric[11] as well as word-level Precision, Recall, and F-measure over the named-entity annotations.[35] I compare the results of the two metrics for consistencies and disagreements in IAA evaluation.

1. **Cohen's kappa** takes an aligned word pair and defines a correctly labeled alignment as:

   $Match_{\text{named\_entity}}$: word pair where both words are assigned the same category label for named entity, or where both words are not labeled.

   Cohen's kappa is defined as:

   $$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{3.1}$$

   where:

   $Pr(a) = \frac{Matches_{\text{named\_entity}}}{\#words}$ is the observed agreement between the annotators

   $Pr(e)$ is the probability of random agreement between the annotators

   $\#words$ is the total number of aligned words between the two languages.

   Cohen's Kappa has a range from $0 - 1.0$ and larger values represent better annotator reliability. In general, $k > 0.70$ is considered satisfactory. I compute Cohen's Kappa results over the entire set of annotations, without a distinction on the different named-entity categories.

2. **Word-level Precision, Recall, and F-measure** are defined as:

   $$Precision = \frac{\text{\#Correct aligned words from each mention marked by evaluated annotator}}{\text{\#Aligned words marked by evaluated annotator}} \tag{3.2}$$

   $$Recall = \frac{\text{\#Correct aligned words from each mention marked by evaluated annotator}}{\text{\#Aligned words marked by reference annotator}} \tag{3.3}$$

   $$F - measure = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

I define in turns each of the annotators to be the reference annotator, and I consider the remaining annotators to be the evaluated annotator. A *correct* word w_A is a word aligned to a word w_B in the reference annotations that has the same named-entity annotation as w_B.

I compute Precision, Recall, and F-measure results over each named-entity category. I report the overall IAA performance as the unweighted average over Precision, Recall, and F-measure.

3.5.1 Inter-annotator agreement results

Comparison of annotator decision against the final gold standard

Table 3.4 presents the results of evaluating each annotator decisions on named-entity recognition against the gold standard for the respective language. For each of the three languages, the precision results are higher than the recall results (approximatively 95% precision and .90 recall), but the F-measure results are around 93% for all three languages. The high evaluation results for named-entity recognition convey that the annotators are very close to the gold standard in their annotation decisions.

	P	R	F
English	97.7	89.51	93.43
French	94.69	92.41	93.54
German	98.47	89.89	93.98

Table 3.4: Named-entity recognition evaluation against the gold standard. P = Precision, R = Recall, F = F-measure.

Table 3.5 presents the evaluation results for the coreference chains created by the annotator against the gold standard coreference chains. The languages with annotations closest to the gold standard are English and German. Across the three coreference resolution evaluation metrics, the annotators present an approximatively 83% F-measure on English, approximatively 80% F-measure on French, and approximatively 84% F-measure on German.

	MUC			B-CUBED			CEAF		
	P	R	F	P	R	F	P	R	F
English	88.99	77.94	82.81	85.49	78.32	81.75	84.29	84.29	84.29
French	83.49	83.49	83.49	78.44	76.55	77.48	82.89	75.98	79.28
German	96.08	83.49	89.35	90.12	65.81	76.07	84.68	86.77	85.72

Table 3.5: Coreference resolution annotation evaluation against the gold standard. P = Precision, R = Recall, F = F-measure.

Comparison of inter-annotator decisions

Table 3.6 presents the *kappa* results for English-German, English-French, and German-French on the training and test corpus. The IAA results range from 0.65 to 0.77. For the training corpus, the best IAA comes from the English-German language pair (0.77), while the worst IAA is observed for the German-French language pair (0.73). The test corpus has larger IAA results, with a best IAA of 0.87 on the English-German language pair. The observed *kappa* values for the English-French language pair of the training corpus are larger due to the larger percentage of words that are not part of a mention and are annotated by both annotators with a *Not Mention* tag: approximatively 80% of the word pairs in the training corpus are labeled with *Not Mention* by both annotators, compared to 74% of the word pairs in the test corpus. In general, the IAA results show a satisfactory agreement of annotations made across the three languages (i.e., English, French, and German) on both the training and test corpus.

Table 3.7 presents the IAA results in terms of precision, recall, and F-measure. In general, the IAA results are higher when the reference annotator is the English annotator, mainly due to better word-alignenment results. The difference in IAA results when the reference and evaluated annotator are switched is 1% for English-German and for English-French, and 7% for German-French. The *kappa* results and the overall unweighted F-measure IAA results are consistent with each other: 0.87 *kappa* vs. 0.88 overall unweighted F-measure for English-German, 0.71 *kappa* vs. 0.71 overall unweighed F-measure for English-French, and 0.75 *kappa* vs. 0.74 overall unweighted F-measure for German-French. In general, the person category has the

Language Pair	*kappa*	Not Mention
Training set		
English - German	0.77	80%
English - French	0.75	80%
German - French	0.73	78%
Test set		
English - German	0.87	74%
English - French	0.71	75%
German - French	0.75	74%

Table 3.6: IAA results on named-entity recognition: *kappa* and percentage of word pairs labeled with *Not Mention* by both annotators.

highest IAA F-measure, followed by the location, and organization categories.

Reference	Evaluated	Person			Location			Organization			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
English	German	0.95	0.91	0.93	0.89	0.89	0.89	0.86	0.79	0.82	0.9	0.86	0.88
German	English	0.91	0.90	0.90	0.89	0.87	0.88	0.96	0.82	0.84	0.92	0.86	0.87
English	French	0.89	0.70	0.79	0.62	0.84	0.71	0.72	0.53	0.62	0.74	0.69	0.70
French	English	0.74	0.87	0.81	0.88	0.61	0.72	0.54	0.68	0.61	0.72	0.72	0.71
German	French	0.93	0.95	0.94	0.55	0.76	0.64	0.79	0.54	0.64	0.75	0.75	0.74
French	German	0.75	0.87	0.81	0.81	0.45	0.58	0.57	0.70	0.63	0.71	0.67	0.67

Table 3.7: Precision (P), Recall (R), and F-measure (F) IAA results on named-entity recognition.

3.6 Discussion

In general, the annotators found the content of the documents less ambiguous when annotating documents in English. I consequently ran an experiment where annotators were first given an English document to annotate and then handed the German equivalent of the same document. Both the English and the German documents were new to the annotator working on them. The German document was much easier to work with after the annotator had already annotated the document in English, in terms of both the time required for annotation and the amount of uncertainty in annotations. Thus, even for the human experts there is a valuable gain to be achieved in retrieving information from parallel documents simultaneously, compared to retrieving information from a single language.

The process of reconciliation showed that annotator disagreements are caused by language subtleties and are inherently difficult to resolve. Depending on context and language, mentions would be missing or would not be annotated for coreference as the context did not explicitly include reference to a previous mention. Some of the languages were more verbose, and consequently it was more difficult to track the coreference links and to correctly identify the entire span of a mention.

Given that named-entity recognition and coreference resolution are challenging tasks for human experts, I expect automated systems to be similarly hindered by the complexity of those tasks.

3.7 Conclusions

I present a parallel multilingual corpus annotated for named entities and coreference resolution in English, German, and French. The annotations are evaluated for quality and consistency. I also investigate the difficulty of retrieving information from a text available in a single language compared to multiple languages. The generated annotations are consistent across the three languages with an average *kappa* of 0.75 on the training corpus and 0.77 on the test corpus. From empirical observations I conclude that the human experts benefit from using parallel text available in multiple languages.

The multilingual corpus is made available for research purposes at `http://web.mit.edu/andreeab/www/multilingual_coreference.html`.

Chapter 4

Multilingual end-to-end coreference resolution

4.1 Chapter overview

Humans make use of prior knowledge when communicating through natural language. This knowledge can be grouped intro linguistic knowledge that helps form the discourse, and contextual knowledge that helps disambiguate discourse meaning. In general, the human brain can perform multiple language-related tasks simultaneously - like reading a text and following the main character's evolution, extracting relationships between concepts mentioned inside the text, or performing disambiguation of ambiguous language. This behavior is not currently implemented in many NLP systems, as most processing is done sequentially rather than in parallel. I aim to tackle the challenge of knowledge incorporation and joint solving of related NLP tasks across languages with respect to the named-entity recognition and coreference resolution task.

My goal is to develop a coreference resolution system that builds on the syntactic knowledge previously discussed, and further integrates contextual knowledge in the form of soft linguistic constraints induced in a Bayesian setting. The model presented in this chapter performs joint learning of named entities and coreference resolution relations.

In the remainder of this chapter, I present an introduction and overview of related work on coreference resolution, and specifically on multilingual coreference resolution (see Section 4.2 and Section 4.3). I introduce the end-to-end multilingual coreference resolution system in Section 4.4, followed by the experimental setup in Section 4.5. I discuss the experiment results (Section 4.6), followed by a detailed analysis of model performance across the investigated languages (Section 4.7), and I end the chapter with some conclusions in Section 4.8.

4.2 Introduction

Coreference resolution determines whether two expressions are coreferent, that is, linked by an *identity* or *equivalence* relation. In order to develop coreference resolution systems, one must first obtain the expressions of interest. In the first scenario, the system receives the expressions of interest as input and identifies the relevant relations. In the second scenario (the end-to-end coreference resolution) the system identifies expressions from text and performs coreference resolution on the identified expressions. The problem encountered by both scenarios is the lack of information flow between the NLP system solving the named-entity recognition task and the NLP system solving the coreference resolution. A coreference resolution system could guide its decisions on the retrieval of coreference chains using the decisions made on the named entities, and optimize the entire process across the two layers of retrieved information.

Coreference resolution systems can be built solely from the training data provided. They may also include unbounded amounts of additional world knowledge from external sources such as web sites, dictionaries, ontologies, etc. In general, external knowledge is required by coreference resolution systems in order to compensate for the lack of contextual information required for anaphora disambiguation. An ideal language processing framework would allow for context-sensitive modeling of language, but the computational complexity of context-sensitive models is a large burden for natural language processing systems.

In this thesis I approach the problem of joint modeling of named-entity recognition and coreference resolution in an end-to-end coreference resolution system using linguistic knowledge induced through soft constraints. I further discuss multilingual approaches to named-entity recognition and I present a review of the state-of-the-art in multilingual coreference resolution.

4.3 Related work

In order to understand the progress made by the research community in multilingual coreference resolution, I review the two steps usually undertaken by coreference resolution systems: named-entity recognition and coreference resolution on the identified named entities.

4.3.1 Multilingual named-entity recognition

Several research initiatives have investigated the named-entity recognition task for parallel corpora. Yarowsky et al.[95] investigated the feasibility of projecting English named entity data over French, Chinese, Czech, and Spanish. They showed that resources developed on a source language can be used to automatically induce stand-alone text analysis tools. With a slightly different goal, Klementiev and Roth (2008) proposed an algorithm for cross-lingual multiword named-entity discovery in a bilingual weakly temporally aligned corpus. Richman et al.[74] used the multilingual properties of Wikipedia to annotate a corpus with named entities using little human intervention and no linguistic expertise. Shah et al.[78] used machine translation techniques to develop a named-entity recognition system in which named-entity annotations are projected based on word-alignment. In general, the methods proposed for multilingual named-entity recognition assume the availability of a multilingual parallel or synchronous comparable corpora on which learning is based.[41, 40] Alternatively, they are based the model development on top of other linguistic processing resources,[73] which might not be available for resource-poor languages.

4.3.2 Multilingual coreference resolution

The field of multilingual coreference resolution has moved forward due to multilingual coreference resolution shared tasks. The first shared task was organized during the ACE 2004/2005 evaluations. It was followed by the SemEval-2010 Multilingual Shared Task 1[73] which addressed coreference resolution in six languages (i.e., Catalan, Dutch, English, German, Italian, Spanish), and the 2012 CoNLL Shared Task[70] which handled only three languages (i.e., English, Chinese, Arabic). Both shared tasks evaluate system performance without making a difference on the category of the named entities. Results are reported across all possible chains instead of making a separate classification of performance based on the type of the named entities.

The SemEval-2010 Shared Task provided multilingual corpora annotated for several layers of linguistic knowledge including part-of-speech, morphological features, syntactic dependency information, semantic information, named entities, and coreference resolution. The Shared Task evaluated the participating system's performance when the provided linguistic knowledge was generated by human experts (gold-standard setting) as well as when the linguistic knowledge was generated by state-of-the-art NLP systems (regular setting). It also considered how systems performed when integrating knowledge outside the provided corpora (open setting) versus when using only corpus-specific information (closed setting). Two of the best performing systems in the SemEval-2010 shared task are the SUCRE[42] and the UBIU system.[98] The SUCRE system is a supervised classification system trained in a monolingual setting (i.e., it requires gold standard annotations for the target language). It improves on previous state-of-the art systems by its feature engineering technique built on top of a relational database. On the task of coreference resolution, SUCRE obtains a 67.3 unweighted average F-measure on Catalan, 72.5 unweighted average F-measure on English, 67.5 unweighted average F-measure on Spanish, and 65.2 unweighted average F-measure on Dutch under the closed gold-standard settings. It obtains 45.2 unweighted average F-measure on Catalan, 60.76 unweighted average F-measure on English, 48.26 unweighted average F-measure on Spanish, and 19.1

unweighted average F-measure on Dutch in the closed regular setting. Similar to SUCRE, the UBIU system is also a classification system that makes use of a rich feature set extracted from the corpora and the available annotations. UBIU obtains coreference resolution scores ranging from 40.93 unweighted average F-measure on Catalan to 53.66 on English in the closed gold-standard setting, and from 29.3 unweighted average F-measure on Catalan to 42.16 on English in the closed regular setting.

A common trend of the shared tasks was the design of supervised monolingual coreference resolution systems, or generic coreference resolution systems easily adapted to the target language. The problem of these approaches is the expectation that annotated resources are available in the source languages. One could replace the supervised approaches with unsupervised or heuristic methods. But in this case the burden is placed on external linguistic knowledge necessary for designing coreference rules[71, 57] or generative models.[29] To address the corpus annotation bottleneck, Rahman and Ng[72] use a translation-based projection approach to coreference resolution. The authors first automatically translate the target language into the source language, produce annotations in the translated source language text using a language specific coreference resolver. Finally they project the annotations from the source language to the target language. Their system achieved 90% of the performance of a supervised coreference resolver in Spanish and Italian, when only a mention extractor for the target language was available. A projection-based approach was also taken by de Souza and Orăsan[22], but the authors used parallel corpora for performing the annotation projection. Harabagiu and Maioreanu[32] discuss the performance of a system trained on a bilingual English-Romanian corpus that outperforms the coreference resolution results of a monolingual baseline.

4.3.3 Evaluation metrics

Mention evaluation

System performance on mention extraction is commonly evaluated using the precision (P), recall (R), and F-measure (F) metrics. These metrics are computed based on the true positives (TPs), false positives (FPs), and false negatives (FNs) of named entities retrieved by a system. I define the TP, FP, and FN differently for mentions that exactly overlap the gold standard mentions (i.e., exact overlap), and for mentions that at least partially overlapped the gold standard mentions (i.e., at least partial overlap).

For exact overlap, I define TP, FP, and FN as:

- TP: system mentions that exactly match with a gold standard mention annotation, in both word offset and named-entity category.
- FP: system mentions that do not exactly agree with any gold standard mention annotation, in either word offset or named-entity category.
- FN: gold standard mention annotations that do not exactly agree with any system mention annotation, in either word offset or named-entity category.

For at least partial overlap, I define TP, FP, and FN as:

- TP: system mentions that at least partially match with a gold standard mention annotation, in both word offset and named-entity category.
- FP: system mentions that do not at least partially agree with any gold standard mention annotation, in either word offset or named-entity category.
- FN: gold standard mention annotations that do not at least partially agree with any system mention annotation, in either word offset or named-entity category.

Precision, recall, and F-measure are then defined as:

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

$$F\text{-}measure = \frac{2 * precision * recall}{recall + precision} \tag{4.3}$$

Coreference resolution evaluation

I evaluate the systems' performance on coreference resolution using three evaluation metrics: MUC,[90] B^3,[5] and CEAF.[49] Following common practices in the literature,[88, 70] I use the unweighted average of the MUC, B^3, and CEAF metrics as a measure of system performance on coreference chains.

The MUC metric evaluates the set of system chains by looking at the minimum number of coreference pair additions and removals required to match the gold standard coreference pairs. The pairs to be added represent false negatives, while the pairs to be removed represent false positives. Let K represent the gold standard chains set, and R the system chains set. Given chains k and r from K and R, respectively, MUC recall and precision of R are:

$$Recall_{MUC} = \frac{\sum_k (|k| - m(k, R))}{\sum_k (|k| - 1)} \tag{4.4}$$

$$Precision_{MUC} = \frac{\sum_k (|r| - m(k, K))}{\sum_k (|r| - 1)} \tag{4.5}$$

where $m(r, K)$, by definition, represents the number of chains in K that intersected the chain r. The MUC F-measure is given by:

$$F\text{-}measure = \frac{2 * precision * recall}{recall + precision} \tag{4.6}$$

B^3 metrics evaluate system performance by measuring the overlap between the chains predicted by the system and the gold standard chains. Let C be a collection of n documents, d a document in C, and m a mention in document d. I define the gold standard chain that includes m as G_m and the system chain that contains m as

S_m. O_m is the intersection of G_m and S_m. B^3 recall and precision are defined as:

$$Recall_{\mathrm{B}^3} = \frac{1}{n} \sum_{d \in C} \sum_{m \in d} \frac{|O_m|}{|G_m|} \tag{4.7}$$

$$Precision_{\mathrm{B}^3} = \frac{1}{n} \sum_{d \in C} \sum_{m \in d} \frac{|O_m|}{|S_m|} \tag{4.8}$$

The B^3 F-measure is defined identically to the MUC F-measure, but on the B^3-specific definitions of precision and recall.

The CEAF metric first computes an optimal alignment ($\boldsymbol{\phi}(g^*)$) between the system chains and the gold standard chains based on a similarity score. This score could be based on the mentions or on the chains. The chains-based score has two variants, ϕ_3 and ϕ_4; in reporting the results for this thesis I use ϕ_4, unless otherwise specified.

$$\phi_3 = |K_i \cap R_j| \tag{4.9}$$

$$\phi_4 = \frac{2 * |K_i \cap R_j|}{|K_i| + |R_j|} \tag{4.10}$$

The CEAF precision and recall are defined as:

$$Recall_{CEAF} = \frac{\boldsymbol{\phi}(g^*)}{\sum_i \phi(R_i, R_i)} \tag{4.11}$$

$$Precision_{CEAF} = \frac{\boldsymbol{\phi}(g^*)}{\sum_i \phi(K_i, K_i)} \tag{4.12}$$

The CEAF F-measure is defined identically to the MUC F-measure, but on the CEAF-specific definitions of precision and recall.

4.4 Model overview

I design a system for joint learning of named entities and coreference relations from a given document. I represent my solution using a graphical model, specifically a factorial hidden markov model (FHMM),[26] in which separate hidden node variables are used to model the named entities and coreference relations, respectively. My approach is motivated by the work of Li et al. for pronoun anaphora resolution.[48] I extend their work to noun-phrase mentions, and use a different model representation in order to reduce the number of externally-induced model constraints. Specifically, I encode in the model an additional queue storing the mention history, that helps capture context-sensitive knowledge. In the rest of this chapter, I refer to my named entity and coreference resolution FHMM model by *NECR*.

NECR is trained over entire sentences, where each time step t corresponds to a word position inside the sentence and the transitions between time steps correspond to sequentially moving through the words in a sentence. For the languages I consider in this thesis, the states transitions are equivalent to moving left-to-right through a sentence.

The *NECR* model consists of two hidden states (mention hidden state m and coreference hidden state *cr*) and two observed states (part-of-speech observed state *pos* and dependency *head* observed state *dep*) (see Figure 4.1). The m state has three possible values {*beginning, inside, outside*}; the *beginning* state value specifies the beginning of a mention; the *inside* state value specifies the inside of a mention; and the *outside* state value specifies there is no mention. At each time step, the model stores the most recently discovered mentions in a history queue H. The history queue stores at most $n = 6$ mentions, in order to limit the model complexity. As new mentions are discovered, they are added to the history queue, and the last mention inside the history queue is removed if the queue is full.

The *cr* state is defined in relation to the mentions stored inside the history queue. It specifies the mention inside the queue to which the mention at the current state corefers. The *cr* state takes values from 1 to n, or *none*, where n is the size of the history queue. Each state value represents an index inside the queue, or no index (*none*) when no coreference relation applies.

It is worth pointing out that some of the *NECR* hidden state transitions are not possible. For example, the m state cannot generate a sequence *outside* → *inside* → *beginning* and the *cr* state cannot generate a sequence such as *none* → *2* → *3* → *2*. Similarly, a *cr* state with a value other than *none* cannot exist at time step t, unless the m state at time step t has a value different from *outside*. Such restrictions on the possible state sequences are captured by a set of soft constraints.

The *NECR* model has two observations: part-of-speech *pos* and dependency *head* observed state *dep*. The *pos* observed state takes values from the universal POS tag set discussed in Chapter 2. The *dep* observed state takes values from 1 to the sentence length, and represents the *head* position inside the sentence.

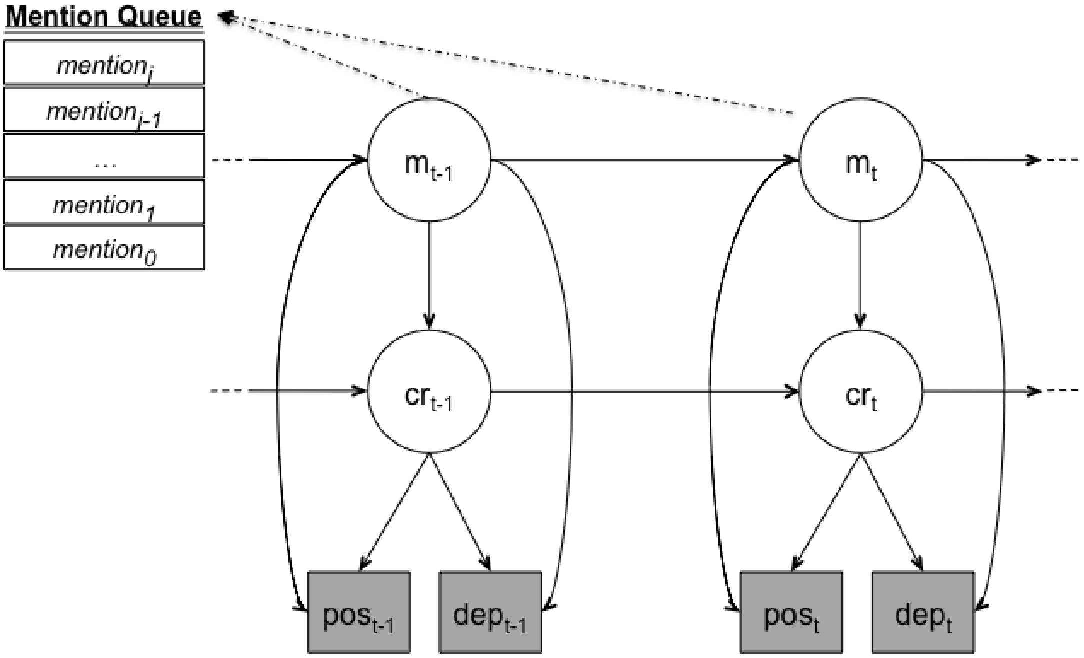

Figure 4.1: *NECR* model design

The transition probability of the *NECR* model is defined in Equation 4.13. The observed states depend on the two hidden states at each time step. In designing the *NECR* model, I assume that the two observed states are independent of each other given the hidden states. The observation model is defined in Equation 4.14.

$$P(hidden\_state_t|hidden\_state_{t-1}) = P(m_t|m_{t-1}) * P(cr_t|cr_{t-1}, m_t) \tag{4.13}$$

$$P(observation_t|hidden\_state_t) = P(pos_t|m_t, cr_t) * P(dep_t|m_t, cr_t) \tag{4.14}$$

My *NECR* model is delexicalized, as the observations are constructed from a combination of universal part of speech and dependency attributes at each time step t and no lexical information is used during the model learning stage. The model assumes the existence of universal POS tags and dependency annotations. The SemEval corpus has associated POS annotations for each language, and I generate POS annotations for the EuroParl$_{parallel}$ corpus using the state-of-the-art Stanford POS Tagger.[86] Because the input corpora are not annotated with universal POS tags, I map the corpus specific POS tags to universal tags using the mapping proposed by Naseem et al.[60] I also generate dependency annotations for each sentence inside a corpus using the $Predictor_{KL\_BEST}$ model presented in Chapter 2, in order to have a universal parser for all languages in the input corpora.

The FHMM model is enhanced with a set of soft constraints that guide the learning process. The use of soft constraints has been previously shown to improve performance of HMM models on the task of named-entity recognition[12] and relation extraction.[8] A sample set of the constraints used in the *NECR* model is presented in Table 4.1.

The *NECR* model is implemented on top of the HMM-learning with constraints framework available from [12].

Name	Description
CoreferMaintained	The coreference label has to be maintained across a mention
CoreferInQueue	If a coreference label is assigned, then a mention must exist at the respective position inside the queue
CorefNoEntity	Coreference labels cannot be assigned if a named-entity label does not exist at the specific time point
NEMaintained	The named-entity label has to be maintained across a mention
NoPunct	Punctuation signs should not be annotated
NoStartInside	Named entities cannot begin with an *inside* label

Table 4.1: Sample *NECR* model constraints.

4.5 Experiments

4.5.1 Corpus and Annotations

I evaluate the performance of the *NECR* model on the coreference resolution treebank released by the SemEval-2010 Shared Task.[73] The treebank contains newswire texts for Catalan, Spanish, Dutch, English, German, and Italian. It is annotated on several linguistic layers including named entities, coreference resolution, syntactic dependency trees, prepositions, and word sense. The treebank is gold-standard annotated for named entities and coreference resolution for all of the languages. The other layers of annotations are generated using state-of-the-art automated systems. I exclude Italian when running comparison experiments in this thesis, as there is no granular-to-universal POS mapping available for the granular Italian POS tagset. The analysis presented in this thesis also excludes the German language, as it was not available for public release together with the remaining languages.

Each SemEval language-specific corpus was annotated based on independent annotation guidelines. Consequently, certain inconsistencies in annotations will be present. In a cross-lingual learning setting those inconsistencies might hinder the system performance. The named entities are not annotated for category, but only for the span of the named entity.

The number of documents, sentences, and tokens for each of the SemEval languages used in this thesis is included in Table 4.2. The largest portion of the corpus is represented by the Spanish language, with 875 documents and $284,179$ tokens in the training set and 168 documents and $51,040$ tokens in the test set, followed by Catalan with 829 documents and $253,513$ tokens in the training set and 167 documents and $49,260$ tokens in the test set.

Language	Type	Train	Development	Test
Catalan	Documents	829	142	167
	Sentences	8,709	1,445	1,698
	Tokens	253,513	42,072	49,260
Spanish	Documents	875	140	168
	Sentences	9,022	1,419	1,705
	Tokens	284,179	44,460	51,040
English	Documents	229	39	85
	Sentences	3,648	741	1,141
	Tokens	79,060	17,044	24,206
Dutch	Documents	145	23	72
	Sentences	2,544	496	2,410
	Tokens	46,894	9,165	48,007

Table 4.2: Number of documents, sentences, and tokens in the 2010 SemEval corpus used in this thesis.

I also report system results for named-entity recognition and coreference resolution on the EuroParl$_{parallel}$ corpus developed by the author.

4.5.2 Experiment setup

In order to analyze the *NECR* model performance and identify its specific contributions, I evaluate the system under the following experimental settings:

Setting 1: Monolingual system I investigate how the *NECR* model performs when trained in a monolingual setting. Under this setting, the model *dep* observed state values are obtained from the multilingual *Predictor*$_{KL\_BEST}$ parser. I report the model performance for both named-entity recognition and coreference resolution on the EuroParl$_{\text{parallel}}$ and SemEval corpora.

I refer to this setting of the model as the $NECR^{monolingual}_{KL\_BEST}$ model.

Setting 2: Monolingual system in cross-lingual evaluation I investigate how the monolingual$NECR^{monolingual}_{KL_BEST}$ model performs when directly transferred to other languages. Specifically, I use the $NECR^{monolingual}_{KL\_BEST}$ model trained in a monolingual setting on language L, and report its performance on the remaining languages $L_k \neq L$. I report the model performance for both named-entity recognition and coreference resolution on the EuroParl$_{\text{parallel}}$ and SemEval corpora.

Setting 3: Monolingual training with language specific parsers In general, NLP systems trained on gold standard annotations perform better than systems for which training data is not available. This experiment investigates whether the multilingual dependency parsers are detrimental to the final *NECR* system performance, or whether they are as good or better than the parsers trained on language-specific annotations (the language specific parsers).

In order to obtain language specific parsers I learn a dependency parsing model from the available gold standard annotations for each specific language using the MSTParser. The parser is trained on the CoNLL 2006/2007 corpus for each respective language. The parser is thus transferred across corpora. I use the output from the language-specific parsers for the observed states of the *NECR* system, and refer to this setting as the $NECR^{monolingual}_{MST}$ model.

I report model results on the SemEval corpus only, because gold standard dependency annotations are not available for the EuroParl$_{\text{parallel}}$ corpus and I consequently cannot train a language-specific parser for the latter corpus.

Setting 4: Multilingual source training In this experiment, I investigate whether joint training on more than one source language can help improve the model performance on a target language. I train the *NECR* with state values obtained from the $Predictor_{KL\_BEST}$ parser on a subset $S_{source\_training}$ of k source languages and report its performance on the remaining target languages $T_j \notin S_{source\_training}$.

I experiment with different values for the number of source training languages k and allow source languages to be selected from (1) the EuroParl$_{\text{parallel}}$ corpus,

(2) the SemEval corpus, or (3) from a combination of the two corpora. I report results on the target languages from the EuroParl$_{\text{parallel}}$ corpus only, in order to facilitate an informed analysis of model performance without variance caused by differences in the input texts specific to each language.

I refer to this experimental setting of the *NECR* model as $NECR^{multi\_source}_{KL\_BEST}$.

4.6 Results

4.6.1 Setting 1: Monolingual system

System results on the EuroParl$_{\text{parallel}}$ corpus

The named-entity recognition results of the $NECR^{monolingual}_{KL\_BEST}$ model on the EuroParl$_{\text{parallel}}$ corpus are included in Table 4.3. The system evaluates at 43.18 F-measure on English, 21.83 F-measure on French, and 10.98 F-measure on German on exact overlap. The partial overlap results are better across all languages: 59.84 F-measure on English, 58.07 F-measure on French, and 18.43 F-measure on German for partial overlap.

	Exact			Partial		
	P	R	F	P	R	F
English	48.3	39.04	43.18	66.94	54.10	59.84
French	29.76	17.24	21.83	79.16	45.86	58.07
German	27.85	6.77	10.98	47.14	11.45	18.43

Table 4.3: $NECR^{monolingual}_{KL\_BEST}$: named-entity recognition results on the EuroParl$_{\text{parallel}}$ corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) over exact and partial overlap.

Table 4.4 presents the $NECR^{monolingual}_{KL\_BEST}$ coreference resolution results on the EuroParl$_{\text{parallel}}$ corpus. The unweighted average F-measure is 19.98 on English, 8.11 on French, and 3.61 on German. In general, the system evaluates better in terms of the $CEAF$ metric, reporting a 22.58 $CEAF$ F-measure on English, 13.1 $CEAF$ F-measure on French, and 6.14 $CEAF$ F-measure on German.

	MUC			B^3			CEAF			Overall		
	P	R	F	P	R	F	P	R	F	P	R	F
English	16.56	12.56	14.28	25.84	20.89	23.1	23.43	21.78	22.58	21.93	18.41	19.98
French	7.27	1.94	3.06	12.53	6.06	8.17	17.85	10.34	13.1	12.55	6.11	8.11
German	5.71	0.96	1.65	12.08	1.75	3.06	15.71	3.81	6.14	11.16	2.17	3.61

Table 4.4: $NECR^{monolingual}_{KL\_BEST}$: coreference resolution results on the $\text{EuroParl}_{\text{parallel}}$ corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) of the MUC, B^3, and CEAF metrics respectively, as well as the unweighted average of precision, recall, and F-measure over the three metrics.

System results on the SemEval corpus

The named-entity recognition results of the $NECR^{monolingual}_{KL\_BEST}$ model on the SemEval corpus are included in Table 4.5. For exact overlap, the system reports a 27.3 F-measure on Catalan, a 33.13 F-measure on Spanish, a 52.56 F-measure on English, and a 28.94 F-measure on Dutch. The partial overlap results are larger across all languages, with a 55.39 F-measure on Catalan, a 59.32 F-measure on Spanish, a 75.47 F-measure on English, and a 53.64 F-measure on Dutch. For English, the named-entity recognition results on the SemEval corpus are larger than the results on the $\text{EuroParl}_{\text{parallel}}$ corpus. This behavior is explained by the larger size of the SemEval training corpus, compared to the $\text{EuroParl}_{\text{parallel}}$ training corpus.

	Exact			Partial		
	P	R	F	P	R	F
Catalan	27.75	26.87	27.3	56.3	54.52	55.39
Spanish	30.83	35.8	33.13	55.2	64.11	59.32
English	50.46	54.6	52.56	72.75	78.42	75.47
Dutch	26.94	31.26	28.94	49.94	57.95	53.64

Table 4.5: $NECR^{monolingual}_{KL\_BEST}$: named-entity recognition results on the SemEval corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) over exact and partial overlap.

The coreference resolution results of the $NECR^{monolingual}_{KL\_BEST}$ model on the SemEval corpus are included in Table 4.6. The system evaluates at 15.15 unweighted average F-measure on Catalan, 17.11 on Spanish, 35.70 on English, and 12.63 on Dutch. English is the language with the best coreference resolution performance, while Dutch is the language with the lowest results for coreference resolution.

	MUC			B^3			CEAF			Overall		
	P	R	F	P	R	F	P	R	F	P	R	F
Catalan	2.98	3.34	3.15	21.6	20.92	21.25	22.61	19.69	21.05	15.73	14.65	15.15
Spanish	2.35	3.41	2.78	20.3	25.69	22.68	26.1	25.64	25.87	16.26	18.24	17.11
English	31.43	17.53	22.51	44.68	41.76	43.17	36.57	47.82	41.44	37.56	35.70	35.70
Dutch	3.14	3.73	3.41	13.74	16.47	14.98	18.51	20.63	19.51	11.79	13.61	12.63

Table 4.6: $NECR^{monolingual}_{KL\_BEST}$: coreference resolution results on the SemEval corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) of the MUC, B^3, and CEAF metrics respectively, as well as the unweighted average of precision, recall, and F-measure over the three metrics.

Across both corpora, English obtains the best results on coreference resolution due to the fact that it also manages to identify the highest percentage of exact overlap mentions. The rest of the languages exhibit lower results on named entity recognition, and those results impact the final performance on coreference resolution.

4.6.2 Setting 2: Monolingual system in cross-lingual evaluation

System results on the EuroParl$_{\textbf{parallel}}$ corpus

Table 4.7 presents the $NECR^{monolingual}_{KL\_BEST}$ results for named-entity recognition on the EuroParl$_{parallel}$ corpus. The system performance on the target language varies based on the source language. For all three languages, the best source $NECR^{monolingual}_{KL\_BEST}$ model is English when the task is evaluated over exact overlap (43.18 F-measure on English, 36.05 F-measure on French, and 24.2 F-measure on German). The best source $NECR^{monolingual}_{KL\_BEST}$ is French when the task is evaluated over partial overlap (64.39 F-measure on English, 58.07 F-measure on French, and 49.31 F-measure on German). When German is the source language, the $NECR^{monolingual}_{KL\_BEST}$ model reports the lowest results on all the target languages, including on German.

The $NECR^{monolingual}_{KL\_BEST}$ coreference resolution results on the EuroParl$_{\text{parallel}}$ corpus are presented in Table 4.8. The best performing $NECR^{monolingual}_{KL\_BEST}$ model is based on the English source language for all target languages. It reports an unweighted average F-measure of 19.98 on English, 21.78 on French, and 8.72 on German. From

		Target					
		English					
		Exact			Partial		
		P	R	F	P	R	F
Source	English	48.3	39.04	43.18	66.94	54.10	59.84
	French	43.55	36.46	40.53	72.03	58.21	64.39
	German	34.61	9.24	14.59	48.71	13.01	20.54
		French					
		Exact			Partial		
		P	R	F	P	R	F
Source	English	42.68	31.2	36.05	64.15	46.89	54.18
	French	29.76	17.24	21.83	79.16	45.86	58.07
	German	30.86	8.62	13.47	46.91	13.10	20.48
		German					
		Exact			Partial		
		P	R	F	P	R	F
Source	English	18.67	34.37	24.2	34.52	63.54	44.74
	French	17.26	26.73	20.98	40.58	62.84	49.31
	German	27.85	6.77	10.98	47.14	11.45	18.43

Table 4.7: $NECR^{monolingual}_{KL\_BEST}$: named-entity recognition results on the EuroParl$_{\text{parallel}}$ corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) over exact and partial overlap. The row labels represent the source language, and the column labels represent the target language.

all the target languages, the highest scoring language is French when the English $NECR^{monolingual}_{KL\_BEST}$ model is used (21.78 unweighted average F-measure). German proves to be the most difficult language to model for coreference resolution: when used as a source language, it does not manage to perform better than any of the other source languages.

System results on the SemEval corpus

Table 4.9 presents the $NECR^{monolingual}_{KL\_BEST}$ named-entity recognition results on the SemEval corpus. The best results on exact overlap come from the English $NECR^{monolingual}_{KL\_BEST}$ model for all target languages: 49.13 F-measure on Catalan, 49.2 F-measure on Spanish, 52.56 F-measure on English, and 34.39 F-measure on Dutch. The English $NECR^{monolingual}_{KL\_BEST}$ model reports the best partial overlap results on the Catalan, Spanish, and English target languages (67.31 F-measure, 69.08 F-measure, and 75.47 F-measure, respectively). The best source $NECR^{monolingual}_{KL\_BEST}$ model for Dutch is the Dutch $NECR^{monolingual}_{KL\_BEST}$ model, with a 53.64 F-measure on partial overlap.

		Target											
		English											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	English	16.56	12.56	14.28	25.84	20.89	23.1	23.43	21.78	22.58	21.93	18.41	19.98
	French	7.23	5.31	6.12	18.05	11.84	14.3	21.18	17.12	18.93	15.48	11.42	13.11
	German	10	2.41	3.89	14.08	2.89	4.8	21.79	5.82	9.18	15.29	3.71	5.96
		French											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	English	16.99	12.62	14.48	12.95	17.28	14.81	30.28	44.82	36.05	16.74	14.90	21.78
	French	7.27	1.94	3.06	12.53	6.06	8.17	17.85	10.34	13.1	12.55	6.11	8.11
	German	3.84	0.97	1.55	6.91	2.28	3.43	16.04	4.48	7	12.26	2.57	3.99
		German											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	English	6.01	13.04	8.23	4.25	15.78	6.69	8.67	15.97	11.24	6.31	14.93	8.72
	French	0.34	0.48	0.4	6.54	8.69	7.4	9.86	15.27	11.98	5.58	8.14	6.59
	German	5.71	0.96	1.65	12.08	1.75	3.06	15.71	3.81	6.14	11.16	2.17	3.61

Table 4.8: $NECR^{monolingual}_{KL\_BEST}$: coreference resolution results on the EuroParl$_{\text{parallel}}$ corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) of the MUC, B^3, and CEAF metrics respectively, as well as the unweighted average of precision, recall, and F-measure over the three metrics. The row labels represent the source language, and the column labels represent the target language.

Table 4.10 presents the $NECR^{monolingual}_{KL\_BEST}$ coreference resolution results on the SemEval corpus. The best system performance is given by the English $NECR^{monolingual}_{KL\_BEST}$ model, with a 30.53 unweighted average F-measure on the Catalan target language, 30.80 on the Spanish target language, 35.70 on the English target language, and 18.83 on the Dutch target language. The second best performing $NECR^{monolingual}_{KL_BEST}$model is the Spanish $NECR^{monolingual}_{KL\_BEST}$ model for the Catalan, Spanish, and English target languages, and the Dutch $NECR^{monolingual}_{KL\_BEST}$ model for the Dutch target language.

4.6.3 Setting 3: Monolingual training with language specific parsers

Table 4.11 presents the $NECR^{monolingual}_{MST}$ named-entity recognition results for the SemEval corpus, when the *dep* observed variable is obtained from the MSTParser. The best performing $NECR^{monolingual}_{MST}$ model is the English $NECR^{monolingual}_{MST}$ model for the Catalan, Spanish, and English target languages, while Dutch is best predicted by the Dutch $NECR^{monolingual}_{MST}$ model. The best exact overlap results are 48.66 F-measure on

		Target					
		Catalan					
		Exact			Partial		
		P	R	F	P	R	F
Source	Catalan	27.75	26.87	27.3	56.3	54.52	55.39
	Spanish	31.28	36.5	33.69	64.07	54.91	59.14
	English	44.06	55.51	49.13	60.37	76.05	67.31
	Dutch	37.63	22.94	28.51	65.62	40.01	49.71
		Spanish					
		Exact			Partial		
		P	R	F	P	R	F
Source	Catalan	28.39	26.12	27.21	57.57	52.95	55.16
	Spanish	30.83	35.8	33.13	55.2	64.11	59.32
	English	45.11	54.11	49.2	63.34	75.97	69.08
	Dutch	38.65	22.72	28.62	66.20	38.92	49.02
		English					
		Exact			Partial		
		P	R	F	P	R	F
Source	Catalan	27.34	29.05	28.17	53.5	56.85	55.12
	Spanish	26.81	34.28	30.09	49.76	63.63	55.84
	English	50.46	54.6	52.56	72.75	78.42	75.47
	Dutch	37.5	23.85	29.15	69.38	44.11	53.93
		Dutch					
		Exact			Partial		
		P	R	F	P	R	F
Source	Catalan	16.04	31.15	21.26	33.20	65.22	44.01
	Spanish	16.86	40.2	23.75	30.21	72.04	42.57
	English	25.82	51.47	34.39	39.00	77.72	51.93
	Dutch	26.94	31.26	28.94	49.94	57.95	53.64

Table 4.9: $NECR_{KL\_BEST}^{monolingual}$: named-entity recognition results on the SemEval corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) over exact and partial overlap. The row labels represent the source language, and the column labels represent the target language.

target Catalan, 48.35 F-measure on target Spanish, and 51.15 F-measure on target English. The Dutch $NECR_{MST}^{monolingual}$ model obtains a 28.94 F-measure on Dutch. The best partial overlap results are 67.87 F-measure on target Catalan, 68.37 F-measure on target Spanish, and 74.63 F-measure on target English. The Dutch $NECR_{MST}^{monolingual}$ model reports the best partial overlap results on the Dutch target language, with a 55.00 F-measure.

The $NECR_{MST}^{monolingual}$ coreference resolution results on the SemEval corpus are presented in Table 4.12. The English $NECR_{MST}^{monolingual}$ model performs best on all the target languages. On Catalan, it reports a 31.34 unweighted average F-measure, a 32.17 unweighted average F-measure on Spanish, a 36.11 unweighted average F-

		Target											
		Catalan											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	Catalan	2.98	3.34	3.15	21.6	20.92	21.25	22.61	19.69	21.05	15.73	14.65	15.15
	Spanish	2.53	3.57	2.96	21.26	26.48	23.58	27.09	27.45	27.27	16.96	19.16	17.93
	English	23.66	14.38	17.89	38.95	36.89	37.89	28.64	47.85	35.83	30.41	33.04	30.53
	Dutch	4.07	3.97	4.02	19.21	16.33	17.65	24.11	14.7	18.26	15.79	11.66	13.31
		Spanish											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	Catalan	2.39	2.44	2.42	19.38	18.66	19.01	22.22	19.06	20.52	14.66	13.36	13.98
	Spanish	2.35	3.41	2.78	20.3	25.69	22.68	26.1	25.64	25.87	16.26	18.24	17.11
	English	27.13	14.74	19.1	33.62	40.33	36.67	29.7	47.87	36.65	30.15	34.31	30.80
	Dutch	4.42	4.47	4.34	18.74	15.77	17.13	30.77	10.81	16	17.97	10.35	11.49
		English											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	Catalan	2.91	5.35	3.77	17.6	25.94	20.97	26.98	19.74	22.62	15.83	17.01	15.78
	Spanish	3.36	7.88	4.71	21.96	28.08	26.64	26.81	34.28	30.09	17.37	23.41	20.48
	English	31.43	17.53	22.51	44.68	41.76	43.17	36.57	47.82	41.44	37.56	35.70	35.70
	Dutch	3.63	3.94	3.78	24.32	19.05	21.36	32.54	14.25	19.82	20.16	12.41	14.98
		Dutch											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	Catalan	2.05	3.28	2.53	10.03	14.95	12.01	9.31	24.22	13.45	7.13	14.15	9.33
	Spanish	2.54	5.61	3.5	9.78	24.73	14.61	10.8	29.12	15.76	7.70	19.82	11.29
	English	22.33	14.13	17.28	22.99	23.2	21.82	10.7	46.66	17.41	18.67	27.99	18.83
	Dutch	3.14	3.73	3.41	13.74	16.47	14.98	18.51	20.63	19.51	11.79	13.61	12.63

Table 4.10: $NECR^{monolingual}_{KL\_BEST}$: coreference resolution results on the SemEval corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) of the MUC, B^3, and CEAF metrics respectively, as well as the unweighted average of precision, recall, and F-measure over the three metrics. The row labels represent the source language, and the column labels represent the target language.

measure on English, and a 21.89 unweighted average F-measure on Dutch. When $NECR^{monolingual}_{MST}$ is trained on the same language it is evaluated on, it ranks third best on Catalan, second best on Spanish, best on English, and second best on Dutch.

4.6.4 Setting 4: Multilingual source training

The named-entity recognition results of the $NECR^{multi\_source}_{KL\_BEST}$ model on the EuroParl$_{parallel}$ corpus are presented in Table 4.13. When the source languages are selected from the EuroParl$_{parallel}$ corpus, the best exact overlap results come from the English-French $NECR^{multi\_source}_{KL\_BEST}$ model for the English, French, and German target languages. The best partial overlap results are returned by the French-German $NECR^{multi\_source}_{KL\_BEST}$ model

		Target					
		Catalan					
		Exact			Partial		
		P	R	F	P	R	F
Source	Catalan	27.98	26.55	27.25	55.56	52.66	54.07
	Spanish	31.21	33.63	32.37	55.76	60.07	57.84
	English	42.61	56.7	48.66	59.44	79.01	67.87
	Dutch	37.63	22.94	28.51	60.46	65.29	62.79
		Spanish					
		Exact			Partial		
		P	R	F	P	R	F
Source	Catalan	28.39	26.12	27.21	56.53	51.46	53.88
	Spanish	31.65	33.43	32.52	55.89	59.03	57.42
	English	43.36	54.63	48.35	61.32	77.26	68.37
	Dutch	38.65	22.72	28.62	61.70	63.43	62.55
		English					
		Exact			Partial		
		P	R	F	P	R	F
Source	Catalan	27.34	29.05	28.17	52.41	56.63	54.44
	Spanish	27.67	33.67	30.37	50.36	61.27	55.28
	English	46.39	57.01	51.15	67.68	83.18	74.63
	Dutch	37.5	23.85	29.15	65.25	62.07	63.62
		Dutch					
		Exact			Partial		
		P	R	F	P	R	F
Source	Catalan	16.04	31.51	21.26	33.36	64.47	43.97
	Spanish	16.88	37.91	23.26	30.48	68.45	42.18
	English	21.97	26.09	23.85	37.94	82.77	52.03
	Dutch	26.94	31.26	28.94	43.52	74.74	55.00

Table 4.11: $NECR_{MST}^{monolingual}$: named-entity recognition results of on the SemEval corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) over exact and partial overlap. The row labels represent the source language, and the column labels represent the target language.

for target English, by the English-French-German $NECR_{KL\_BEST}^{multi\_source}$ model for target French, and by the English-French $NECR_{KL\_BEST}^{multi\_source}$ model for target German. For English and German, the best partial overlap $NECR_{KL\_BEST}^{multi\_source}$ model is trained over source languages different from the target language. The combination of all three source languages performs best only in the partial overlap setting, and only on the French target language.

When the source languages are selected from the SemEval corpus only, the best performing $NECR_{KL\_BEST}^{multi\_source}$ model on named-entity recognition is the Catalan-English $NECR_{KL\_BEST}^{multi\_source}$ model for all the target languages on both exact and partial overlap. The Spanish-English $NECR_{KL\_BEST}^{multi\_source}$ model also gives the best exact overlap results

		Target											
		Catalan											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	Catalan	3.06	3.4	3.22	18.85	19.18	19.01	21.84	20.72	21.26	14.58	14.43	14.49
	Spanish	3.12	4.17	3.57	20.53	24.22	22.22	23.49	25.3	24.36	15.71	17.89	16.71
	English	24.16	16.31	19.48	32.53	43.29	37.14	29.4	51.31	37.4	28.69	36.97	31.34
	Dutch	4.07	3.97	4.02	19.21	16.33	17.65	24.11	14.7	18.26	15.79	11.66	13.31
		Spanish											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	Catalan	2.39	2.44	2.42	19.38	18.66	19.01	21.78	20.04	20.87	14.51	13.71	14.1
	Spanish	2.84	3.78	3.24	22.98	24.27	23.61	26.48	23.38	24.83	17.43	17.14	17.22
	English	26.38	16.51	20.31	38.4	38.03	38.22	29.98	49.68	37.39	31.58	34.74	32.17
	Dutch	4.42	4.27	4.34	18.74	15.77	17.13	23.35	13.73	17.29	15.50	10.70	12.92
		English											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	Catalan	2.91	5.35	3.77	17.6	25.94	20.97	22.26	23.65	22.93	14.25	18.31	15.89
	Spanish	3.6	8.02	4.97	17.2	29.95	21.85	22.47	27.34	26.67	14.42	21.77	17.83
	English	29.28	20.35	24.01	41.25	45.11	43.09	34.68	50.81	41.23	35.07	38.75	36.11
	Dutch	3.63	3.94	3.78	24.32	19.05	21.36	27.08	17.22	21.05	18.34	13.40	15.39
		Dutch											
		MUC			B^3			CEAF			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
Source	Catalan	2.05	3.28	2.53	10.52	20.67	13.94	9.31	24.22	13.45	7.29	16.05	9.97
	Spanish	2.78	5.71	3.74	9.83	18.35	12.8	10.97	28.33	15.82	7.86	17.46	10.78
	English	22.38	17.33	19.53	21.97	26.09	23.85	16.26	35.48	22.3	20.20	26.29	21.89
	Dutch	3.14	3.73	3.41	13.74	16.47	14.9	18.51	20.63	19.51	11.79	13.61	12.60

Table 4.12: $NECR_{MST}^{monolingual}$: coreference resolution results on the SemEval corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) of the MUC, B^3, and CEAF metrics respectively, as well as the unweighted average of precision, recall, and F-measure over the three metrics. The row labels represent the source language, and the column labels represent the target language.

on the French target language. For the French and Dutch target languages, the best performing $NECR_{KL\_BEST}^{multi\_source}$ model is trained over source languages different from the target language.

Source		Target																	
k	Langs	English						French						German					
		Exact			Partial			Exact			Partial			Exact			Partial		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EuroParl$_{parallel}$																			
2	En-Fr	45.94	29.1	35.63	68.10	43.50	52.83	55.1	26.03	35.36	68.61	32.41	43.11	23.23	23.95	23.58	41.75	43.05	42.39
	En-Ge	45.51	24.31	31.69	62.82	33.56	43.75	45.83	22.75	30.41	63.19	31.37	41.93	18.18	20.31	19.53	36.01	38.88	37.39
	Fr-Ge	46.42	27.22	32.18	65.23	46.91	54.58	37.23	12.06	18.22	69.14	22.41	33.85	22.29	25	23.56	37.77	42.36	39.93
3	En-Fr-Ge	39.35	29.1	33.45	63.86	44.17	52.22	42.8	21.55	28.66	64.38	32.41	43.11	19.08	29.68	23.23	34.37	53.47	41.84
SemEval																			
2	Ca-Sp	18.15	32.02	23.17	31.65	55.82	40.39	10.68	42.06	17.03	19.87	78.27	31.70	15.1	33.5	20.81	28.16	62.50	38.83
	Ca-En	23.41	53.08	32.49	38.97	88.35	54.08	17.94	48.44	26.18	30.01	81.03	43.80	19.8	51.56	28.61	34.93	90.97	50.48
	Ca-Du	17.65	36.81	23.86	32.01	66.78	43.28	10.64	42.03	16.9	20.39	78.26	32.38	15.41	36.45	21.67	28.34	67.01	39.83
	Sp-En	17.97	43.15	25.37	33.23	79.79	46.92	17.94	48.44	26.18	23.67	89.65	37.46	15.22	40.45	22.12	33.85	89.93	49.19
	Sp-Du	17.09	41.43	24.19	27.54	66.78	39.00	10.64	41.03	16.9	17.42	82.06	28.74	13.6	38.88	20.16	25.88	73.95	38.34
	En-Du	20.24	44.86	27.9	34.77	77.05	47.92	13.15	47.75	20.62	24.02	87.24	37.36	15.44	38.36	22.12	33.75	83.33	48.04
3	Ca-Sp-En	17.1	33.21	22.58	30.68	59.58	40.51	11.3	44.82	18.05	19.91	78.96	31.80	14.68	34.72	20.63	28.04	66.31	39.42
	Ca-Sp-Du	17.89	36.64	24.04	31.10	63.69	41.79	13.21	50.43	20.93	20.58	78.27	32.59	14.47	35.59	20.58	27.11	66.66	38.55
	Ca-En-Du	17.17	36.81	23.42	31.16	67.80	43.13	10.34	41.37	16.55	19.74	78.96	31.58	14.68	36.8	20.99	28.25	70.83	40.39
	Sp-En-Du	17.99	41.09	25.02	33.13	75.68	46.09	13.21	50.34	20.93	23.61	90	37.41	14.81	39.93	21.61	32.86	88.54	47.93
4	Ca-Sp-En-Du	17.64	36.98	23.89	31.53	66.09	42.69	10.48	43.27	16.87	19.88	82.06	32.01	14.16	36.11	20.35	27.24	69.44	39.13

Table 4.13: $NECR^{multi\_source}_{KL\_BEST}$: named-entity recognition results on the EuroParl$_{parallel}$ corpus, when the multiple source languages are taken from the EuroParl$_{parallel}$ corpus only (see the first table section) and from the SemEval corpus only (see the second table section). Results are reported in terms of precision (P), recall (R), and F-measure (F) over exact and partial overlap. The row labels represent the source language, and the column labels represent the target language. The first column represents k, the number of source languages used in training, and the second column mentions the source language name abbreviations.

Table 4.14 presents the $NECR^{multi\_source}_{KL\_BEST}$ coreference resolution results on the EuroParl$_{parallel}$ corpus. The English-French $NECR^{multi\_source}_{KL\_BEST}$ model reports the best unweighted average F-measure on all three target languages, when the source languages are selected from the EuroParl$_{\text{parallel}}$ corpus only. The Catalan-English $NECR^{multi\_source}_{KL\_BEST}$ model reports the best unweighted average F-measure on all three target languages, when the source languages are selected from the SemEval corpus only. The second best performing model is not consistent across target languages. When the source languages are selected from the SemEval corpus only, the Catalan-English $NECR^{multi\_source}_{KL\_BEST}$ model gives the best results on named-entity recognition and coreference resolution across all target languages.

Source		**Target**								
k	Langs	English			French			German		
		P	R	F	P	R	F	P	R	F
EuroParl										
2	En-Fr	24.74	14.34	18.13	11.43	24.94	15.64	11.35	10.10	10.67
	En-Ge	11.04	6.66	8.28	10.5	7.81	8.42	8.18	5.11	5.96
	Fr-Ge	21.59	11.13	14.32	13.41	3.8	5.89	9.24	7.92	8.32
3	En-Fr-Ge	13.37	8.08	10.02	10.10	6.01	7.45	9.19	5.02	6.44
SemEval										
2	Ca-Sp	11.30	6.15	7.93	2.61	4.31	13.03	11.36	4.51	6.41
	Ca-En	20.19	13.70	15.84	18.92	9.68	12.59	18.18	12.22	14.01
	Ca-Du	14.83	6.92	9.39	12.81	2.69	4.41	12.9	5.21	7.4
	Sp-En	14.64	6.09	8.58	15.44	3.26	5.32	14.74	4.65	6.99
	Sp-Du	18.49	7.08	10.19	18.78	3.12	5.3	13.48	4.53	6.78
	En-Du	16.75	9.33	11.89	14.87	4.11	6.43	14.53	6.25	8.73
3	Ca-Sp-En	15.58	6.66	9.27	15.6	3.44	5.61	13.01	5.09	7.29
	Ca-Sp-Du	14.47	6.76	9.16	14.26	3.18	5.18	12.95	5.09	7.3
	Ca-En-Du	15	6.8	9.31	15.04	3.11	5.12	12.64	4.7	6.83
	Sp-En-Du	15.55	5.93	8.5	19.22	3.7	6.06	14.74	4.65	6.99
4	Ca-Sp-En-Du	15.58	6.66	9.27	16.16	3.06	5.09	12.13	4.24	6.26

Table 4.14: $NECR^{multi\_source}_{KL\_BEST}$: coreference resolution results on the EuroParl$_{parallel}$ corpus, when the multiple source languages are taken from the EuroParl$_{parallel}$ corpus only (see the first table section) and from the SemEval corpus only (see the second table section). Results are reported in terms of the unweighted average of precision (P), recall (R), and F-measure (F) over the MUC, B^3, and CEAF metrics. The row labels represent the source language, and the column labels represent the target language. The first column represents k, the number of source languages used in training, and the second column mentions the source language name abbreviations.

Table 4.15 contains the $NECR^{multi\_source}_{KL\_BEST}$ named-entity recognition results on the EuroParl$_{parallel}$ corpus when the source languages are selected from both the EuroParl$_{parallel}$

and the SemEval corpus. The best exact overlap performance for the English target language is given by the Catalan-English$_{SemEval}$-English$_{EuroParl}$ $NECR^{multi\_source}_{KL\_BEST}$ model, with a 45.84 F-measure. The best partial overlap performance is given by the Dutch-French $NECR^{multi\_source}_{KL\_BEST}$ model, with a 64.70 F-measure. The Catalan-English$_{SemEval}$-English$_{EuroParl}$ $NECR^{multi\_source}_{KL\_BEST}$ model also gives the best exact overlap performance for the French target language (34.82 F-measure) and the German target language (31.55 F-measure). The best partial overlap performance for target French is given by the English$_{SemEval}$-English$_{EuroParl}$ $NECR^{multi\_source}_{KL\_BEST}$ model (66.05 F-measure). The Catalan-English-German $NECR^{multi\_source}_{KL\_BEST}$ model gives the best partial overlap performance on German (54.72 F-measure). In general, the best exact overlap results are given by a system modeled over a combination of two source languages. The best partial overlap results are given by a system modeled over two source languages for the English and French target languages, and by a combination of three source languages for target German.

The $NECR^{multi\_source}_{KL\_BEST}$ coreference resolution results are reported in Table 4.16. The best performing system for target English is the Catalan-English$_{SemEval}$-English$_{EuroParl}$ $NECR^{multi\_source}_{KL\_BEST}$ model, with a 21.68 unweighted average F-measure. For the French target language, the best performing system is the Catalan-English$_{SemEval}$-French $NECR^{multi\_source}_{KL\_BEST}$ model (17.16 unweighted average F-measure). For the German target language the best performing system is the Catalan-English$_{SemEval}$-German $NECR^{multi\_source}_{KL\_BEST}$ model (15.83 unweighted average F-measure). The best performing models are trained over a set of source languages that contains the target language. In general, the number of source languages that give the best model performance is $k = 3$ source languages.

4.7 Discussion

Across all experiment settings, the results on named-entity recognition are substantially larger than the results on coreference resolution, regardless of the corpus on which the experiments are run. For example, in Setting 2, the English named-entity

#	Langs	English						French						German					
		Exact			Partial			Exact			Partial			Exact			Partial		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EuroParl & SemEval																			
2	Ca-En2	22.13	23.8	22.93	41.08	44.18	42.57	16.75	21.72	18.91	30.59	39.66	34.53	15.1	22.91	18.2	35.93	54.51	43.31
	Ca-Fr	21.38	22.77	22.05	39.55	42.12	40.80	19.74	18.79	19.25	33.70	32.07	32.86	14.19	19.61	16.47	36.43	50.35	42.27
	Ca-Ge	21.4	23.4	22.36	38.20	34.93	36.49	21.24	19.48	20.32	36.09	33.10	34.53	16.36	18.57	17.39	34.86	39.58	37.07
	Sp-En2	22.26	19.06	20.53	36.36	42.47	39.18	12.19	22.41	15.79	24.77	45.52	32.08	15.48	22.04	18.19	31.71	45.14	37.25
	Sp-Fr	19.69	19.55	19.62	37.76	38.01	37.88	16.38	19.82	17.94	31.34	37.93	34.32	18.12	18.75	18.43	35.23	36.46	35.84
	Sp-Ge	21.4	18.76	20	36.10	38.70	37.36	15.47	22.41	18.3	27.62	40.00	32.68	17.84	20.13	18.92	33.23	37.50	35.24
	En-En2	45.11	34.76	39.26	71.56	55.14	62.28	32.15	26.72	29.19	63.07	52.41	57.25	19.25	22.39	20.7	38.81	45.14	41.73
	En-Fr	30.33	34.58	32.32	59.16	67.47	63.04	20.08	17.24	18.55	71.49	61.38	66.05	15.14	28.12	19.68	36.64	68.06	47.63
	En-Ge	40.3	32.02	35.68	53.02	42.12	46.95	32.51	16.03	21.47	67.83	33.45	44.80	27.45	24.3	25.78	37.25	32.99	34.99
	Du-En2	34.76	42.64	38.3	69.75	56.85	62.64	38.16	27.24	31.79	62.80	44.83	52.31	16.53	22.56	19.08	45.55	62.15	52.57
	Du-Fr	44.52	45.93	45.21	65.72	63.70	64.70	32.22	15	20.47	69.63	32.41	44.24	21.6	32.63	26	41.84	63.19	50.35
	Du-Ge	15.58	35	21.56	53.08	23.63	32.70	40.85	13.44	20.25	66.32	21.72	32.73	32.72	12.5	18.09	54.55	20.83	30.15
3	Ca-Sp-En2	16.96	22.6	19.38	33.42	44.52	38.18	11.96	27.06	16.59	23.48	53.10	32.56	15.58	23.43	18.72	31.64	47.57	38.00
	Ca-Sp-Fr	19.69	18.6	19.13	37.54	39.73	38.60	13.59	18.96	16.08	27.16	36.90	31.29	17.19	18.92	18.01	33.75	37.15	35.37
	Ca-Sp-Ge	18.3	23.8	21.02	36.59	46.23	40.85	12.61	24.13	16.56	23.78	45.52	31.24	16.31	21.18	18.42	30.91	41.32	35.36
	Ca-En-En2	49.79	42.46	45.84	71.49	60.96	65.80	36.11	33.62	34.82	62.22	57.93	60.00	28.12	35.93	31.55	48.10	61.46	53.96
	Ca-En-Fr	38.9	37.84	38.36	60.92	59.25	60.07	38.69	30.68	34.23	64.35	51.03	56.92	25.22	29.51	27.2	50.74	59.38	54.72
	Ca-En-Ge	52.72	36.47	43.11	65.84	45.55	53.85	37	26.03	30.56	64.71	45.52	53.44	26.06	31.77	28.63	46.45	45.49	45.96
	Ca-Du-En2	23.11	21.15	22.09	35.74	39.04	37.32	13.27	27.41	17.88	26.21	54.14	35.32	17.43	24.82	20.48	33.17	47.22	38.97
	Ca-Du-Fr	19.69	19.69	19.69	41.10	41.10	41.10	14.28	19.31	16.24	29.34	39.66	33.72	20.88	18.92	19.85	36.78	33.33	34.97
	Ca-Du-Ge	21.06	19.04	20	37.46	41.44	39.35	15.14	24.65	18.76	28.39	46.21	35.17	18.48	19.44	18.79	33.12	35.42	34.23
	Sp-En-En2	25.51	20.81	22.92	36.87	45.21	40.62	15.42	37.24	21.81	28.86	69.66	40.81	17.43	26.38	20.99	36.01	54.51	43.37
	Sp-En-Fr	19.86	20.64	20.24	39.15	37.67	38.39	16.57	21.37	18.67	32.35	41.72	36.45	20.55	19.27	19.89	38.89	36.46	37.63
	Sp-En-Ge	21.06	22.44	21.73	37.23	34.93	36.04	16.35	28.96	21.05	28.54	50.00	36.34	19.87	21.52	20.66	36.86	39.93	38.33
	Sp-Du-En2	25.51	21.34	23.24	37.54	44.86	40.87	12.74	31.72	18.18	25.21	62.76	35.97	16.26	25.52	19.86	30.97	48.61	37.84
	Sp-Du-Fr	20.2	19.79	19.99	38.59	39.38	38.98	16.26	21.03	18.34	30.67	39.66	34.59	19.28	19.61	19.44	35.49	36.11	35.80
	Sp-Du-Ge	21.4	18.76	20	36.94	42.12	39.36	14.36	22.44	17.45	26.50	41.03	32.21	17.63	21	19.17	32.94	39.24	35.82
	En-Du-En2	26.71	20.96	23.94	42.74	54.45	47.89	13.97	40.34	20.76	28.08	81.03	41.70	13.96	24.82	17.87	37.50	66.67	48.00
	En-Du-Fr	21.06	27.21	23.74	44.69	34.59	39.00	20.94	18.27	19.52	40.32	35.17	37.57	24.76	18.05	20.88	42.38	30.90	35.74
	En-Du-Ge	22.6	23.82	23.19	40.07	38.01	39.02	16	28.96	20.61	29.90	54.14	38.53	18.82	22.22	20.38	35.59	42.01	38.54
4	Ca-Sp-En-En2	23.11	17.95	20.2	32.98	42.47	37.13	12.7	29.13	17.69	24.96	57.24	34.76	15.94	25.69	19.68	30.82	49.65	38.03
	Ca-Sp-En-Fr	19.34	18.58	18.95	37.50	39.04	38.26	13.68	20.86	16.53	26.92	41.03	32.51	18.28	19.61	18.92	34.63	37.15	35.85
	Ca-Sp-En-Ge	21.74	17.83	19.59	34.55	42.12	37.96	12.18	26.89	16.77	23.28	51.38	32.04	17.08	23.61	19.82	32.41	44.79	37.61
	Ca-Sp-Du-En2	23.11	17.9	20.17	33.69	43.49	37.97	12.37	30.68	17.64	23.78	58.97	33.89	14.87	23.61	18.25	29.98	47.57	36.78
	Ca-Sp-Du-Fr	20.03	17.83	18.87	37.80	42.47	40.00	13.55	21.55	16.64	26.25	41.72	32.22	18.85	19.44	19.14	34.34	35.42	34.87
	Ca-Sp-Du-Ge	22.6	18.53	20.73	36.24	44.18	39.81	12.73	25.51	16.99	24.78	49.66	33.07	17.59	20.83	19.07	32.84	38.89	35.61
	Ca-En-Du-En2	22.77	21.66	22.2	37.13	39.04	38.06	12.86	31.89	18.33	25.31	62.76	36.08	18.69	28.29	22.51	34.63	52.43	41.71
	Ca-En-Du-Fr	20.03	20.74	20.38	41.84	40.41	41.11	15.7	21.55	18.16	31.16	42.76	36.05	22.1	18.57	20.18	38.43	32.29	35.09
	Ca-En-Du-Ge	22.08	19.25	20.57	35.52	40.75	37.96	13.06	23.79	16.87	26.52	48.28	34.23	17.44	19.44	18.30	30.53	34.03	32.18
	Sp-En-Du-En2	25.34	20.44	22.62	37.29	46.23	41.28	15.21	32.58	20.74	28.02	60.00	38.20	16.58	24.65	19.83	34.58	51.39	41.34
	Sp-En-Du-Fr	20.03	21.29	20.59	40.22	38.01	39.08	16.09	20.86	18.16	32.45	42.07	36.64	21.18	19.27	20.18	38.55	35.07	36.73
	Sp-En-Du-Ge	21.23	21.37	21.3	39.66	39.38	39.52	15.21	32.58	20.74	28.16	54.48	37.13	19.4	21.35	20.33	36.91	40.62	38.68
5	Ca-Sp-En-Du-En2	22.77	19.79	21.17	36.90	42.47	39.49	12.11	32.93	17.71	22.72	61.72	33.21	16.91	27.25	20.87	32.11	51.74	39.63
	Ca-Sp-En-Du-Fr	20.71	19.7	20.2	39.41	41.44	40.40	13.91	22.41	17.17	27.41	44.14	33.82	19	19.79	19.38	35.33	36.81	36.05
	Ca-Sp-En-Du-Ge	22.77	17.78	19.96	32.89	42.12	36.94	11.7	30	16.84	22.07	56.55	31.75	16.03	23.61	19.1	29.48	43.40	35.11

Table 4.15: $NECR^{multi\_source}_{KL\_BEST}$: named-entity recognition results on the EuroParl$_{parallel}$ corpus, when the multiple source languages are taken from both the EuroParl$_{parallel}$ corpus and the SemEval corpus. Results are reported in terms of precision (P), recall (R), and F-measure (F) over exact and partial mentions. The row labels represent the source language, and the column labels represent the target language. The first column represents k, the number of source languages used in training, and the second column mentions the source language name abbreviations. Note: The $En2$ language represents the EuroParl$_{parallel}$ version of the English language.

recognition F-measure results range from 43.18 exact overlap F-measure to 59.84 partial overlap F-measure on the EuroParl$_{parallel}$ corpus, and from 52.56 exact overlap F-measure to 75.47 partial overlap F-measure on the SemEval corpus. Meanwhile, the English coreference resolution results range from 19.98 unweighted average F-measure on the EuroParl$_{parallel}$ corpus to 35.70 unweighted average F-measure on the SemEval corpus. The *NECR* system is, in general, better at identifying mentions with partial overlap. It over-generates the mention spans for Romance languages, and under-generates the mention spans for Germanic languages like German.

The *NECR* system performance varies across corpora and languages. For English, the only language common across corpora, I observe better results for both named-entity recognition and coreference resolution on the SemEval corpus. This behavior is explained by the larger size of the SemEval training corpus, compared to the EuroParl$_{\text{parallel}}$ training corpus. The system performance varies across languages: for both the EuroParl$_{\text{parallel}}$ and the SemEval corpus, English is the target language with the best performance, while German is the target language with the lowest performance from the EuroParl$_{\text{parallel}}$ corpus and Dutch is the target language with the lowest performance from the SemEval corpus.

I do not observe large differences in system results when the MST language specific parsers are used, compared to when the $Predictor_{KL\_BEST}$ multilingual parsers are used. On named-entity recognition, the English $NECR^{monolingual}_{\text{MST}}$ model reports a 67.87 F-measure on partial overlap for the English target language, compared to 67.31 partial overlap F-measure obtained by the English $NECR^{monolingual}_{KL\_BEST}$ model. The English $NECR^{monolingual}_{KL\_BEST}$ model obtains better results on exact overlap for the English target language: 49.13 F-measure compared to 48.66 exact overlap F-measure obtained by the English $NECR^{monolingual}_{\text{MST}}$ model. In general, the English $NECR^{monolingual}_{KL\_BEST}$ model obtains as good as or better F-measure results than the English $NECR^{monolingual}_{\text{MST}}$ model on both exact and partial overlap when Catalan, Spanish, and English are used as source languages. These results validate the contribution of my $Predictor_{KL\_BEST}$ parsing model to transferring syntactic information across languages, as it obtains as good as or better results than a parsing model that has access to gold standard annotations.

The experiments conducted in Setting 4 show that, overall, combining source languages from the EuroParl$_{\text{parallel}}$ and SemEval corpora contributes to a performance increase on the named-entity recognition task, compared to the results obtained by the $NECR_{KL\_BEST}^{monolingual}$ model on named-entity recognition. For the English target language, the $NECR_{KL\_BEST}^{multi\_source}$ model obtains a best of 45.84 F-measure on exact overlap and a best of 64.70 F-measure on partial overlap - both results larger than the best results of the $NECR_{KL\_BEST}^{monolingual}$ model. The best performance on German of the $NECR_{KL\_BEST}^{multi\_source}$ model is larger than the best performance of the $NECR_{KL\_BEST}^{monolingual}$ model evaluated on German, on both exact and partial overlap. The best performance of the $NECR_{KL\_BEST}^{multi\_source}$ model does not perform better than the best $NECR_{KL\_BEST}^{monolingual}$ model results on French for exact overlap, but only for partial overlap. The best performing $NECR_{KL\_BEST}^{multi\_source}$ models on named-entity recognition are generated by a combination of two or three source languages, and usually do not include the target language among the set of source languages.

Regarding the performance of the $NECR_{KL\_BEST}^{multi\_source}$ model on the coreference resolution task when the source languages are selected from both corpora, the results show that combining several source languages results in performance results better than results of the $NECR_{KL\_BEST}^{monolingual}$ model results for German only. The best $NECR_{KL\_BEST}^{multi\_source}$ model for German includes the German language among the source languages, together with the Catalan and English languages. Even though German has the lowest performance as a source language alone when tested on itself, in combination with Catalan and English it helps generate the best performance of the $NECR_{KL\_BEST}^{multi\_source}$ model when evaluated on the German target language.

When $NECR_{KL\_BEST}^{multi\_source}$ is built from a combination of SemEval source languages alone, it cannot perform better than the $NECR_{KL\_BEST}^{monolingual}$ model on the English and French target languages. Nevertheless, the mixed combination of SemEval and EuroParl$_{\text{parallel}}$ source languages helps the $NECR_{KL\_BEST}^{multi\_source}$ model perform as well as or better than the $NECR_{KL\_BEST}^{monolingual}$ model on those languages. The $NECR_{KL\_BEST}^{multi\_source}$ model obtains best results that outperform the best results of the $NECR_{KL\_BEST}^{monolingual}$ model on German, for both exact and partial named-entity recognition and for coreference resolution. Similarly, when $NECR_{KL\_BEST}^{multi\_source}$ is built from a combination of EuroParl$_{\text{parallel}}$ source languages alone it outperforms the $NECR_{KL\_BEST}^{monolingual}$ best coreference resolution results on German only.

4.7.1 Comparison to baseline systems

I include a comparison of the *NECR* system to two other baseline systems presented in the literature,[42, 98] on the tasks of named-entity recognition and coreference resolution on the SemEval data (see Table 4.17). I compare the SUCRE[42] and UBIU[98] system results obtained in the closed regular setting of the SemEval-2010 Shared Task to the $NECR^{monolingual}_{KL\_BEST}$ built on the English source language for the Catalan, Spanish, and Dutch target languages, and on the Spanish source language for the English target language. My system does not use the gold standard of the target language during training, and does not use any other linguistic resources other than the multilingual parser constructed from a set of languages different from the target language, and the universal POS information.

On named-entity recognition, my system performs better than the UBIU system on Catalan, Spanish, and Dutch. On Dutch, it manages to perform better than the SUCRE and the UBIU system on named-entity recognition. It does not manage to outperform either the SUCRE or the UBIU system on English. This behavior is explained by the feature set design of the two baseline systems, that is well-tailored to languages like English. On coreference resolution, my system performs better than the UBIU system on Catalan, Spanish, and Dutch.

Without making use of training data for the target language or any other annotated linguistic information, *NECR* manages to perform better than the second best state-of-the-art system, UBIU. *NECR* does not perform better than the SUCRE system on any of the languages, and it performs less than both the SUCRE and UBIU systems on English. On Dutch, a language that was reported as difficult to model during the SemEval-2010 Shared Task, my system manages to show improved performance on both named-entity recognition and coreference resolution.

# Source	Langs	English			French			German		
		P	R	F	P	R	F	P	R	F
EuroParl & SemEval										
2	Ca-En2	9.62	9.29	9.28	5.89	5.2	5.48	7.13	5.70	6.23
	Ca-Fr	8.7	8.42	8.47	5.39	4.83	5.07	4.61	6.1	5.17
	Ca-Ge	10.21	7.71	8.7	6.12	4.94	5.43	5.66	5.53	6.17
	Sp-En2	7.82	7.08	7.27	5.73	3.76	4.51	7.73	5.84	6.56
	Sp-Fr	6.39	6.39	6.23	5.32	4.43	4.79	6.02	6.04	5.83
	Sp-Ge	6.62	6.35	6.36	6.74	5.04	5.70	6.92	6.83	6.69
	En-En2	22.53	13.27	16.74	16.51	11.59	13.53	10.55	8.92	9.38
	En-Fr	19.43	15.6	16.85	12.16	7.19	8.73	10.29	11.22	10.05
	En-Ge	24.54	13.28	16.74	16.00	6.42	9.08	17.56	10.05	12.25
	Du-En2	16.72	11.42	11.89	8.75	6.41	7.39	6.35	5.19	5.7
	Du-Fr	21.41	12.59	15.25	9.95	3.41	4.98	10.79	10.1	9.98
	Du-Ge	10.88	4.68	6.54	9.78	2.57	4.04	9.37	3.29	4.87
3	Ca-Sp-En2	7.69	6.57	6.92	8.15	4.01	5.37	8.33	6.32	7.06
	Ca-Sp-Fr	6.4	6.39	6.18	5	3.73	4.17	6.43	6.32	6.13
	Ca-Sp-Ge	7.98	8.01	7.86	7.09	4.13	5.19	6.97	6.28	6.48
	Ca-En-En2	27.24	18.28	21.68	18.74	15.61	16.99	15.25	14.80	14.83
	Ca-En-Fr	23.61	17.45	19.81	19.97	15.09	17.16	16.37	14.08	14.90
	Ca-En-Ge	26.37	18.07	21.43	17.48	12.40	14.49	15.54	16.29	15.83
	Ca-Du-En2	7.59	7.13	7.27	7.82	4.57	5.73	8.21	6.84	7.39
	Ca-Du-Fr	6.86	6.92	6.74	4.70	3.58	4.04	5.94	6.82	6.28
	Ca-Du-Ge	8.99	7.7	7.9	7.77	5.39	6.33	6.48	6.26	6.23
	Sp-En-En2	7.14	7.39	7.19	9.61	4.22	5.85	8.25	6.59	7.25
	Sp-En-Fr	6.59	7.16	6.88	5.35	4.65	4.95	6.04	7.57	6.57
	Sp-En-Ge	8.05	6.88	7.28	5.81	4.15	4.82	7.95	6.74	7.16
	Sp-Du-En2	10.35	9.20	9.60	11.62	5.08	7.04	9.21	6.43	7.52
	Sp-Du-Fr	7.05	6.98	6.86	5.1	5.09	5.01	6.83	7.03	6.81
	Sp-Du-Ge	8.61	7.41	7.71	8.03	6.19	6.87	7.51	6.69	6.97
	En-Du-En2	8.36	8.57	8.4	11.39	4.54	6.47	7.93	5.04	6.14
	En-Du-Fr	9.07	6.71	7.7	5.38	4.57	4.94	7.06	5.07	5.88
	En-Du-Ge	9.41	7.51	8.32	8.39	4.84	6.13	6.16	6.24	6.16
4	Ca-Sp-En-En2	8.28	7.09	7.48	8.78	4.13	5.6	9.15	6.5	7.52
	Ca-Sp-En-Fr	6.38	6.54	6.21	6.12	4.33	5.01	6.85	7.01	7.29
	Ca-Sp-En-Ge	12.21	10.16	10.61	7.82	4.07	5.32	7.89	6.54	7.06
	Ca-Sp-Du-En2	12.56	9.84	10.64	9.03	4.17	5.75	8.19	5.42	6.48
	Ca-Sp-Du-Fr	6.88	6.37	6.63	5.68	3.97	4.66	6.58	6.85	7.30
	Ca-Sp-Du-Ge	8.34	7.67	7.78	7.89	4.47	5.67	7.28	6.81	6.89
	Ca-En-Du-En2	9.83	9.27	6.91	9.58	4.68	6.24	9.94	7.86	8.68
	Ca-En-Du-Fr	6.88	7.20	6.91	5.98	4.65	5.2	5.9	7.2	6.45
	Ca-En-Du-Ge	9.57	8.46	8.58	6.61	4.11	5.04	6.22	5.99	5.97
	Sp-En-Du-En2	10.06	9.48	9.59	9.64	3.94	5.54	8.84	6.84	7.65
	Sp-En-Du-Fr	7.49	7.01	7.12	4.7	5.46	5.02	6.74	5.89	6.22
	Sp-En-Du-Ge	8.35	7.73	7.89	9.05	5.82	7.02	7.59	7.05	7.19
5	Ca-Sp-En-Du-En2	9.2	8.35	8.55	9.80	3.92	5.6	9.87	6.42	7.76
	Ca-Sp-En-Du-Fr	6.88	7.32	6.86	6.16	4.1	4.9	6.67	6.92	6.64
	Ca-Sp-En-Du-Ge	12.21	10.16	10.61	9.52	3.97	5.59	8.06	5.92	6.75

Table 4.16: $NECR^{multi\_source}_{KL\_BEST}$: coreference resolution results on the EuroParl$_{parallel}$ corpus, when the multiple source languages are taken from both the EuroParl$_{parallel}$ corpus and the SemEval corpus. Results are reported in terms of the unweighted average of precision (P), recall (R), and F-measure (F) over the MUC, B^3, and CEAF metrics. The row labels represent the source language, and the column labels represent the target language. The first column represents k, the number of source languages used in training, and the second column mentions the source language name abbreviations. Note: The $En2$ language represents the EuroParl$_{parallel}$ version of the English language.

System Name	Catalan		English		Spanish		Dutch	
	ne	cr	ne	cr	ne	cr	ne	cr
$NECR_{multi\_parser}$	67.31	30.53	55.84	20.48	69.08	30.80	51.93	18.83
SUCRE	69.7	45.2	80.7	60.76	70.3	48.26	42.3	19.1
UBIU	59.6	29.3	74.2	42.16	60	30.33	34.7	14.1

Table 4.17: Systems comparison results. Note: *ne* represents the F-measure results on named-entity recognition and *cr* represents the unweighted average F-measure result on coreference resolution.

4.8 Conclusions

I present *NECR*, a system for joint learning of named-entities and coreference resolution in a multilingual setting. I show that the *NECR* system benefits from linguistic information gathered from multiple languages. Even though *NECR* does not make use of gold standard annotations on the target language, it performs second best among monolingual supervised state-of-the-art systems for three out of four target languages. The performance of the *NECR* system shows that language modeling can be performed in a multilingual setting even for deep NLP tasks. Due to its design, the *NECR* system can be applied to resource-poor languages for which linguistic information is unavailable or is very sparse.

Chapter 5

Conclusions and Future Work

In this thesis I introduce (1) an NLP system for granular learning of syntactic information from multilingual language sources, (2) a corpus annotated for named entities and coreference resolution, and (3) an NLP system for joint-learning of named entities and coreference resolution in a multilingual setting. The design of these multilingual systems and resources represents a step forward in the development of natural language processing systems for resource-poor languages and furthers the understanding and analysis of linguistic phenomena shared across languages.

By learning syntactic information at the granular level of a sentence, my syntactic parsing system improves over current state-of-the-art multilingual dependency parsing systems. An automated parsing system can more finely identify the syntactic rules common among languages when comparing lower units of language - in this case sentences. This implies that due to the large diversity inherent within a language, modeling multilingual NLP systems at a language level is not sufficient for capturing all the possible similarities between languages. In addition, high-performing dependency parsers can be built on top of source languages from different language families than the target language. I attribute this behavior to both a diversity in treebank annotations across languages and to the degree of diversity inherent in the natural language generation process.

Even with no human annotations available for a resource-poor language, one can build a system for syntactic parsing and coreference resolution with comparable performance to state-of-the-art systems. The systems I present take advantage of underlying syntactic properties shared by languages in order to improve on final system performance on the task of interest. It is worth pointing out that a system built for the coreference resolution task, commonly known as a difficult task to solve in both the monolingual and multilingual setting, manages to perform as well as or better than current state-of-the-art systems when modeled with little syntactic information. This is due to the delexizcalized joint-learning framework that ties together the tasks of named-entity recognition and coreference resolution (similar to how the human brain actually approaches them) and to the comprehensive characterization of language structure done by the model representation through universal linguistic information.

The multilingual corpus I present for named entities and coreference resolution in English, French, and German represents a valuable resource for benchmarking future multilingual systems on their performance across languages. By having a corpus with semantically similar content across languages, one can perform a more informed analysis of system performance. The fact that the annotation guidelines are universally applied across languages guarantees that the same underlying linguistic phenomena are consistently annotated across languages. It also guarantees that NLP systems are not differently penalized during evaluation on account of differences in the annotation guidelines.

5.1 Future work

The contributions presented in this thesis represent advancements to the state of the art in multilingual parsing and coreference resolution. Yet, I envision several directions for future work:

- In this work I was limited by the availability of coreference resolution annotations to four Indo-European languages, and I do not show results of system performance on a wider range of language families. I envision future work to investigate system performance on a larger set of languages and on different language families, as well as on documents from different genres. One first step towards achieving this goal would be to further annotate the EuroParl corpus for coreference resolution on a larger set of languages. Because the EuroParl corpus contains only Indo-European languages, a further step would be to identify resources for creating corpora on non Indo-European languages.

- A common trend in the generation of multilingual annotations is to develop annotations across several natural language processing tasks (e.g., part-of-speech, dependency parsing, coreference resolution). In order to facilitate an informed analysis of the performance of computational systems for each of the linguistic tasks across languages, this analysis should be carried across documents that are semantically equivalent across languages. Thus, future work should invest into generating additional layers of annotations for portion of the EuroParl corpus already annotated for coreference resolution.

- The coreference resolution system presented in this thesis does not thoroughly investigate the cross-lingual modeling of coreference relations. A more general direction for future work is to incorporate explicit modeling of linguistic information shared across languages when solving the coreference resolution task. Specifically, this could be done by using parallel corpora to guide the learning of coreference relations on target languages. Given parallel corpora, one could enforce the model to *(i)* mainly predict coreference relations on mentions equivalent between the target and source languages, *(ii)* predict coreference chains on the target language that maintain similar properties to chains observed in the source languages, in terms of chain length, average distance between mentions involved in a chain, etc.

- I also envision an extension to the current modeling of the coreference resolution hidden state, to better incorporate information available on the mentions stored in the model queue. Specifically, similarity functions could be computed over the queue mention and the current mention predicted by the model. The similarity functions could incorporate morphological, syntactic, or external knowledge. These similarity functions can be either *(i)* language independence or *(ii)* language dependent. They would bring additional information in the coreference resolution model, by biasing coreference relations to take place between mentions that are more similar.

- One deficiency of my multilingual models is that they do not adjust the model parameters to accommodate for the lexicon of the target language. It would be interesting to investigate how the models perform when they are first learned as a multilingual instance and then specialized to the syntactic and semantic structure of the target language, both when annotated information is available and when it is missing. Specializing the models to a specific lexicon could allow for incorporation of contextual cues, as well as external knowledge from resources like Wikipedia, online dictionaries, or large collections of documents.

Appendix A

Annotation Guidelines

A.1 Overview

Rationale: The task of this project is to capture two layers of information about expressions occurring inside a document. The first layer captures expressions as they occur inside a document, based on their type. The second layer, the coreference layer, links together all expressions of a given type that are identical to each other.

Document Structure These guidelines describe the specific type of information that should be annotated for named entity extraction and coreference resolution and provides examples similar to those that may be found in the EuroParl documents. The instances that should be marked along with the examples in the surrounding text that should be included in the annotations are described. Instances in this guideline marked in BLUE are correctly annotated named entities. Instances marked in RED are terms that should not be marked. Coreference pairs will be linked by a connecting line.

Annotation Tool: www.notableapp.com/

A.2 General Guidelines for Named Entities

1. **What things to annotate**

(a) Only complete noun phrases (NPs) and adjective phrases (APs) should be marked. Named entities that fit the described rules, but are only used as modifiers in a noun phrase should not be annotated.

- Media-conscious David Servan-Schreiber was not the first ...
- Deaths were recorded in Europe. Various European capitals ...

2. **How much to annotate**

(a) Include all modifiers with named entities when they appear in the same phrase except for assertion modifiers, i.e., modifiers that change the meaning of an assertion as in the case of negation.

- some of our Dutch colleagues
- Committee on the Environment
- no criminal court

(b) Include up to one prepositional phrase following a named entity. If the prepositional phrase contains a named entity by itself, but it is the first prepositional phrase following a named entity, then it is included as a prepositional phrase and not annotated as a stand-alone named entity.

- President of the council of Ecuador
- President of Ecuador
- members of the latest strike

(c) Include articles and possessives.

- **the** European Union
- **an** executive law
- **his** proposed law

(d) Do not annotate generic pronouns like *we*, *it*, *one* that refer to generic entities.

- **It** must be rainy today.
- **We** must oppose the vote.

3. **Hypothetical mentions** do not annotate hypothetical and vague mentions.

   - A potential candidate to the European Union.

A.3 Categories of Entities

Concepts are defined in three general categories that are each annotated separately: Location, Organization, and Person. Named entities of other entity types should be ignored. In general, an entity is an object in the world like a place or person and a named entity is a phrase that uniquely refers to an object by its proper name ("Hillary Clinton"), acronym ("IBM"), nickname ("Oprah") or abbreviation ("Minn.").

A.3.1 Location

Location entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments. Compound expressions in which place names are separated by a comma are to be tagged as the same instance of Location (see "Kaohsiung, Taiwan", "Washington, D.C."). Also tag "generic" entities like "the renowned city", "an international airport", "the outbound highway".

A.3.2 Organization

Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure. Some examples are businesses ("Bridgestone Sports Co."), stock ticker symbols ("NASDAQ"), multinational organizations ("European Union"), political parties ("GOP") non-generic government entities ("the State Department"), sports teams ("the Yankees"), and military groups (the Tamil Tigers). Also tag "generic" entities like "the government", "the sports team".

A.3.3 Person

Person entities are limited to humans (living, deceased, fictional, deities, ...) identified by name, nickname or alias. Include titles or roles ("Ms.", "President", "coach") and names of family members ("father", "aunt"). Include suffixes that are part of a name (Jr.,Sr. or III). There is no restriction on the length of a title or role (see "Saudi Arabia's Crown Prince Salman bin Abdul Aziz"). Also tag "generic" person expressions like "the patient", "the well-known president".

NOTE: some expressions tend to be ambiguous in the category to which they belong (see "Paris", both the capital of France (Location) and a proper name (Person); "Peugeot", both an organization (Organization) and a proper name (Person)). We ask that you specifically disambiguate those cases, and annotate the expression with the category best defined by the context in which it is used.

A.4 General Guidelines for Coreference Resolution

The general principle for annotating coreference is that two named entities are coreferential if they both refer to an identical expression. Only named entities of the same type can corefer. Named entities should be paired with their nearest preceding coreferent named entity.

NOTE: For ease of annotation, the pronouns in each document have been annotated. If a pronoun is involved in a coreference relation with a named entity annotated in step 1, then a coreference link should be created. See the examples below for when a pronoun should be linked to a named entity.

1. **Bound Anaphors:** Mark a coreference link between a "bound anaphor" and the noun phrase which binds it.

    - Most Politicians prefer their.

- Every institution reported its profits yesterday. They plan to realease full quaterly statements tomorrow.

2. **Apposition:** Typical use of an appositional phrase is to provide an alternative description or name for an object. In written text, appositives are generally set off by commas.

    - Herman Van Rompuy, the well-known president...
    - Herman Van Rompuy, president.
    - Martin Schultz, who was formerly president of the European Union, became president of the European Parliament.

    Mark negated appositions:

    - Ms. Ima Head, never a reliable attendant...

    Also mark if there is only partial overlap between the named entities:

    - The criminals, often legal immigrants...

3. **Predicate Nominals and Time-dependent Identity:** Predicate nominals are typically coreferential with the subject.

    - Bill Clinton was the President of the United States.
    - ARPA program managers are nice people.

Do **NOT** annotate if the text only asserts the possibility of identity:

- Phinneas Flounder may be the dumbest man who ever lived.
- Phinneas Flounder was almost the first president of the corporation.
- If elected, Phinneas Flounder would be the first Californian in the Oval Office.

A.4.1 Coreference Annotation Arbitration

Each batch of documents will be annotated by two independent human annotators. The merged document batches will then will then undergo arbitration by a third annotator.

Bibliography

[1] CoNLL 2012. CoNLL 2012- Modelling unrestricted coreference in OntoNotes. `http://conll.cemantix.org/2012/`. Accessed July 22, 2012.

[2] David J Allerton. *Essentials of grammatical theory: A consensus view of syntax and morphology.* Routledge & Kegan Paul London, 1979.

[3] Abhishek Arun and Frank Keller. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 306–313. Association for Computational Linguistics, 2005.

[4] Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. Coreference resolution in a multilingual information extraction system. In *Proceedings of the First Language Resources and Evaluation Conference (LREC)*, 1998.

[5] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, 1998.

[6] Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics, 2008.

[7] Taylor Berg-Kirkpatrick and Dan Klein. Phylogenetic grammar induction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1288–1297, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[8] Andreea Bodnari. A medication extraction framework for electronic health records. Master's thesis, Massachusetts Institute of Technology, September 2010.

[9] Sabine Buchholz and Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL*, pages 149–164, 2006.

[10] David Burkett and Dan Klein. Two languages are better than one (for syntactic parsing). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 877–886, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[11] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[12] Ming-Wei Chang, Lev Ratinov, and Dan Roth. Guiding semi-supervision with constraint-driven learning. *Urbana*, 51:61801, 2007.

[13] David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, June 2007.

[14] Shay B Cohen, Dipanjan Das, and Noah A Smith. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 50–61. Association for Computational Linguistics, 2011.

[15] Michael John Collins. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 184–191. Association for Computational Linguistics, 1996.

[16] CoNLL. CoNLL: the conference of SIGNLL. `http://ifarm.nl/signll/conll/`. Accessed July 19, 2012.

[17] V.J. Cook. *Chomsky's Universal Grammar: An Introduction*. Applied Language Studies. Blackwell, 1988.

[18] Michael A Covington. *A dependency parser for variable-word-order languages*. Citeseer, 1990.

[19] Michael A. Covington. A fundamental algorithm for dependency parsing. In *In Proceedings of the 39th Annual ACM Southeast Conference*, pages 95–102, 2000.

[20] Brooke Cowan and Michael Collins. Morphology and reranking for the statistical parsing of spanish. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 795–802. Association for Computational Linguistics, 2005.

[21] David Crystal. *Language and the Internet*. Cambridge University Press, 2001.

[22] José Guilherme Camargo de Souza and Constantin Orăsan. Can projected chains in parallel corpora help coreference resolution? In *Anaphora Processing and Applications*, pages 59–69. Springer, 2011.

[23] Amit Dubey and Frank Keller. Probabilistic parsing for german using sister-head dependencies. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 96–103. Association for Computational Linguistics, 2003.

[24] Jenny Rose Finkel, Alex Kleeman, and Christopher D. Manning. Efficient, feature-based, conditional random field parsing. In *In Proceedings ACL/HLT*, 2008.

[25] Ryan Georgi, Fei Xia, and William Lewis. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 385–393. Association for Computational Linguistics, 2010.

[26] Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. *Machine learning*, 29(2-3):245–273, 1997.

[27] Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. Relation extraction and the influence of automatic named-entity recognition. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(1):2, 2007.

[28] Dick Grune and Ceriel Jacobs. Parsing techniques–a practical guide. *VU University. Amsterdam*, 1990.

[29] Aria Haghighi and Dan Klein. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[30] Aria Haghighi and Dan Klein. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, EMNLP '09, pages 1152–1161, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[31] Sanda M Harabagiu, Rzvan C Bunescu, and Steven J Maiorano. Text and knowledge mining for coreference resolution. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

[32] Sanda M Harabagiu and Steven J Maiorano. Multilingual coreference resolution. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 142–149. Association for Computational Linguistics, 2000.

[33] Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. *The World Atlas of Language Structures*. Oxford University Press, 2005.

[34] David G Hays. Dependency theory: A formalism and some observations. *Language*, 40(4):511–525, 1964.

[35] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298, 2005.

[36] Xuedong D Huang, Yasuo Ariki, and Mervyn A Jack. *Hidden Markov models for speech recognition*, volume 2004. Edinburgh University Press, 1990.

[37] John Hutchins. Machine translation: general overview. *The Oxford Handbook of Computational Linguistics*, pages 501–511, 2003.

[38] Zurb Inc. Notable app, April 2014.

[39] Thomas Jackson, Sara Tedmori, Chris Hinde, and Anoud Bani-Hani. The boundaries of natural language processing techniques in extracting knowledge from emails. *Journal of Emerging Technologies in Web Intelligence*, 4(2), 2012.

[40] Alexandre Klementiev and Dan Roth. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 817–824. Association for Computational Linguistics, 2006.

[41] Alexandre Klementiev and Dan Roth. Named entity transliteration and discovery in multilingual corpora. *Learning Machine Translation*, 2008.

[42] Hamidreza Kobdani and Hinrich Schütze. Sucre: A modular system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 92–95. Association for Computational Linguistics, 2010.

[43] Terry Koo and Michael Collins. Efficient third-order dependency parsers. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1–11, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[44] Jonathan K Kummerfeld and Dan Klein. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013.

[45] Sadao Kurohashi and Makoto Nagao. Kn parser: Japanese dependency/case structure analyzer. In *Proceedings of the Workshop on Sharable Natural Language Resources*, pages 48–55, 1994.

[46] Gregory W Lesher, Bryan J Moulton, D Jeffery Higginbotham, et al. Effects of ngram order and training text size on word prediction. In *Proceedings of the RESNA'99 Annual Conference*, pages 52–54, 1999.

[47] Philip M Lewis II and Richard Edwin Stearns. Syntax-directed transduction. *Journal of the ACM (JACM)*, 15(3):465–488, 1968.

[48] Dingcheng Li, Tim Miller, and William Schuler. A pronoun anaphora resolution system based on factorial hidden markov models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1169–1178. Association for Computational Linguistics, 2011.

[49] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 25–32, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[50] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330, 1993.

[51] Ryan Mcdonald. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, 2007.

[52] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. *Proceedings of ACL, Sofia, Bulgaria*, 2013.

[53] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 523–530, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[54] Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–72. Association for Computational Linguistics, 2011.

[55] Paul McNamee and James Mayfield. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1-2):73–97, 2004.

[56] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, pages 128–144, 2008.

[57] Ruslan Mitkov. Multilingual anaphora resolution. *Machine Translation*, 14(3-4):281–299, 1999.

[58] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[59] Tahira Naseem, Regina Barzilay, and Amir Globerson. Selective sharing for multilingual dependency parsing. In *Proceedings of the 2012 ACL*. Association for Computational Linguistics, 2013.

[60] Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010*

Conference on Empirical Methods in Natural Language Processing, pages 1234–1244. Association for Computational Linguistics, 2010.

[61] Tahira Naseem, Chen Harr, Regina Barzilay, and Mark Johnson. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of EMNLP*, pages 1234–1244. Association for Computational Linguistics, 2010.

[62] Roberto Navigli, David Jurgens, and Daniele Vannella. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (* SEM 2013)*, pages 222–231, 2013.

[63] Joakim Nivre. Dependency grammar and dependency parsing. *MSI report*, 5133(1959):1–32, 2005.

[64] Joakim Nivre, Johan Hall, Sandra Kubler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, 2007.

[65] Joakim Nivre and Jens Nilsson. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 99–106, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[66] Franz Josef Och and Hermann Ney. Giza++: Training of statistical translation models, 2000.

[67] Kemal Oflazer. Dependency parsing with an extended finite-state approach. *Computational Linguistics*, 29(4):515–544, 2003.

[68] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.

[69] Amy E Pierce. *Language acquisition and syntactic theory*. Springer, 1992.

[70] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 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, 2012.

[71] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics, 2010.

[72] Altaf Rahman and Vincent Ng. 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*, NAACL HLT '12, pages 720–730, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[73] Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[74] Alexander E Richman and Patrick Schone. Mining wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1–9, 2008.

[75] R. H. Robins. *A short history of Linguistic*. Longman, 1967.

[76] Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. McGraw-Hill New York, 1983.

[77] SemEval-2014 Task 10. SemEval-2014 Task 10, April 2014.

[78] Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. Synergy: A named entity recognition system for resource-scarce languages such as swahili using online machine translation. In *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26, 2010.

[79] Benjamin Snyder, Tahira Naseem, and Regina Barzilay. Unsupervised multilingual grammar induction. 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*, ACL '09, pages 73–81, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[80] Anders Søgaard. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 682–686, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[81] Oliver Streiter, Kevin P Scannell, and Mathias Stuflesser. Implementing nlp projects for noncentral languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4):267–289, 2006.

[82] Oscar Täckström, Ryan McDonald, and Joakim Nivre. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 NAACL HLT*. Association for Computational Linguistics, 2013.

[83] Pasi Tapanainen and Timo Järvinen. A non-projective dependency parser. In *In Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71, 1997.

[84] Jörg Tiedemann. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (vol V)*, pages 237–248. John Benjamins, Amsterdam/Philadelphia, 2009.

[85] Michal Toman, Roman Tesar, and Karel Jezek. Influence of word normalization on text classification. *Proceedings of InSciT*, pages 354–358, 2006.

[86] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[87] Ö. Uzuner, B.R. South, S. Shen, and S.L. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[88] Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett South. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of American Medical Informatics Association*, 17:514–518, February 2010.

[89] B Vauquois. A survey of formal grammars and algorithms for recognition and transformation in machine translation', ifip congress-68, edinburgh, 254-260; reprinted in ch. *Bernard Vauquois et la TAO: Vingt-cinq Ans de Traduction Automatique-Analectes*, pages 201–213, 1968.

[90] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A Model-theoretic Coreference Scoring Scheme. In *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pages 45–52, Morristown, NJ, USA, 1995. Association for Computational Linguistics.

[91] Atro Voutilainen. Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232, 2003.

[92] David Weiss and Ben Taskar. Structured Prediction Cascades. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[93] Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403, 1997.

[94] Hiroyasu Yamada and Yuji Matsumoto. Statistical dependency analysis with support vector machines. In *In Proceedings of IWPT*, pages 195–206, 2003.

[95] David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, HLT '01, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[96] Hao Zhang and Ryan McDonald. Generalized higher-order dependency parsing with cube pruning. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 320–331, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[97] Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. Learning to map into a universal pos tagset. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1368–1378. Association for Computational Linguistics, 2012.

[98] Desislava Zhekova and Sandra Kübler. Ubiu: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99. Association for Computational Linguistics, 2010.