Second Language Learning from a Multilingual Perspective

by

Yevgeni Berzak

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

February 2018

Author . . . . . .
Department of Electrical Engineering and Computer Science
December 28, 2017

Certified by . . . . . .
Boris Katz
Principal Research Scientist
Computer Science & Artificial Intelligence Lab
Thesis Supervisor

Accepted by . . . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Second Language Learning from a Multilingual Perspective

by

Yevgeni Berzak

Submitted to the Department of Electrical Engineering and Computer Science
on December 28, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

How do people learn a second language? In this thesis, we study this question through an examination of cross-linguistic transfer: the role of a speaker's native language in the acquisition, representation, usage and processing of a second language. We present a computational framework that enables studying transfer in a unified fashion across language production and language comprehension. Our framework supports bidirectional inference between linguistic characteristics of speakers' native languages, and the way they use and process a new language. We leverage this inference ability to demonstrate the systematic nature of cross-linguistic transfer, and to uncover some of its key linguistic and cognitive manifestations. We instantiate our framework in language production by relating syntactic usage patterns and grammatical errors in English as a Second Language (ESL) to typological properties of the native language, showing its utility for automated typology learning and prediction of second language grammatical errors. We then introduce eye tracking during reading as a methodology for studying cross-linguistic transfer in second language comprehension. Using this methodology, we demonstrate that learners' native language can be predicted from their eye movement while reading free-form second language text. Further, we show that language processing during second language comprehension is intimately related to linguistic characteristics of the reader's first language. Finally, we introduce the Treebank of Learner English (TLE), the first syntactically annotated corpus of learner English. The TLE is annotated with Universal Dependencies (UD), a framework geared towards multilingual language analysis, and will support linguistic and computational research on learner language. Taken together, our results highlight the importance of multilingual approaches to the scientific study of second language acquisition, and to Natural Language Processing (NLP) applications for non-native language.

Thesis Supervisor: Boris Katz
Title: Principal Research Scientist
Computer Science & Artificial Intelligence Lab

Acknowledgments

This thesis would not have been possible without the help of many wonderful people. First and foremost, I would like to thank my advisor Boris Katz for his support, guidance, enthusiasm, kindness, and for encouraging me to pursue the ideas that I was most passionate about. I would also like extend my deepest gratitude to my thesis committee, Anna Korhonen and Peter Szolovits for their valuable advice on this work.

I have been very fortunate to work with and learn from fantastic collaborators, Andrei Barbu, Suzanne Flynn, Danny Harari, Yan Huang, Anna Korhonen, Chie Nakamura, Roi Reichart, Carrie Spadine and Shimon Ullman. I am also very grateful to have had the opportunity to mentor and collaborate with extremely talented undergraduate students at MIT, Run Chen, Sebastian Garza, Emily Kellison-Linn, Jessica Kenney, Lucia Lam, Sophie Mori, Amelia Smith, Jing Wang and Emily Weng.

Special thanks to my labmates, colleagues and friends at MIT for all the feedback on this work and for all the support during my grad school years, Andreea Gane, Candace Ross, Charles Gruenwald, Gabi Melo, Gabriel Zaccak, Guy Ben Yosef, Jason Tong, Jon Malmaud, Karthik Narasimhan, Leo Rosenstein, Matt Sniatinsky, Michelle Spektor, Mirella Lapata, Nate Kushman, Richard Futrell, Roger Levy, S.R.K. Branavan, Sue Felshin, Tao Lei, Tomer Ullman, Yen-Ling Kuo, Yonatan Belinkov, Yoong Keok Lee and Yuan Zhang.

Before coming to MIT, I was lucky to have had the chance to work with Michael Fink and Maya Bar-Hillel at the Hebrew University, and Caroline Sporleder at Saarland University. Their influence on my academic path was transformative.

Finally, I would like to thank my parents Galia Berzak and Alik Berzak, and my brother Leon Berzak for being there for me.

Dedication

This thesis is dedicated to my grandfather Valeri Somorov, and late grandmother Alla Somorov who have been a constant source of inspiration in my life.

Bibliographic Note

This thesis is based on prior work that appeared in peer-reviewed publications. Chapter 2 on reconstruction of typology from ESL text is based on material that was published in Berzak et al. [10]. Chapter 3 on prediction of grammatical errors from typology is based on Berzak et al. [11]. Chapter 4 on inferring first language from gaze is based on Berzak et al. [9]. Chapter 5 on the Treebank of Learner English (TLE) is based on Berzak et al. [8]. Appendix A on methodology of syntactic annotation is based on Berzak et al. [7].

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Today, the majority of the English speakers in the world learn and use English as a Second Language (ESL)[1][21]. This prevalence of learner English is often overlooked in linguistics and psychology, which have traditionally focused on the study of first language acquisition. Similarly, Natural Language Processing (NLP) tools are typically calibrated towards native speakers, making them of limited use for learner language. As the number of ESL learners continues to grow, learner language becomes of paramount importance to the scientific study of language, as well as to NLP applications which will need to address the unique characteristics of non-native language and support language learning. Increasing our understanding of the ways in which adult speakers acquire, represent and process new languages will be crucial for the success of these endeavors.

In this thesis we investigate multilingualism through the prism of *cross-linguistic transfer*. The study of cross-linguistic transfer (also known as cross-linguistic influence and language interference) seeks to establish the role of a speaker's first language in learning and using a second language. This topic is essential to understanding multilingualism as it is tightly connected to the processes by which speakers acquire new languages. It is also pivotal for understanding the extent to which cognitive resources are shared in the mental

[1] https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

representation of multiple languages. From this perspective, transfer phenomena can be used not only for studying second language acquisition, but also for probing pertinent aspects of first language knowledge. Finally, understanding transfer can improve modeling of language usage and language processing, making this topic highly relevant for both psycholinguistic research as well as practical NLP tools which seek to process learner language in a sound and robust manner.

Cross-linguistic transfer has been previously studied in Psychology, Second Language Acquisition (SLA) and Linguistics. Much of this research used *comparative* approaches, focusing on case study analyses of native language influence in second language performance [75, 36, 64, 44, 2]. These studies provided important conceptual frameworks for understanding transfer, most notably the Contrastive Analysis (CA) approach, which seeks to understand second language usage in light of differences between the second language and the native language of the speaker. However, work in these fields were often limited in the range of examined linguistic phenomena, used small sets of languages and rarely addressed the transfer problem from a computational perspective.

Computational approaches to cross-linguistic transfer were introduced more recently in NLP, where this topic was mainly studied using *detection* based approaches [43]. These approaches revolve around the task of Native Language Identification (NLI), which requires to predict the native language of a non-native writer. While high performance on the NLI task [101], and subsequent feature analyses [105, 13, 103, 97, 98, 59, 14] confirmed the ability to extract native language signal from second language text, its linguistic characteristics as well as its scope are far from being established.

Our work seeks to advance the scientific study of multilingualism, by addressing a set of interrelated open questions which unfold from the topic of cross-linguistic transfer: Do we acquire new languages independently from our existing linguistic knowledge or only in reference to it? If the answer is the latter, which parts of this knowledge are used, and how? Does transfer affect both language production and language comprehension? We leverage insights from investigating these questions to deepen our understanding of fundamental computational, cognitive and linguistic principles of multilingualism, and connect these principles to development of linguistic resources and NLP applications for learner

language.

1.2 This Thesis

This thesis presents several key results which advance the state of the art in the study of cross-linguistic transfer, and improve our understanding of multilingualism. First, in chapters 2 and 3 we provide a novel conceptual and empirical framework for determining first language influence in foreign language production. The core idea of this framework is connecting native language specific production patterns in ESL to typological properties of native languages. Using this framework, we demonstrate that linguistic characteristics of the first language *systematically* affect syntactic usage and grammatical errors in second language usage.

We further show that cross-linguistic transfer can be leveraged for two fundamental NLP tasks, prediction of typology in the first language, and prediction of grammatical errors in the second language. In chapted 2 we address typology prediction, demonstrating that traces of cross-linguistic transfer can be used to recover native language typological similarity structure directly from ESL text. We use this ESL based similarity structure to perform prediction of typological features in an unsupervised fashion with respect to the target languages. This approach enables predicting the full typological profile of a language without any prior knowledge about its characteristics.

Reversing the direction of this analysis, in chapter 3 we demonstrate that language specific error distributions in ESL writing differ across native languages, and can be predicted from the typological properties of the native language and their relation to the typology of English. Our typology driven model enables obtaining accurate estimates of such distributions without access to any ESL data for the target languages. Furthermore, we show that our framework is instrumental for linguistic inquiry seeking to identify first language factors that contribute to a wide range of difficulties in second language acquisition. We also show that typology learning and grammatical error prediction can be combined in a bootrstrapping strategy, resulting in a system that is able to predict both typological features in the first language and grammatical errors in a second language using only unannonated

ESL texts as its input.

In chapter 4 we introduce eyetracking while reading free-form text in a foreign language as a general methodology for studying first language influence in second language comprehension. Using this methodology, we show for the first time that native language can be predicted from the reader's gaze patterns in a second language. We provide analysis of classifier uncertainty and learned features, which indicates that differences in English reading are likely to be rooted in linguistic characteristics of native languages. Our analysis of language comprehension stands in direct parallel to the relation between second language performance and first language typology in language production, established in chapters 2 and 3. This parallelism offers new ground for advancing both empirical and theoretical research of multilingualism.

Finally, in chapter 5 we introduce the Treebank of Learner English (TLE) - the first publicly available treebank of learner language. Based on our investigation of cross-linguistic transfer, we argue for the multilingual syntactic formalism Universal Dependencies (UD) as a suitable representation of learner language syntax, and delineate ESL annotation guidelines that support consistent syntactic treatment of ungrammatical English. A first of its kind resource, the TLE provides manually annotated POS tags and syntactic trees for over 5,000 learner sentences. The UD annotations are tied to annotations of grammatical errors, whereby full syntactic analyses are provided for both the original and error corrected versions of each sentence. Further on, we benchmark POS tagging and dependency parsing performance on the TLE dataset and measure the effect of grammatical errors on parsing accuracy. We envision the treebank to support a wide range of linguistic and computational research on learner syntax, linguistic typology, and automatic processing of ungrammatical language.

In Appendix A we describe a study on methodology and characteristics of human syntactic annotations, which branched out of our syntactic annotations project presented in chapter 5. In this study we address two key characteristics of human syntactic annotations: anchoring and agreement. Anchoring is a well known cognitive bias in human decision making, where judgments are drawn towards preexisting values. We study the influence of anchoring on a standard approach to creation of syntactic resources where syntactic an-

notations are obtained via human editing of tagger and parser output. Our experiments demonstrate a clear anchoring effect and reveal unwanted consequences, including overestimation of parsing performance and lower quality of annotations in comparison with human based annotations. Using sentences from the Penn Treebank WSJ, we also report systematically obtained inter-annotator agreement estimates for English dependency parsing. Our agreement results control for parser bias, and are consequential in that they are on par with state of the art parsing performance for English newswire. We discuss the impact of our findings on strategies for future annotation efforts and parser evaluations

1.3 Contributions

This thesis presents a unified framework for studying multilingualism, focusing on identification of *linguistic regularities* in learner language. To uncover such regularities, we introduce three broad conceptual and methodological contributions. First, we present a novel approach to the study of multilingualism and cross linguistic transfer which combines *linguistics, NLP and psycholinguistics* into one coherent framework. Linguistic theory introduces fundamental research frameworks and linguistic categories which are instrumental for distilling the core questions to be addressed. Linguistic research also provides valuable resources for linguistic documentation of languages of the world. NLP utilizes linguistic corpora and provides computational machinery for automated processing of textual data. Finally, psycholinguistics enables broad coverage analysis of behavioral traces of linguistic performance and online processing by language learners. We combine theory, data and methodologies from these three domains into a computational framework which translates core questions on the nature of multilingualism into empirical data-driven prediction tasks. Our framework enables gaining insight into the nature of these tasks by analyzing model behavior and learned features. Our integrative approach provides powerful means for understanding how humans acquire and process multiple languages, and has thus far been underexplored, in part, due to disciplinary boundaries.

The second broad contribution of this thesis is connecting *language learning* to the *linguistic characterization of languages of the world* via language usage and behavioral data.

While such performance traces have been used in prior work for studying both language learning and linguistic typology, these topics were traditionally examined in separation. Our approach establishes an explicit link between seemingly unrelated objects of study, interference signal during second language processing and the documentation of language typology. Connecting cognitive aspects of second language processing with first language typology paves the way for novel methodologies for inducing typology from cognitive state, and well as utilization of typological information for analyzing language learning.

Our third contribution is the unification of *language production* and *language comprehension* in one framework for studying cross-linguistic transfer in multilingualism. We establish a common ground consisting of textual materials, NLP tools and machine learning models, which enables bridging between textual production and gaze patterns during second language reading. Our approach supports direct comparisons of results across production and comprehension and their integration within a broader theory of language learning. This framework deepens our fundamental understanding of human language processing and both improves and broadens the reach of NLP technology.

Chapter 2

From ESL Production to Native Language Typology

2.1 Introduction

In this chapter, we examine the hypothesis that cross-linguistic structure transfer is governed by the *typological properties of the native language*. We provide empirical evidence for this hypothesis by showing that language similarities derived from structural patterns of ESL usage are strongly correlated with similarities obtained directly from the typological features of the native languages. This correlation suggests that the structure of a speaker's first language systematically affects structural usage in a second language.

Our finding has broad implications on the ability to perform inference from native language structure to second language performance and vice versa. In particular, it paves the way for a novel and powerful framework for *comparing native languages through second language performance*. This framework overcomes many of the inherent difficulties of direct comparison between languages, and the lack of sufficient typological documentation for the vast majority of the world's languages.

Further on, we utilize this transfer enabled framework for the task of reconstructing typological features. Automated prediction of language typology is extremely valuable for both linguistic studies and NLP applications which rely on such information (e.g. [71, 99],

see [76] for an overview on the use of typological information in NLP). Furthermore, this task provides an objective external testbed for the quality of our native language similarity estimates derived from ESL texts.

Treating native language similarities obtained from ESL as an approximation for typological similarities, we use them to predict typological features without relying on typological annotation for the target languages. Our ESL based method yields 71.4% – 72.2% accuracy on the typology reconstruction task, as compared to 69.1% – 74.2% achieved by typology based methods which depend on pre-existing typological resources for the target languages.

To summarize, this chapter offers two main contributions. First, we provide an empirical result that validates the systematic existence of linguistic transfer, tying the typological characteristics of the native language with the structural patterns of foreign language usage. Secondly, we show that ESL based similarities can be directly used for prediction of native language typology. As opposed to previous approaches, our method achieves strong results without access to any a-priori knowledge about the target language typology.

2.2 Data

Cambridge FCE

We obtain ESL essays from the Cambridge First Certificate in English (FCE) learner corpus [113], a publicly available subset of the Cambridge Learner Corpus (CLC)[1]. The corpus contains upper-intermediate level essays by native speakers of 16 languages[2]. Discarding Swedish and Dutch, which have only 16 documents combined, we take into consideration the remaining following 14 languages, with the corresponding number of documents in parenthesis: Catalan (64), Chinese (66), French (146), German (69), Greek (74), Italian (76), Japanese (82), Korean (86), Polish (76), Portuguese (68), Russian (83), Spanish (200), Thai (63) and Turkish (75). The resulting dataset contains 1228 documents with an average

[1] `http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603`

[2] We plan to extend our analysis to additional proficiency levels and languages when error annotated data for these learner profiles will be publicly available.

of 379 words per document.

World Atlas of Language Structures

We collect typological information for the FCE native languages from WALS. The database contains information about 2,679 of the world's 7,105 documented living languages [54]. The typological feature list has 188 features, 175 of which are present in our dataset. The features are associated with 9 linguistic categories: Phonology, Morphology, Nominal Categories, Nominal Syntax, Verbal Categories, Word Order, Simple Clauses, Complex Sentences and Lexicon. Table 2.1 presents several examples for WALS features and their range of values.

ID	Type	Feature Name	Values
26A	Morphology	Prefixing vs. Suffixing in Inflectional Morphology	Little affixation, Strongly suffixing, Weakly suffixing, Equal prefixing and suffixing, Weakly prefixing, Strong prefixing.
30A	Nominal Categories	Number of Genders	None, Two, Three, Four, Five or more.
83A	Word Order	Order of Object and Verb	OV, VO, No dominant order.
111A	Simple Clauses	Non-periphrastic Causative Constructions	Neither, Morphological but no compound, Compound but no morphological, Both.

Table 2.1: Examples of WALS features. As illustrated in the table examples, WALS features can take different types of values and may be challenging to encode.

One of the challenging characteristics of WALS is its low coverage, stemming from lack of available linguistic documentation. It was previously estimated that about 84% of the language-feature pairs in WALS are unknown [22, 38]. Despite the prevalence of this issue, it is important to bear in mind that some features do not apply to all languages by definition. For instance, feature 81B *Languages with two Dominant Orders of Subject, Object, and Verb* is relevant only to 189 languages (and has documented values for 67 of them).

We perform basic preprocessing, discarding 5 features that have values for only one language. Further on, we omit 19 features belonging to the category Phonology as comparable phonological features are challenging to extract from the ESL textual data. After

this filtering, we have 151 features, 114.1 features with a known value per language, 10.6 languages with a known value per feature and 2.5 distinct values per feature.

Following previous work, we binarize all the WALS features, expressing each feature in terms of k binary features, where k is the number of values the original feature can take. Note that beyond the well known issues with feature binarization, this strategy is not optimal for some of the features. For example, the feature 111A *Non-periphrastic Causative Constructions* whose possible values are presented in table 2.1 would have been better encoded with two binary features rather than four. The question of optimal encoding for the WALS feature set requires expert analysis and is left for future research.

2.3 Inferring Language Similarities from ESL

Our first goal is to derive a notion of similarity between languages with respect to their native speakers' distinctive structural usage patterns of ESL. A simple way to obtain such similarities is to train a probabilistic NLI model on ESL texts, and interpret the uncertainty of this classifier in distinguishing between a pair of native languages as a measure of their similarity.

NLI Model

The log-linear NLI model is defined as follows:

$$p(y|x;\theta) = \frac{\exp(\theta \cdot f(x,y))}{\sum_{y' \in Y} \exp(\theta \cdot f(x,y'))} \tag{2.1}$$

where y is the native language, x is the observed English document and θ are the model parameters. The parameters are learned by maximizing the L2 regularized log-likelihood of the training data $D = \{(x_1, y_1), ..., (x_n, y_n)\}$.

$$L(\theta) = \sum_{i=1}^{n} \log p(y_i|x_i;\theta) - \lambda\|\theta\|^2 \tag{2.2}$$

The model is trained using gradient ascent with L-BFGS-B [15]. We use 70% of the FCE data for training and the remaining 30% for development and testing of the classifier.

Features

As our objective is to relate native language and target language *structures*, we seek to control for biases related to the content of the essays. Such biases may arise from the essay prompts as well as from various cultural factors. For example, our experiments with a typical NLI feature set show that the strongest features for predicting Chinese are strings such as *China* and *in China*. Similar features dominate the weights of other languages as well. Such content features boost classification performance, but are hardly relevant for modeling linguistic phenomena. We therefore define the model using only *unlexicalized* morpho-syntactic features, which capture structural properties of English usage.

Feature Type	**Examples**
Unlexicalized labeled dependencies	Relation = *prep* Head = *VBN* Dependent = *IN*
Ordering of head and dependent	Ordering = *right* Head = *NNS* Dependent = *JJ*
Distance between head and dependent	Distance = 2 Head = *VBG* Dependent = *PRP*
POS sequence between head and dependent	Relation = *det* POS-between = *JJ*
POS n-grams (up to 4-grams)	POS bigram = *NN VBZ*
Inflectional morphology	Suffix = *ing*
Derivational morphology	Suffix = *ity*

Table 2.2: Examples of syntactic and morphological features of the NLI model. The feature values are set to the number of occurrences of the feature in the document.

Our syntactic features are derived from the output of the Stanford parser. A comprehensive description of the Stanford parser dependency annotation scheme can be found in the Stanford dependencies manual [25]. The feature set, summarized in table 2.2, contains features which are strongly related to many of the structural features in WALS. In particular, we use features derived from labeled dependency parses. These features encode properties such as the types of dependency relations, ordering and distance between the head and the dependent. Additional syntactic information is obtained using POS n-grams. Finally, we consider derivational and inflectional morphological affixation. The annotations required for our syntactic features are obtained from the Stanford POS tagger [104] and the Stanford

parser [24]. The morphological features are extracted heuristically.

ESL Based Native Language Similarity Estimates

Given a document x and its author's native language y, the conditional probability $p(y'|x;\theta)$ can be viewed as a measure of confusion between languages y and y', arising from their similarity with respect to the document features. Under this interpretation, we derive a language similarity matrix S'_{ESL} whose entries are obtained by averaging these conditional probabilities on the training set documents with the true label y, which we denote as $D_y = \{(x_i, y) \in D\}$.

$$S'_{ESL_{y,y'}} = \begin{cases} \frac{1}{|D_y|} \sum\limits_{(x,y)\in D_y} p(y'|x;\theta) & \text{if } y' \neq y \\ 1 & \text{otherwise} \end{cases} \tag{2.3}$$

For each pair of languages y and y', the matrix S'_{ESL} contains an entry $S'_{ESL_{y,y'}}$ which captures the average probability of mistaking y for y', and an entry $S'_{ESL_{y',y}}$, which represents the opposite confusion. We average the two confusion scores to receive the matrix of pairwise language similarity estimates S_{ESL}.

$$S_{ESL_{y,y'}} = S_{ESL_{y',y}} = \frac{1}{2}(S'_{ESL_{y,y'}} + S'_{ESL_{y',y}}) \tag{2.4}$$

Note that comparable similarity estimates can be obtained from the confusion matrix of the classifier, which records the number of misclassifications corresponding to each pair of class labels. The advantage of our probabilistic setup over this method is its robustness with respect to the actual classification performance of the model.

Language Similarity Tree

A particularly informative way of representing language similarities is in the form of hierarchical trees. This representation is easier to inspect than a similarity matrix, and as such, it can be more instrumental in supporting linguistic inquiry on language relatedness. Additionally, as we show in section 2.6, hierarchical similarity trees can outperform raw

similarities when used for typology reconstruction.

We perform hierarchical clustering using the Ward algorithm [107]. Ward is a bottom-up clustering algorithm. Starting with a separate cluster for each language, it successively merges clusters and returns the tree of cluster merges. The objective of the Ward algorithm is to minimize the total within-cluster variance. To this end, at each step it merges the cluster pair that yields the minimum increase in the overall within-cluster variance. The initial distance matrix required for the clustering algorithm is defined as $1 - S_{ESL}$. We use the Scipy implementation[3] of Ward, in which the distance between a newly formed cluster $a \cup b$ and another cluster c is computed with the Lance-Williams distance update formula [51].

2.4 WALS Based Language Similarities

In order to determine the extent to which ESL based language similarities reflect the typological similarity between the native languages, we compare them to similarities obtained directly from the typological features in WALS.

The WALS based similarity estimates between languages y and y' are computed by measuring the cosine similarity between the binarized typological feature vectors.

$$S_{WALS_{y,y'}} = \frac{v_y \cdot v_{y'}}{\|v_y\|\|v_{y'}\|} \tag{2.5}$$

As mentioned in section 2.2, many of the WALS features are either not defined or do not have values for all the FCE languages. To address this issue, we experiment with two different strategies for choosing the WALS features to be used for language similarity computations. The first approach, called *shared-all*, takes into account only the 32 features that have defined and known values in all the 14 languages of our dataset. In the second approach, called *shared-pairwise*, the similarity estimate for a pair of languages is determined based on the features shared between these two languages.

As in the ESL setup, we use the two matrices of similarity estimates to construct WALS

[3] `http://docs.scipy.org/.../scipy.cluster.hierarchy.linkage.html`

based hierarchical similarity trees. Analogously to the ESL case, a WALS based tree is generated by the Ward algorithm with the input distance matrix $1 - S_{WALS}$.

2.5 Comparison Results

After independently deriving native language similarity matrices from ESL texts and from typological features in WALS, we compare them to one another. Figure 2-1 presents a scatter plot of the language similarities obtained using ESL data, against the equivalent WALS based similarities. The scores are strongly correlated, with a Pearson Correlation Coefficient of 0.59 using the *shared-pairwise* WALS distances and 0.50 using the *shared-all* WALS distances.

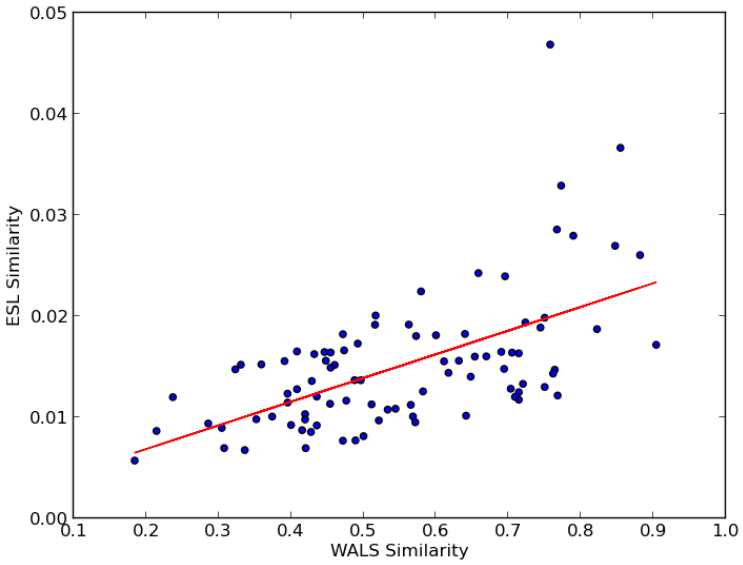


Figure 2-1: *shared-pairwise* WALS based versus ESL based language similarity scores. Each point represents a language pair, with the vertical axis corresponding to the ESL based similarity and the horizontal axis standing for the typological *shared-pairwise* WALS based similarity. The scores correlate strongly with a Pearson's coefficient of 0.59 for the *shared-pairwise* construction and 0.50 for the *shared-all* feature-set.

This correlation provides appealing evidence for the hypothesis that distinctive structural patterns of English usage arise via cross-linguistic transfer, and to a large extent reflect the typological similarities between the respective native languages. The practical conse-

quence of this result is the ability to use one of these similarity structures to approximate the other. Here, we use the ESL based similarities as a proxy for the typological similarities between languages, allowing us to reconstruct typological information without relying on a-priori knowledge about the target language typology.

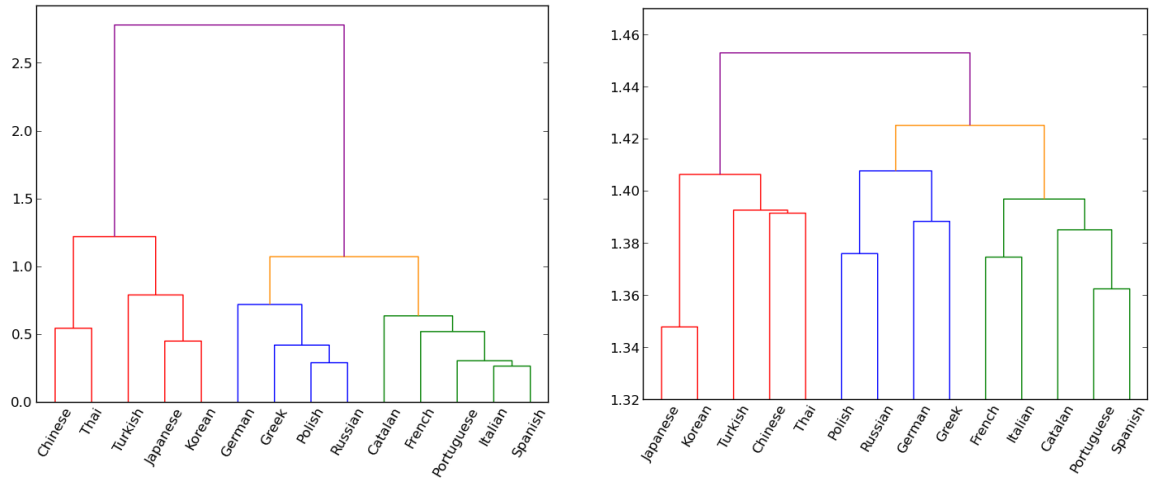
In figure 2-2 we present, for illustration purposes, the hierarchical similarity trees obtained with the Ward algorithm based on WALS and ESL similarities. The trees bear strong resemblances to one other. For example, at the top level of the hierarchy, the Indo-European languages are discerned from the non Indo-European languages. Further down, within the Indo-European cluster, the Romance languages are separated from other Indo-European subgroups. Further points of similarity can be observed at the bottom of the hierarchy, where the pairs Russian and Polish, Japanese and Korean, and Chinese and Thai merge in both trees. We note that hierarchical clustering algorithms are known for not being stable with respect to their input, and different ESL trees may be obtained with different feature sets and distance mertic choices. Nonetheless, the similarities of the trees presented in 2-2 indicate that the typological similarity structure of our languages is likely to be reflected to a large extent in the uncertainty of the ESL classifier.

In the next section we evaluate the quality of our ESL based similarity estimates with respect to their ability to support accurate nearest neighbors based reconstruction of native language typology.

2.6 Typology Prediction

Although pairwise language similarities derived from structural features in ESL texts are highly correlated with similarities obtained directly from native language typology, evaluating the absolute quality of such similarity matrices and trees is challenging.

We therefore turn to typology prediction based evaluation, in which we assess the quality of the induced language similarity estimates by their ability to support accurate prediction of unseen typological features. In this evaluation mode we project unknown WALS features to a target language from the languages that are closest to it in the similarity structure. The underlying assumption of this setup is that better similarity structures will lead to



(a) Hierarchical clustering using WALS based *shared-pairwise* distances.

(b) Hierarchical clustering using ESL based distances.

Figure 2-2: Language Similarity Trees. Both trees are constructed with the Ward agglomerative hierarchical clustering algorithm. Tree (a) uses the WALS based *shared-pairwise* language distances. Tree (b) uses the ESL derived distances.

better accuracies in the feature prediction task.

Typological feature prediction not only provides an objective measure for the quality of the similarity structures, but also has an intrinsic value as a stand-alone task. The ability to infer typological structure automatically can be used to create linguistic databases for low-resource languages, and is valuable to NLP applications that exploit such resources, most notably multilingual parsing [71, 99].

Prediction of typological features for a target language using the language similarity matrix is performed by taking a majority vote for the value of each feature among the K nearest languages of the target language. In case none of the K nearest languages have a value for a feature, or given a tie between several values, we iteratively expand the group of nearest languages until neither of these cases applies.

To predict features using a hierarchical cluster tree, we set the value of each target language feature to its majority value among the members of the parent cluster of the target language, excluding the target language itself. For example, using the tree in figure 2-2(a), the feature values for the target language French will be obtained by taking majority votes between Portuguese, Italian and Spanish. Similarly to the matrix based prediction, missing

values and ties are handled by backing-off to a larger set of languages, in this case by proceeding to subsequent levels of the cluster hierarchy. For the French example in figure 2-2(a), the first fall-back option will be the Romance cluster.

Following the evaluation setups in Daumé III [22] and Georgi et al. [38], we evaluate the WALS based similarity estimates and trees by constructing them using 90% of the WALS features. We report the average accuracy over 100 random folds of the data. In the *shared-all* regime, we provide predictions not only for the remaining 10% of features shared by all languages, but also for all the other features that have values in the target language and are not used for the tree construction.

Importantly, as opposed to the WALS based prediction, our ESL based method does not require any typological features for inferring language similarities and constructing the similarity tree. In particular, no typological information is required for the target languages. Typological features are needed only for the neighbors of the target language, from which the features are projected. This difference is a key advantage of our approach over the WALS based methods, which presuppose substantial typological documentation for all the languages involved.

Table 2.3 summarizes the feature reconstruction results. The ESL approach is highly competitive with the WALS based results, yielding comparable accuracies for the *shared-all* prediction, and lagging only 1.7% – 3.4% behind the *shared-pairwise* construction. Also note that for both WALS based and ESL based predictions, the highest results are achieved using the hierarchical tree predictions, confirming the suitability of this representation for accurately capturing language similarity structure.

Figure 2-3 presents the performance of the strongest WALS based typological feature completion method, WALS *shared-pairwise* tree, as a function of the percentage of features used for obtaining the language similarity estimates. The figure also presents the strongest result of the ESL method, using the ESL tree, which does not require any such typological training data for obtaining the language similarities. As can be seen, the WALS based approach would require access to almost 40% of the currently documented WALS features to match the performance of the ESL method.

The competitive performance of our ESL method on the typology prediction task un-

Method	NN	3NN	Tree
WALS *shared-all*	71.6	71.4	69.1
WALS *shared-pairwise*	73.1	74.1	**74.2**
ESL	71.4	70.7	**72.2**

Table 2.3: Typology reconstruction results. Three types of predictions are compared, nearest neighbor (NN), 3 nearest neighbors (3NN) and nearest tree neighbors (Tree). WALS *shared-all* are WALS based predictions, where only the 32 features that have known values in all 14 languages are used for computing language similarities. In the WALS *shared-pairwise* predictions the language similarities are computed using the WALS features shared by each language pair. ESL results are obtained by projection of WALS features from the closest languages according to the ESL language similarities.

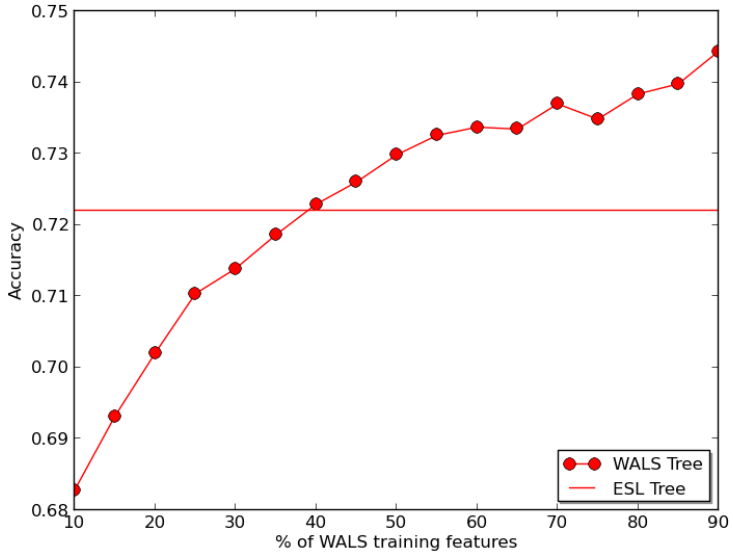


Figure 2-3: Comparison of the typological feature completion performance obtained using the WALS tree with *shared-pairwise* similarities and the ESL tree based typological feature completion performance. The dotted line represents the WALS based prediction accuracy, while the horizontal line is the ESL based accuracy. The horizontal axis corresponds to the percentage of WALS features used for constructing the WALS based language similarity estimates.

derlines its ability to extract strong typologically driven signal, while being robust to the partial nature of existing typological annotation which hinders the performance of the baselines. Given the small amount of ESL data at hand, these results are highly encouraging with regard to the prospects of our approach to support typological inference, even in the absence of any typological documentation for the target languages.

2.7 Related Work

Our investigation integrates two areas of computational research, cross-linguistic transfer and linguistic typology.

Cross-linguistic Influence in NLP

In NLP, the topic of linguistic transfer was mainly addressed in relation to the Native Language Identification (NLI) task, which requires predicting the native language of an ESL text's author. Following the work of Koppel et al. [48], NLI and related feature analysis [105, 13, 103, 97, 98, 59, 14], have been gaining increasing interest in NLP, including two shared tasks in 2013 [101] and 2017 [60]. The overall high performance on this classification task is considered to be a central piece of evidence for the existence of cross-linguistic transfer [43].

While high performance on the NLI task confirms the ability to extract native language signal from second language text, it offers limited insight on the systematicity of this process. Furthermore, as mentioned previously, some of the discriminative power of high performing NLI classifiers stems from cultural and task related artifacts. Our work incorporates an NLI component, but departs from the performance optimization orientation towards leveraging computational analysis for better understanding of the relations between native language typology and ESL usage. In particular, our choice of NLI features is driven by their relevance to linguistic typology rather than their contribution to classification performance.

The work described in this chapter is closest to [70] and [67] who use a language modeling approach to demonstrate that linguistic similarities are preserved in ESL usage. Our work provides converging evidence for this result and extends it in two ways. First, we introduce a new classification based approach for capturing ESL based language usage similarities. Secondly, and more crucially, we tie these similarities to linguistic typology rather than linguistic genealogy, a choice that has a better linguistic justification, and enables fine grained analysis of first language properties in relation to second language usage.

Finally, we note that recent work has demonstrated that linguistic trees can be reconstructed not only from second language usage, but also from source language signal in translations of foreign text to the native language of the writer [81]. Additional investigation of the relation between this result and first language influence will further advance our understanding of the interaction in the representation and usage of multiple languages by the same speaker.

Language Typology

The second area of research, language typology, deals with the documentation and comparative study of language structures [96]. Much of the descriptive work in the field is summarized in the World Atlas of Language Structures (WALS)[4] [29] in the form of structural features. We use the WALS features as our source of typological information.

Several previous studies have used WALS features for hierarchical clustering of languages and typological feature prediction. Most notably, Teh et al. [100] and subsequently Daumé III [22] predicted typological features from language trees constructed with a Bayesian hierarchical clustering model. In Georgi et al. [38] additional clustering approaches were compared using the same features and evaluation method. In addition to the feature prediction task, these studies also evaluated their clustering results by comparing them to genetic language clusters.

Our approach differs from this line of work in several aspects. First, similarly to our WALS based baselines, the clustering methods presented in these studies are affected by the sparsity of available typological data. Furthermore, these methods rely on existing typological documentation for the target languages. Both issues are obviated in our English based framework which does not depend on any typological information to construct the native language similarity structures, and does not require any knowledge about the target languages except from the ESL essays of a sample of their speakers. Finally, we do not compare our clustering results to genetic groupings, as to our knowledge, there is no firm theoretical ground for expecting typologically based clustering to reproduce language phylogenies. The empirical results in Georgi et al. [38], which show that typology based

[4] http://wals.info/

clustering differs substantially from genetic groupings, support this assumption.

2.8 Conclusion

In this chapter, we presented a novel framework for utilizing cross-linguistic transfer to infer language similarities from morpho-syntactic features of ESL text. Trading laborious expert annotation of typological features for a modest amount of ESL texts, we are able to reproduce language similarities that strongly correlate with the equivalent typology based similarities, and perform competitively on a typology reconstruction task.

Our study leaves multiple questions for future research. For example, while the current work examines structure transfer, additional investigation is required to better understand lexical and phonological transfer effects. Our focus in this chapter is on native language typology, where we use English as the foreign language. This limits our ability to study the constraints imposed on cross-linguistic transfer by the foreign language. An intriguing future research direction would be to explore other foreign languages and compare the outcomes to our results on English.

Chapter 3

From Typology to ESL Production

3.1 Introduction

Much of the linguistic work on cross-linguistic influence was carried out within the comparison based framework of Contrastive Analysis (CA), a theoretical approach that aims to explain difficulties in second language learning in terms of the relations between structures in the native and foreign languages. The basic hypothesis of CA was formulated by Lado [50], who suggested that "we can predict and describe the patterns that will cause difficulty in learning, and those that will not cause difficulty, by comparing systematically the language and culture to be learned with the native language and culture of the student". In particular, Lado postulated that divergences between the native and foreign languages will negatively affect learning and lead to increased error rates in the foreign language. This and subsequent hypotheses were soon met with criticism, targeting their lack of ability to provide reliable predictions, leading to an ongoing debate on the extent to which foreign language errors can be explained and predicted by examining native language structure.

Differently from the SLA tradition, which emphasizes manual analysis of error case studies [75], we address the heart of this controversy from a computational data-driven perspective, focusing on the issue of predictive power. We provide a formalization of the CA framework, and demonstrate that the relative frequency of grammatical errors in ESL can be reliably predicted from the typological properties of the native language and their

relation to the typology of English using a regression model.

Tested on 14 languages in a leave-one-out fashion, our model achieves a Mean Average Error (MAE) reduction of 21.8% in predicting the language specific relative frequency of the 20 most common ESL structural error types, as compared to the relative frequency of each of the error types in the training data, yielding improvements across all the languages and the large majority of the error types. Our regression model also outperforms a stronger, nearest neighbor based baseline, that projects the error distribution of a target language from its typologically closest language.

While our method presupposes the existence of typological annotations for the test languages, we also demonstrate its viability in low-resource scenarios for which such annotations are not available. To address this setup, we present a bootstrapping framework in which the typological features required for prediction of grammatical errors are approximated from automatically extracted ESL morpho-syntactic features using the method presented in chapter 2. Despite the noise introduced in this process, our bootstrapping strategy achieves an error reduction of 13.9% compared to the average frequency baseline.

The utilization of typological features as predictors also enables shedding light on linguistic factors that could give rise to different error types in ESL. For example, in accordance with common linguistic knowledge, feature analysis of the model suggests that the main contributor to increased rates of determiner omission in ESL is the lack of determiners in the native language. A more complex case of missing pronouns is intriguingly tied by the model to native language subject pronoun marking on verbs.

To summarize, the main contribution of this chapter is a CA inspired computational framework for learning language specific grammatical error distributions in ESL. Our approach is both predictive and explanatory. It enables us obtaining improved estimates for language specific error distributions without access to ESL error annotations for the target language. Coupling grammatical errors with typological information also provides meaningful explanations to some of the linguistic factors that drive the observed error rates.

3.2 Data

Similarely to chapter 2, we use the FCE as our source of ESL texts and WALS as our knowldge base of typological information.

ESL Corpus

We obtain ESL essays from the Cambridge First Certificate in English (FCE) learner corpus [113], a publicly available subset of the Cambridge Learner Corpus (CLC)[1]. As descibed in section 2.2, the essay authors represent 16 native languages. We discarded Dutch and Swedish speakers due to the small number of documents available for these languages (16 documents in total). The remaining documents are associated with the following 14 native languages: Catalan, Chinese, French, German, Greek, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Thai and Turkish. The resulting corpus is identical to the one used in chapter 2, with overall 1228 documents, corresponding to an average of 87.7 documents per native language.

In addition to the texts themselves, the FCE corpus has an elaborate error annotation scheme [73] and high quality of error annotations, making it particularly suitable for our investigation in this chapter. The annotation scheme encompasses 75 different error types, covering a wide range of grammatical errors on different levels of granularity. As the typological features used in this work refer mainly to *structural* properties, we filter out spelling errors, punctuation errors and open class semantic errors. This filtering step leaves us with a list of grammatical errors that are typically related to language structure. We focus on the 20 most frequent error types[2] in this list, which are presented and exemplified in table 3.1. In addition to concentrating on the most important structural ESL errors, this cutoff prevents us from being affected by data sparsity issues associated with less frequent errors.

[1] `http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603`

[2] Filtered errors that would have otherwise appeared in the top 20 list, with their respective rank in brackets: Spelling (1), Replace Punctuation (2), Replace Verb (3), Missing Punctuation (7), Replace (8), Replace Noun (9) Unnecessary Punctuation (13), Replace Adjective (18), Replace Adverb (20).

Rank	Code	Name	Example	Count	KW	MW
1	TV	Verb Tense	I hope I **give** *have given* you enough details	3324	**	34
2	RT	Replace Preposition	**on** *in* July	3311	**	31
3	MD	Missing Determiner	I went for *the* interview	2967	**	57
4	FV	Wrong Verb Form	had time to **played** *play*	1789	**	21
5	W	Word Order	**Probably** our homes will *probably* be	1534	**	34
6	MT	Missing Preposition	explain *to* you	1435	**	22
7	UD	Unnecessary Determiner	a course at **the** Cornell University	1321		
8	UT	Unnecessary Preposition	we need it **on** each minute	1079		
9	MA	Missing Pronoun	because *it* is the best conference	984	**	33
10	AGV	Verb Agreement	the teachers **was** *were* very experienced	916	**	21
11	FN	Wrong Form Noun	because of my **study** *studies*	884	**	24
12	RA	Replace Pronoun	she just met Sally, **which** *who*	847	**	17
13	AGN	Noun Agreement	two **month** *months* ago	816	**	24
14	RD	Replace Determiner	of **a** *the* last few years	676	**	35
15	DJ	Wrongly Derived Adjective	The mother was **pride** *proud*	608	*	8
16	DN	Wrongly Derived Noun	**working place** *workplace*	536		
17	DY	Wrongly Derived Adverb	**Especial** *Especially*	414	**	14
18	UA	Unnecessary Pronoun	feel **ourselves** comfortable	391	*	9
19	MC	Missing Conjunction	reading, *and* playing piano at home	346	*	11
20	RC	Replace Conjunction	not just the car, **and** *but* also the train	226		

Table 3.1: The 20 most frequent error types in the FCE corpus that are related to language structure. In the Example column, words marked in italics are corrections for the words marked in bold. The Count column lists the overall count of each error type in the corpus. The KW column depicts the result of the Kruskal-Wallis test whose null hypothesis is that the relative error frequencies for different native languages are drawn from the same distribution. Error types for which this hypothesis is rejected with $p < 0.01$ are denoted with '*'. Error types with $p < 0.001$ are marked with '**'. The MW column denotes the number of language pairs (out of the total 91 pairs) which pass the post-hoc Mann-Whitney test with $p < 0.01$.

Typological Database

As in chapter 2, we use the World Atlas of Language Structures (WALS), a repository of typological features of the world's languages, as our source of linguistic knowledge about the native languages of the ESL corpus authors. We perform several preprocessing steps in order to select the features for this study. First, as our focus is on structural features that can be expressed in written form, we discard all the features associated with the categories Phonology, Lexicon[3] and Other. We further discard 24 features which either have a documented value for only one language, or have the same value in all the languages. The resulting feature-set contains 119 features, with an average of 2.9 values per feature, and

[3]The discarded Lexicon features refer to properties such as the number of words in the language that denote colors, and identity of word pairs such as "hand" and "arm".

92.6 documented features per language.[4]

3.3 Variance Analysis of Grammatical Errors in ESL

To motivate a native language based treatment of grammatical error distributions in ESL, we begin by examining whether there is a statistically significant difference in ESL error rates based on the native language of the learners. This analysis provides empirical justification for our approach, and to the best of our knowledge was not conducted in previous studies.

To this end, we perform a Kruskal-Wallis (KW) test [49] for each error type[5]. We treat the relative error frequency per word in each document as a sample[6] (i.e. the relative frequencies of all the error types in a document sum to 1). The samples are associated with 14 groups, according to the native language of the document's author. For each error type, the null hypothesis of the test is that error fraction samples of all the native languages are drawn from the same underlying distribution. In other words, rejection of the null hypothesis implies a significant difference between the relative error frequencies of at least one language pair.

As shown in table 3.1, we can reject the null hypothesis for 16 of the 20 grammatical error types with $p < 0.01$, where Unnecessary Determiner, Unnecessary Preposition, Wrongly Derived Noun, and Replace Conjunction are the error types that do not exhibit dependence on the native language. Furthermore, the null hypothesis can be rejected for 13 error types with $p < 0.001$. These results suggest that the relative error rates of the majority of the common structural grammatical errors in our corpus indeed differ between native speakers of different languages.

We further extend our analysis by performing pairwise post-hoc Mann-Whitney (MW) tests [61] in order to determine the number of language pairs that significantly differ with

[4]Note that due to the removal of features with the same value for all languages, and the feature categories Lexicon and Other, the number of features used in this chapter is smaller than in chapter 2.

[5]We chose the non-parametric KW rank-based test over ANOVA, as according to the Shapiro-Wilk [93] and Levene [53] tests, the assumptions of normality and homogeneity of variance do not hold for our data. In practice, the ANOVA test yields similar results to those of the KW test.

[6]We also performed the KW test on the absolute error frequencies (i.e. raw counts) per word, obtaining similar results to the ones reported here on the relative frequencies per word.

respect to their native speakers' error fractions in ESL. Table 3.1 presents the number of language pairs that pass this test with $p < 0.01$ for each error type. This inspection suggests Missing Determiner as the error type with the strongest dependence on the author's native language, followed by Replace Determiner, Verb Tense, Word Order, Missing Pronoun and Replace Preposition.

3.4 Predicting Language Specific Error Distributions in ESL

3.4.1 Task Definition

Given a language $l \in L$, our task is to predict for this language the relative error frequency $y_{l,e}$ of each error type $e \in E$, where L is the set of all native languages, E is the set of grammatical errors, and $\sum_e y_{l,e} = 1$.

3.4.2 Model

In order to predict the error distribution of a native language, we train regression models on individual error types:

$$\hat{y}'_{l,e} = \theta_{l,e} \cdot f(t_l, t_{eng}) \tag{3.1}$$

In this equation $\hat{y}'_{l,e}$ is the predicted relative frequency of an error of type e for ESL documents authored by native speakers of language l, and $f(t_l, t_{eng})$ is a feature vector derived from the typological features of the native language t_l and the typological features of English t_{eng}.

The model parameters $\theta_{l,e}$ are obtained using Ordinary Least Squares (OLS) on the training data D, which consists of typological feature vectors paired with relative error frequencies of the remaining 13 languages:

$$D = \{(f(t_{l'}, t_{eng}), y_{e,l'}) | l' \in L, l' \neq l\} \tag{3.2}$$

To guarantee that the individual relative error frequency estimates sum up to 1 for each language, we renormalize them to obtain the final predictions:

$$\hat{y}_{l,e} = \frac{\hat{y}'_{l,e}}{\sum_e \hat{y}'_{l,e}} \tag{3.3}$$

3.4.3 Features

Our feature set can be divided into two subsets. The first subset, used in a version of our model called *Reg*, contains the typological features of the native language. In a second version of our model, called *RegCA*, we also utilize additional features that explicitly encode differences between the typological features of the native language, and the typological features of English.

Typological Features In the *Reg* model, we use the typological features of the native language that are documented in WALS. As mentioned in section 3.2, WALS features belong to different variable types, and are hence challenging to encode. We address this issue by binarizing all the features. Given k possible values v_k for a given WALS feature t_i, we generate k binary typological features of the form:

$$f_{i,k}(t_l, t_{eng}) = \begin{cases} 1 & \text{if } t_{l,i} = v_k \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

When a WALS feature of a given language does not have a documented value, all k entries of the feature for that language are assigned the value of 0. This process transforms the original 119 WALS features into 340 binary features.

Divergences from English In the spirit of CA, in the model *RegCA*, we also utilize features that explicitly encode differences between the typological features of the native language and those of English. These features are also binary, and take the value 1 when the value of a WALS feature in the native language is different from the corresponding

value in English:

$$f_i(t_l, t_{eng}) = \begin{cases} 1 & \text{if } t_{l,i} \neq t_{eng,i} \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

We encode 104 such features, in accordance with the typological features of English available in WALS. These features are encoded only when a typological feature of English has a corresponding documented feature in the native language. The addition of these divergence features brings the total number of features in our feature set to 444.

3.4.4 Results

We evaluate the model predictions using two metrics. The first metric, Absolute Error, measures the distance between the predicted and the true relative frequency of each grammatical error type[7]:

$$\text{Absolute Error} = |\hat{y}_{l,e} - y_{l,e}| \tag{3.6}$$

When averaged across different predictions we refer to this metric as Mean Absolute Error (MAE).

The second evaluation score is the Kullback-Leibler divergence D_{KL}, a standard measure for evaluating the difference between two distributions. This metric is used to evaluate the predicted grammatical error distribution of a native language:

$$D_{KL}(y_l||\hat{y}_l) = \sum_e y_{l,e} \ln \frac{y_{l,e}}{\hat{y}_{l,e}} \tag{3.7}$$

Table 3.2 summarizes the grammatical error prediction results[8]. The baseline model *Base* sets the relative frequencies of the grammatical errors of a test language to the respective relative error frequencies in the training data. We also consider a stronger, language specific model called *Nearest Neighbor (NN)*, which projects the error distribution of a target language from the typologically closest language in the training set, according to

[7]For clarity of presentation, all the reported results on this metric are multiplied by 100.

[8]As described in section 3.4.2, we report the performance of regression models trained and evaluated on relative error frequencies obtained by normalizing the rates of the different error types. We also experimented with training and evaluating the models on absolute error counts per word, obtaining results that are similar to those reported here.

	Base	**NN**	**Reg**	**RegCA**
MAE	1.28	1.11	1.02	**1.0**
Error Reduction	-	13.3	20.4	**21.8**
#Languages	-	9/14	12/14	**14/14**
#Mistakes	-	11/20	**15/20**	14/20
AVG $\mathbf{D_mathitKL}$	0.052	0.046	0.033	**0.032**
#Languages	-	10/14	**14/14**	**14/14**

Table 3.2: Results for prediction of relative error frequencies using the *MAE* metric across languages and error types, and the $D_mathitKL$ metric averaged across languages. *#Languages* and *#Mistakes* denote the number of languages and grammatical error types on which a model outperforms *Base*.

the cosine similarity measure. This baseline provides a performance improvement for the majority of the languages and error types, with an average error reduction of 13.3% on the *MAE* metric compared to *Base*, and improving from 0.052 to 0.046 on the KL divergence metric, thus emphasizing the general advantage of a native language adapted approach to ESL error prediction.

Our regression model introduces further substantial performance improvements. The *Reg* model, which uses the typological features of the native language for predicting ESL relative error frequencies, achieves 20.4% *MAE* reduction over the *Base* model. The *RegCA* version of the regression model, which also incorporates differences between the typological features of the native language and English, surpasses the *Reg* model, reaching an average error reduction of 21.8% from the *Base* model, with improvements across all the languages and the majority of the error types. Strong performance improvements are also obtained on the KL divergence measure, where the *RegCA* model scores 0.032, compared to the baseline score of 0.052.

For various practical purposes, such as the development of educational curricula, one may be interested only in the ranking of the errors for each language rather than their exact frequency. We therefore also evaluate the predictions with respect to the Kendall rank coefficient [45]. On this evaluation metric, the *NN* model achieves an average improvement of 6.4%, with superior rankings for 11 languages relative to the *Base* model. The *Reg* model further improves the average ranking score by 6.5%, with gains on 12 languages. Finally,

the *RegCA* model achieves the best ranking scores, outperforming the *Base* model by 7.3%, with improvements on 13 languages.

To illustrate the outcome of our approach, consider the example in table 3.3, which compares the top 10 predicted errors for Japanese using the *Base* and *RegCA* models. In this example, *RegCA* correctly places Missing Determiner as the most common error in Japanese, with a significantly higher relative frequency than in the training data. Similarly, it provides an accurate prediction for the Missing Preposition error, whose frequency and rank are underestimated by the *Base* model. Furthermore, *RegCA* correctly predicts the frequency of Replace Preposition and Word Order to be lower than the average in the training data.

Rank	Base	Frac.	RegCA	Frac.	True	Frac.
1	**Replace Preposition**	**0.14**	**Missing Determiner**	**0.18**	**Missing Determiner**	**0.20**
2	Tense Verb	0.14	Tense Verb	0.12	Tense Verb	0.12
3	**Missing Determiner**	**0.12**	**Replace Preposition**	**0.12**	**Replace Preposition**	**0.10**
4	Wrong Verb Form	0.07	**Missing Preposition**	**0.08**	**Missing Preposition**	**0.08**
5	**Word Order**	**0.06**	Unnecessary Determiner	0.06	Unnecessary Preposition	0.06
6	**Missing Preposition**	**0.06**	Wrong Verb Form	0.05	Unnecessary Determiner	0.05
7	Unnecessary Determiner	0.06	Unnecessary Preposition	0.05	Replace Determiner	0.05
8	Unnecessary Preposition	0.04	Wrong Noun Form	0.05	Wrong Verb Form	0.05
9	Missing Pronoun	0.04	**Word Order**	**0.05**	**Word Order**	**0.04**
10	Wrong Noun Form	0.04	Verb Agreement	0.04	Wrong Noun Form	0.06

Table 3.3: Comparison between the fractions and ranks of the top 10 predicted error types by the *Base* and *RegCA* models for Japanese. As opposed to the *Base* method, the *RegCA* model correctly predicts Missing Determiner to be the most frequent error committed by native speakers of Japanese. It also correctly predicts Missing Preposition to be more frequent and Replace Preposition and Word Order to be less frequent than in the training data.

3.4.5 Feature Analysis

The clear semantics of the features are an important advantage of our typology-based approach, as they facilitate the interpretation of the model. Inspection of the model parameters allows us to gain insight into the typological features that are potentially involved in causing different types of ESL errors. Although such inspection is unlikely to provide a comprehensive coverage of all the relevant causes for the observed learner difficulties, it

can serve as a valuable starting point for exploratory linguistic analysis and formulation of a cross-linguistic transfer theory.

Table 3.4 lists the most salient typological features, as determined by the feature weights averaged across the models of different languages, for the error types Missing Determiner and Missing Pronoun. In the case of determiners, the model identifies the lack of definite and indefinite articles in the native language as the strongest factors related to increased rates of determiner omission. Conversely, features that imply the presence of an article system in the native language, such as 'Indefinite word same as one' and 'Definite word distinct from demonstrative' are indicative of reduced error rates of this type.

Missing Determiner	
37A Definite Articles: Different from English	.057
38A Indefinite Articles: No definite or indefinite article	.055
37A Definite Articles: No definite or indefinite article	.055
49A Number of Cases: 6-7 case	.052
100A Alignment of Verbal Person Marking: Accusative	-.073
38A Indefinite Article: Indefinite word same as 'one'	-.050
52A Comitatives and Instrumentals: Identity	-.044
37A Definite Articles: Definite word distinct from demonstrative	-.036
Missing Pronoun	
101A Expression of Pronominal Subjects: Subject affixes on verb	.015
71A The Prohibitive: Different from English	.012
38A Indefinite Articles: Indefinite word same as 'one'	.011
71A The Prohibitive: Special imperative + normal negative	.010
104A Order of Person Markers on the Verb: A & P do not or do not both occur on the verb	-.016
102A Verbal Person Marking: Only the A argument	-.013
101A Expression of Pronominal Subjects: Obligatory pronouns in subject position	-.011
71A The Prohibitive: Normal imperative + normal negative	-.010

Table 3.4: The most predictive typological features of the *RegCA* model for the errors Missing Determiner and Missing Pronoun. The right column depicts the feature weight averaged across all the languages. Missing determiners are related to the absence of a determiner system in the native language. Missing pronouns are correlated with subject pronoun marking on the verb.

A particularly intriguing example concerns the Missing Pronoun error. The most predictive typological factor for increased pronoun omissions is pronominal subject marking on the verb in the native language. Differently from the case of determiners, it is not

the lack of the relevant structure in the native language, but rather its different encoding that seems to drive erroneous pronoun omission. Decreased error rates of this type correlate most strongly with obligatory pronouns in subject position, as well as a verbal person marking system similar to the one in English.

3.5 Bootstrapping with ESL-based Typology

Thus far, we presupposed the availability of substantial typological information for our target languages in order to predict their ESL error distributions. However, as previously mentioned, the existing typological documentation for the majority of the world's languages is scarce, limiting the applicability of this approach for low-resource languages. We address this challenge for scenarios in which an unannotated collection of ESL texts authored by native speakers of the target language is available. Given such data, we propose a bootstrapping strategy which uses the method proposed in chapter 2 in order to approximate the typology of the native language from morpho-syntactic features in ESL. The inferred typological features serve, in turn, as a proxy for the true typology of that language in order to predict its speakers' ESL grammatical error rates with our regression model.

To put this framework into effect, we use the log-linear model for native language classification described in section 2.3 to infer the ESL similarity usage matrix S_{ESL}. As shown in section 2.3, given this similarity matrix, one can obtain an approximation for the typology of a native language by projecting the typological features from its most similar languages. Here, we use the typological features of the closest language. In the bootstrapping setup, we train the regression models on the true typology of the languages in the training set, and use the approximate typology of the test language to predict the relative error rates of its speakers in ESL.

Results

Table 3.5 summarizes the error prediction results using approximate typological features for the test languages. As can be seen, our approach continues to provide substantial performance gains despite the inaccuracy of the typological information used for the test lan-

guages. The best performing method, *RegCA* reduces the *MAE* of *Base* by 13.9%, with performance improvements for most of the languages and error types. Performance gains are also obtained on the D_{KL} metric, whereby *RegCA* scores 0.041, compared to the *Base* score of 0.052, improving on 11 out of our 14 languages.

	Base	**NN**	**Reg**	**RegCA**
MAE	1.28	1.12	1.13	**1.10**
Error Reduction	-	12.6	11.6	**13.9**
#Languages	-	**11/14**	**11/14**	**11/14**
#Mistakes	-	10/20	10/20	**11/20**
AVG $\mathbf{D}_{KL}$	0.052	0.048	0.043	**0.041**
#Languages	-	10/14	**11/14**	**11/14**

Table 3.5: Results for prediction of relative error frequencies using the bootstrapping approach. In this setup, the true typology of the test language is substituted with approximate typology derived from morpho-syntactic ESL features.

3.6 Related Work

Rooted in the comparative linguistics tradition, CA was first suggested by Fries [34] and formalized by Lado [50]. In essence, CA examines foreign language performance, with a particular focus on learner difficulties, in light of a structural comparison between the native and the foreign languages. From its inception, CA was criticized for the lack of a solid predictive theory [108, 110], leading to an ongoing scientific debate on the relevance of comparison based approaches. Important to our study is that the type of evidence used in this debate typically relies on small scale manual case study analysis. Our work seeks to reexamine the issue of predictive power of CA based methods using a computational, data-driven approach.

As noted in section 2.7, computational work touching on cross-linguistic transfer was mainly conducted in relation to the Native Language Identification (NLI) task, in which the goal is to determine the native language of the author of an ESL text. Much of this work focuses on experimentation with different feature sets [101], including features derived from the CA framework [112]. A related line of inquiry which is closer to our work deals

with the identification of ESL syntactic patterns that are specific to speakers of different native languages [97, 98]. Our approach differs from this research direction by focusing on grammatical errors, and emphasizing prediction of language specific patterns rather than their identification.

Previous work on grammatical error correction that examined determiner and preposition errors [87, 88] incorporated native language specific priors in models that are otherwise trained on standard English text. Our work extends the native language tailored treatment of grammatical errors to a much larger set of error types. More importantly, the approach in [87, 88] is limited by the availability of manual error annotations for the target language in order to obtain the required error counts. Our framework enables bypassing this annotation bottleneck by predicting language specific priors from typological information. More recent work further confirmed the utility of native language specific information for grammatical error detection [47].

3.7 Conclusion

We present a computational framework for predicting native language specific grammatical error distributions in ESL, based on the typological properties of the native language and their compatibility with the typology of English. Our model achieves substantial performance improvements compared to a language oblivious baseline, as well as a language dependent nearest neighbor baseline. Furthermore, we address scenarios in which the typology of the native language is not available, by bootstrapping typological features from ESL texts. Finally, inspection of the model parameters allows us to identify native language properties which play a pivotal role in generating different types of grammatical errors.

In addition to the theoretical contribution, the outcome of our work has a strong potential to be beneficial in practical setups. In particular, it can be utilized for developing educational curricula that focus on the areas of difficulty that are characteristic of different native languages. Furthermore, the derived error frequencies can be integrated as native language specific priors in systems for automatic error correction. In both application areas, previous work relied on the existence of error tagged ESL data for the languages of

interest. Our approach paves the way for addressing these challenges even in the absence of such data.

Chapter 4

Predicting Native Language in Comprehension

4.1 Introduction

In this chapter, we present a novel framework for studying cross-linguistic influence in language comprehension using *eyetracking during reading of free-form native English text*. We collect and analyze English newswire reading data from 182 participants, including 145 English as Second Language (ESL) learners from four different native language backgrounds: Chinese, Japanese, Portuguese and Spanish, as well as 37 native English speakers. These languages were chosen based on the availability of their native speakers in the Boston area and the linguistic diversity of the resulting language set. Each participant reads 156 English sentences, half of which are shared across all participants, and the remaining half are individual to each participant. All the sentences are manually annotated with part-of-speech (POS) tags and syntactic dependency trees.

We then introduce the task of *Native Language Identification from Reading (NLIR)*, which requires predicting a subject's native language from gaze while reading text in a second language. Focusing on ESL participants and using a log-linear classifier with word fixation times normalized for reading speed as features, we obtain 71.03 NLIR accuracy in the shared sentences regime. We further demonstrate that NLIR can be generalized

effectively to the individual sentences regime, in which each subject reads a different set of sentences, by grouping fixations according to linguistically motivated clustering criteria. In this regime, we obtain an NLIR accuracy of 51.03.

Further on, we provide classification and feature analyses, suggesting that the signal underlying NLIR is likely to be related to *linguistic* characteristics of the respective native languages. First, drawing on previous work on ESL production, we observe that classifier uncertainty in NLIR correlates with global linguistic similarities across native languages. In other words, the more similar are the languages, the more similar are the reading patterns of their native speakers in English. Second, we perform feature analysis across native and non-native English speakers, and discuss structural and lexical factors that could potentially drive some of the non-native reading patterns in each of our native languages. Taken together, our results provide evidence for a systematic influence of native language properties on reading, and by extension, on online processing and comprehension in a second language.

To summarize, we introduce a novel framework for studying cross-linguistic influence in language learning by using eyetracking for reading free-form English text. We demonstrate the utility of this framework in the following ways. First, we obtain the first NLIR results, addressing both the shared and the individual textual input scenarios. We further show that reading preserves linguistic similarities across native languages of ESL readers, and perform feature analysis, highlighting key distinctive reading patterns in each native language. The proposed framework complements and extends production studies, and can inform linguistic inquiry on cross-linguistic influence.

4.2 Experimental Setup

Participants

We recruited 182 adult participants. Of those, 37 are native English speakers and 145 are ESL learners from four native language backgrounds: Chinese, Japanese, Portuguese and Spanish. All the participants in the experiment are native speakers of only one language.

The ESL speakers were tested for English proficiency using the grammar and listening sections of the Michigan English test (MET), which consist of 50 multiple choice questions. The English proficiency score was calculated as the number of correctly answered questions on these modules. The majority of the participants scored in the intermediate-advanced proficiency range. Table 4.1 presents the number of participants and the mean English proficiency score for each native language group. Additionally, we collected metadata on gender, age, level of education, duration of English studies and usage, time spent in English speaking countries and proficiency in any additional language spoken.

	# Participants	**English Score**
Chinese	36	42.0
Japanese	36	40.3
Portuguese	36	41.1
Spanish	37	42.4
English	37	NA

Table 4.1: Number of participants and mean MET English score by native language group.

Reading Materials

We utilize 14,274 randomly selected sentences from the Wall Street Journal part of the Penn Treebank (WSJ-PTB) [62]. To support reading convenience and measurement precision, the maximal sentence length was set to 100 characters, leading to an average sentence length of 11.4 words. Word boundaries are defined as whitespaces. From this sentence pool, 78 sentences (900 words) were presented to all participants (henceforth *shared* sentences) and the remaining 14,196 sentences were split into 182 individual batches of 78 sentences (henceforth *individual* sentences, averaging 880 words per batch).

All the sentences include syntactic annotations from the Universal Dependency Treebank project (UDT) [65]. The annotations include PTB POS tags [91], Google universal POS tags [77] and dependency trees. The dependency annotations of the UDT are converted automatically from the manual phrase structure tree annotations of the WSJ-PTB.

Gaze Data Collection

Each participant read 157 sentences. The first sentence was presented to familiarize participants with the experimental setup and was discarded during analysis. The following 156 sentences consisted of 78 shared and 78 individual sentences. The shared and the individual sentences were mixed randomly and presented to all participants in the same order. The experiment was divided into three parts, consisting of 52 sentences each. Participants were allowed to take a short break between experimental parts.

Each sentence was presented on a blank screen as a one-liner. The text appeared in Times font, with font size 23. To encourage attentive reading, upon completion of sentence reading participants answered a simple yes/no question about its content, and were subsequently informed if they answered the question correctly. Both the sentences and the questions were triggered by a 300ms gaze on a fixation target (fixation circle for sentences and the letter "Q" for questions) which appeared on a blank screen and was co-located with the beginning of the text in the following screen.

Throughout the experiment, participants held a joystick with buttons for indicating completion of sentence reading and answering the comprehension questions. Eye-movement of participants' dominant eye was recorded using a desktop mount Eyelink 1000 eyetracker, at a sampling rate of 1000Hz. Further details on the experimental setup are provided in appendix 4.7.

4.3 Native Language Identification from Reading

Our first goal is to determine whether the native language of ESL learners can be decoded from their gaze patterns while reading English text. We address this question in two regimes, corresponding to our division of reading input into shared and individual sentences. In the *shared regime*, all the participants read the same set of sentences. Normalizing over the reading input, this regime facilitates focusing on differences in reading behavior across readers. In the *individual regime*, we use the individual batches from our data to address the more challenging variant of the NLIR task in which the reading material

given to each participant is different.

4.3.1 Features

We seek to utilize features that can provide robust, simple and interpretable characterizations of reading patterns. To this end, we use speed normalized *fixation duration* measures over word sequences.

Fixation Measures

We utilize three measures of word fixation duration:

- *First Fixation duration (FF)* Duration of the first fixation on a word.
- *First Pass duration (FP)* Time spent from first entering a word to first leaving it (including re-fixations within the word).
- *Total Fixation duration (TF)* The sum of all fixation times on a word.

We experiment with fixations over unigram, bigram and trigram sequences $seq_{i,k} = w_i, ..., w_{i+k-1}, k \in \{1, 2, 3\}$, where for each metric $M \in \{FF, FP, TF\}$ the fixation time for a sequence $M_{seq_{i,k}}$ is defined as the sum of fixations on individual tokens M_w in the sequence[1].

$$M_{seq_{i,k}} = \sum_{w' \in seq_{i,k}} M_{w'} \tag{4.1}$$

Importantly, we control for variation in reading speeds across subjects by normalizing each subjects's sequence fixation times. For each metric M and sequence $seq_{i,k}$ we normalize the sequence fixation time $M_{seq_{i,k}}$ relative to the subject's sequence fixation times in the textual context of the sequence. The context C is defined as the sentence in which the sequence appears for the *Words in Fixed Context* feature-set and the entire textual input for the *Syntactic* and *Information* clusters feature-sets (see definitions of feature-sets below). The normalization term $S_{M,C,k}$ is accordingly defined as the metric's fixation time

[1]Note that for bigrams and trigrams, one could also measure FF and FP for interest regions spanning the sequence, instead, or in addition to summing these fixation times over individual tokens.

per sequence of length k in the context:

$$S_{M,C,k} = \frac{1}{|C|} \sum_{seq_k \in C} M_{seq_k} \tag{4.2}$$

We then obtain a normalized fixation time $Mnorm_{seq_{i,k}}$ as:

$$Mnorm_{seq_{i,k}} = \frac{M_{seq_{i,k}}}{S_{M,C,k}} \tag{4.3}$$

Feature Types

We use the above presented speed normalized fixation metrics to extract three feature-sets, *Words in Fixed Context (WFC)*, *Syntactic Clusters* (SC) and *Information Clusters (IC)*. WFC is a token-level feature-set that presupposes a fixed textual input for all participants. It is thus applicable only in the shared sentences regime. SC and IC are type-level features which provide abstractions over sequences of words. Crucially, they can also be applied when participants read different sentences.

- **Words in Fixed Context (WFC)** The WFC features capture fixation times on word sequences in a specific sentence. This feature-set consists of FF, FP and TF times for each of the 900 unigram, 822 bigram, and 744 trigram word sequences comprising the shared sentences. The fixation times of each metric are normalized for each participant relative to their fixations on sequences of the same length in the surrounding sentence. As noted above, the WFC feature-set is not applicable in the individual regime, as it requires identical sentences for all participants.

- **Syntactic Clusters (SC)** SC features are average globally normalized FF, FP and TF times for word sequences clustered by our three types of syntactic labels: universal POS, PTB POS, and syntactic relation labels. An example of such a feature is the average of speed-normalized TF times spent on the PTB POS bigram sequence DT NN. We take into account labels that appear at least once in the reading input of all participants. On the four non-native languages, considering all three label types, we obtain 104 unigram, 636 bigram and 1,310 trigram SC features per fixation metric in

the shared regime, and 56 unigram, 95 bigram and 43 trigram SC features per fixation metric in the individual regime.

- **Information Clusters (IC)** We also obtain average FF, FP and TF for words clustered according to their *length*, measured in number of characters. Word length was previously shown to be a strong predictor of *information content* [78]. As such, it provides an alternative abstraction to the syntactic clusters, combining both syntactic and lexical information. As with SC features, we take into account features that appear at least once in the textual input of all participants. For our set of non-native languages, we obtain for each fixation metric 15 unigram, 21 bigram and 23 trigram IC features in the shared regime, and 12 unigram, 18 bigram and 18 trigram IC features in the individual regime. Notably, this feature-set is very compact, and differently from the syntactic clusters, does not rely on the availability of external annotations.

In each feature-set, we perform a final preprocessing step for each individual feature, in which we derive a zero mean unit variance scaler from the training set feature values, and apply it to transform both the training and the test values of the feature to Z scores.

4.3.2 Model

The experiments are carried out using a log-linear model:

$$p(y|x;\theta) = \frac{\exp(\theta \cdot f(x,y))}{\sum_{y' \in Y} \exp(\theta \cdot f(x,y'))} \tag{4.4}$$

where y is the reader's native language, x is the reading input and θ are the model parameters. The classifier is trained with gradient descent using L-BFGS [15].

4.3.3 Experimental Results

In table 4.2 we report 10-fold cross-validation results on NLIR in the shared and the individual experimental regimes for native speakers of Chinese, Japanese, Portuguese and Spanish. We introduce two baselines against which we compare the performance of our

	Shared Sentences Regime			Individual Sentences Regime		
Majority Class	25.52			25.52		
Random Clusters	22.76			22.07		
	unigrams	+bigrams	+trigrams	unigrams	+bigrams	+trigrams
Information Clusters (IC)	41.38	44.14	46.21	38.62	32.41	32.41
Syntactic Clusters (SC)	45.52	57.24	58.62	48.97	43.45	48.28
SC+IC	51.72	57.24	60.0	**51.03**	46.21	49.66
Words in Fixed Context (WFC)	64.14	68.28	**71.03**	NA		

Table 4.2: Native Language Identification from Reading results with 10-fold cross-validation for native speakers of Chinese, Japanese, Portuguese and Spanish. In the *Shared* regime all the participants read the same 78 sentences. In the *Individual* regime each participant reads a different set of 78 sentences.

feature-sets. The *majority* baseline selects the native language with the largest number of participants. The *random clusters* baseline clusters words into groups randomly, with the number of groups set to the number of syntactic categories in our data.

In the shared regime, WFC fixations yield the highest classification rates, substantially outperforming the cluster feature-sets and the two baselines. The strongest result using this feature-set, 71.03, is obtained by combining unigram, bigram and trigram fixation times. In addition to this outcome, we note that training binary classifiers in this setup yields accuracies ranging from 68.49 for the language pair Portuguese and Spanish, to 93.15 for Spanish and Japanese. These results confirm the effectiveness of the shared input regime for performing reliable NLIR, and suggest a strong native language signal in non-native reading fixation times.

SC features yield accuracies of 45.52 to 58.62 on the shared sentences, while IC features exhibit weaker performance in this regime, with accuracies of 41.38 to 46.21. Both results are well above chance, but lower than WFC fixations due to the information loss imposed by the clustering step. Crucially, both feature-sets remain effective in the individual input regime, with 43.45 to 48.97 accuracy for SC features and 32.41 to 38.62 accuracy for IC features. The strongest result in the individual regime is 51.03, obtained by concatenating IC and SC features over unigrams. We also note that using this setup in a binary classification scheme yields results ranging from chance level 49.31 for Portuguese versus Spanish, to 84.93 on Spanish versus Japanese.

Generally, we observe that adding bigram and trigram fixations in the shared regime

leads to performance improvements compared to using unigram features only. This trend does not hold for the individual sentences, presumably due to a combination of feature sparsity and context variation in this regime. We also note that IC and SC features tend to perform better together than in separation, suggesting that the information encoded using these feature-sets is to some extent complementary.

The generalization power of our cluster based feature-sets has both practical and theoretical consequences. Practically, they provide useful abstractions for performing NLIR over arbitrary textual input. That is, they enable performing this task using *any* textual input during both training and testing phases. Theoretically, the effectiveness of linguistically motivated features in discerning native languages suggests that linguistic factors play an important role in the ESL reading process. The analysis presented in the following sections will further explore this hypothesis.

4.4 Analysis of Cross-Linguistic Influence in ESL Reading

As mentioned in the previous section, the ability to perform NLIR in general, and the effectiveness of linguistically motivated features in particular, suggest that linguistic factors in the native and second languages are pertinent to ESL reading. In this section we explore this hypothesis further, by analyzing classifier uncertainty and the features learned in the NLIR task.

4.4.1 Preservation of Linguistic Similarity

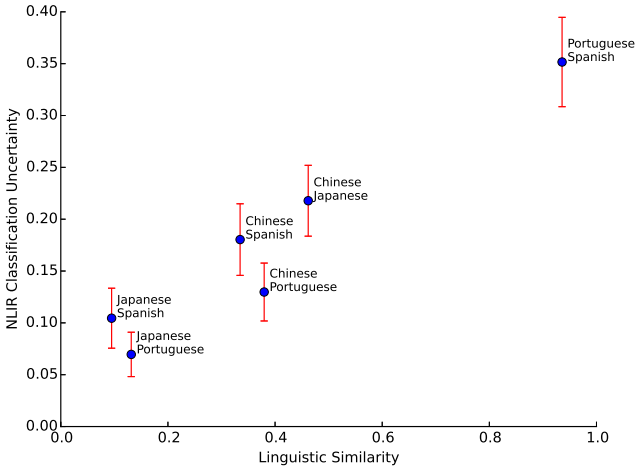
Chapters 2 and 3, as well as [70, 67] suggested a link between textual patterns in ESL production and linguistic similarities of the respective native languages. In particular, chapter 2 has demonstrated that NLI classification uncertainty correlates with similarities between languages with respect to their typological features. Here, we extend this framework and examine if preservation of native language similarities in ESL production is paralleled in reading.

Similarly to section 2.3, we define the classification uncertainty for a pair of native languages y and y' in our data collection D, as the average probability assigned by the NLIR classifier to one language given the other being the true native language. This approach provides a robust measure of classification confusion that does not rely on the actual performance of the classifier. We interpret the classifier uncertainty as a similarity measure between the respective languages and denote it as English Reading Similarity ERS.

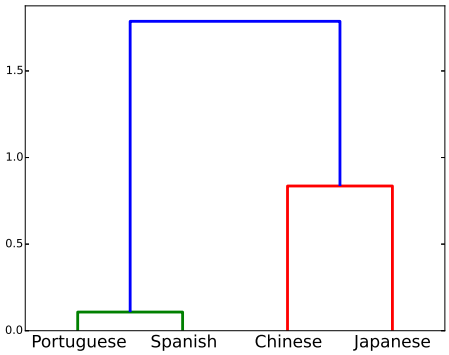
We compare these reading similarities to the linguistic similarities between our native languages. To approximate these similarities, we utilize feature vectors from the URIEL Typological Compendium [56] extracted using the *lang2vec* tool [57]. URIEL aggregates, fuses and normalizes typological, phylogenetic and geographical information about the world's languages. We obtain all the 103 available morpho-syntactic features in URIEL, which are derived from the World Atlas of Language Structures (WALS) [29], Syntactic Structures of the World's Languages (SSWL) [19] and Ethnologue [55]. Missing feature values are completed with a KNN classifier. We also extract URIEL's 3,718 language family features derived from Glottolog [41]. Each of these features represents membership in a branch of Glottolog's world language tree. Truncating features with the same value for all our languages, we remain with 76 features, consisting of 49 syntactic features and 27 family tree features. The linguistic similarity LS between a pair of languages y and y' is then determined by the cosine similarity of their URIEL feature vectors.

Figure 4-1 presents the URIEL based linguistic similarities for our set of non-native languages against the average NLIR classification uncertainties on the cross-validation test samples. The results presented in this figure are based on the unigram IC+SC feature-set in the individual sentences regime. We also provide a graphical illustration of the language similarities for each measure, using the Ward clustering algorithm [107]. We observe a correlation between the two measures which is also reflected in similar hierarchies in the two language trees. Thus, linguistically motived features in English reveal linguistic similarities across native languages. This outcome supports the hypothesis that English reading differences across native languages are related to linguistic factors.

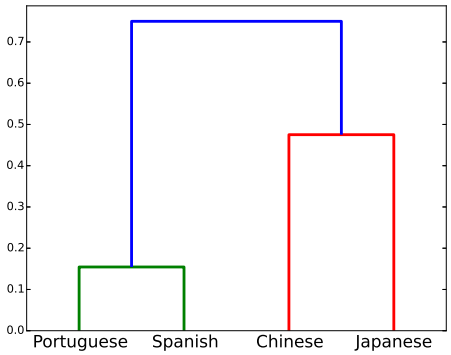
We note that while comparable results are obtained for the IC and SC feature-sets, together and in separation in the shared regime, WFC features in the shared regime do not



(a) Linguistic similarities against mean NLIR classification uncertainty. Error bars denote standard error.



(b) Linguistic tree



(c) English reading tree

Figure 4-1: (a) Linguistic versus English reading language similarities. The horizontal axis represents typological and phylogenetic similarity between languages, obtained by vectorizing linguistic features form URIEL, and measuring their cosine similarity. The vertical axis represents the average uncertainty of the NLIR classifier in distinguishing ESL readers of each language pair. (b) Ward hierarchical clustering of linguistic similarities between languages. (c) Ward hierarchical clustering of NLIR average pairwise classification uncertainties.

exhibit a clear uncertainty distinction when comparing across the pairs Japanese and Spanish, Japanese and Portuguese, Chinese and Spanish, and Chinese and Portuguese. Instead, this feature-set yields very low uncertainty, and correspondingly very high performance ranging from 90.41 to 93.15, for all four language pairs.

4.4.2 Feature Analysis

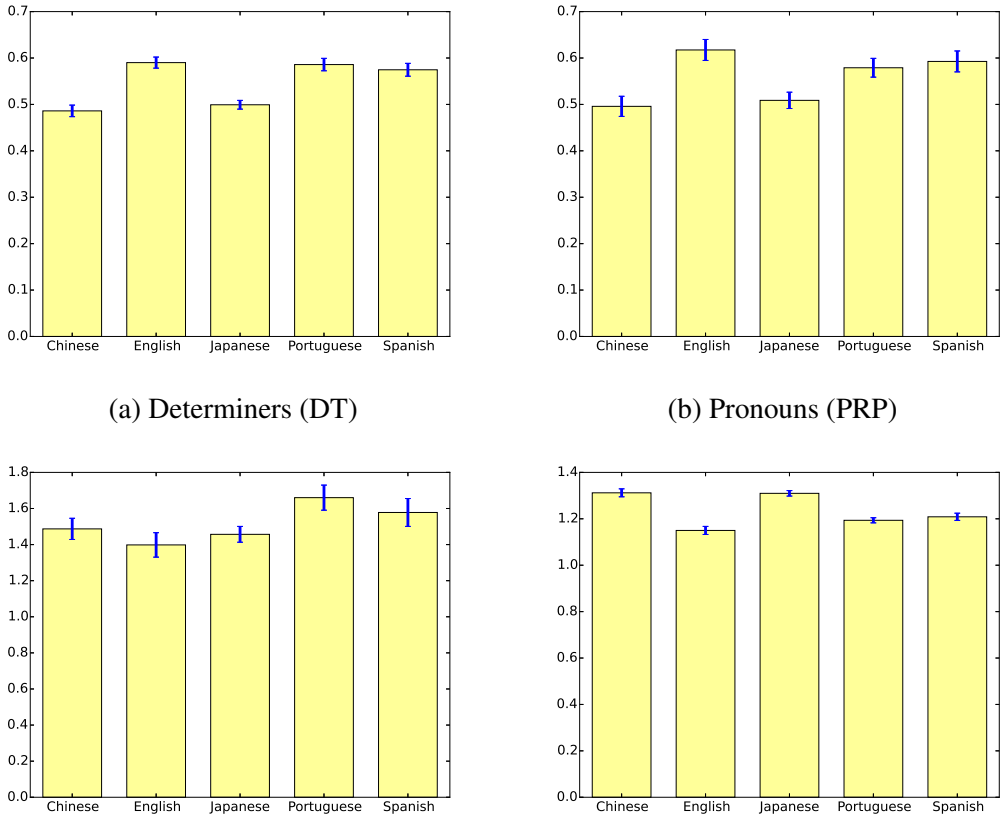
Our framework enables not only native language classification, but also exploratory analysis of native language specific reading patterns in English. The basic question that we examine in this respect is on which features do readers of different native language groups spend more versus less time. We also discuss several potential relations of the observed reading time differences to usage patterns and grammatical errors committed by speakers of our four native languages in production. We obtain this information by extracting grammatical error counts from the CLC FCE corpus [113], and from the ngram frequency analysis in Nagata and Whittaker [70].

In order to obtain a common benchmark for reading time comparisons across non-native speakers, in this analysis we also consider our group of native English speakers. In this context, we train four binary classifiers that discern each of the non-native groups from native English speakers based on TF times over unigram PTB POS tags in the shared regime. The features with the strongest positive and negative weights learned by these classifiers are presented in table 4.3. These features serve as a reference point for selecting the case studies discussed below.

	Negative (Fast)	**Positive (Slow)**
Chinese	**DT**	JJR
	PRP	**NN**
Japanese	**DT**	**NN**
	CD	VBD
Portuguese	NNS	**NN-POS**
	PRP	VBZ
Spanish	NNS	MD
	PRP	RB

Table 4.3: PTB POS features with the strongest weights learned in non-native versus native classification for each native language in the shared regime. Feature types presented in figure 4-2 are highlighted in bold.

Interestingly, some of the reading features that are most predictive of each native language lend themselves to linguistic interpretation with respect to *structural* factors. For example, in Japanese and Chinese we observe shorter reading times for determiners (DT),



(a) Determiners (DT)

(b) Pronouns (PRP)

(c) Possessives (NN+POS)

(d) Nouns (NN)

Figure 4-2: Mean speed-normalized Total Fixation duration for Determiners (DT), Pronouns (PRP), singular noun possessives (NN+POS), and singular nouns (NN) appearing in the shared sentences. Error bars denote standard error.

which do not exist in these languages. Figure 4-2a presents the mean TF times for determiners in all five native languages, suggesting that native speakers of Portuguese and Spanish, which do have determiners, do not exhibit reduced reading times on this structure compared to natives. In ESL production, missing determiner errors are the most frequent error for native speakers of Japanese and third most common error for native speakers of Chinese.

In figure 4-2b we present the mean TF reading times for pronouns (PRP), where we also see shorter reading times by natives of Japanese and Chinese as compared to English natives. In both languages pronouns can be omitted both in object and subject positions. Portuguese and Spanish, in which pronoun omission is restricted to the subject position present a similar albeit weaker tendency.

In figure 4-2c we further observe that differently from natives of Chinese and Japanese, native speakers of Portuguese and Spanish spend more time on NN+POS in head final possessives such as "the *public's* confidence". While similar constructions exist in Chinese and Japanese, the NN+POS combination is expressed in Portuguese and Spanish as a head initial NN *of NN*. This form exists in English (e.g. "the confidence of the public") and is preferred by speakers of these languages in ESL writing [70]. As an additional baseline for this construction, we provide the TF times for NN in figure 4-2d. There, relative to English natives, we observe longer reading times for Japanese and Chinese and comparable times for Portuguese and Spanish.

The reading times of NN in figure 4-2d also give rise to a second, potentially competing interpretation of differences in ESL reading times, which highlights *lexical* rather than structural factors. According to this interpretation, increased reading times of nouns are the result of substantially smaller lexical sharing with English by Chinese and Japanese as compared to Spanish and Portuguese. Given the utilized speed normalization, lexical effects on nouns could in principle account for reduced reading times on determiners and pronouns. Conversely, structural influence leading to reduced reading times on determiners and pronouns could explain longer dwelling on nouns. A third possibility consistent with the observed reading patterns would allow for both structural and lexical effects to impact second language reading. Importantly, in each of these scenarios, ESL reading patterns are related to linguistic factors of the reader's native language.

We note that the presented analysis is preliminary in nature, and warrants further study in future research. In particular, reading times and classifier learned features may in some cases differ between the shared and the individual regimes. In the examples presented above, similar results are obtained in the individual sentences regime for DT, PRP and NN. The trend for the NN+POS construction, however, diminishes in that setup with similar reading times for all languages. On the other hand, one of the strongest features for predicting Portuguese and Spanish in the individual regime is longer reading times for prepositions (IN), an outcome that holds in the shared regime only relative to Chinese and Japanese, but not relative to native speakers of English.

Despite these caveats, our results suggest that reading patterns can potentially be re-

lated to linguistic factors of the reader's native language. This analysis can be extended in various ways, such as inclusion of additional feature types and fixation metrics, as well as utilization of other comparative methodologies. Combined with evidence from language production, this line of investigation can be instrumental for informing a linguistic theory of cross-linguistic influence.

4.5 Related Work

Eyetracking and second language reading Second language reading has been studied using eyetracking, with much of the work focusing on processing of syntactic ambiguities and analysis of specific target word classes such as cognates [30, 84]. In contrast to our work, such studies typically use controlled, rather than free-form sentences. Investigation of global metrics in free-form second language reading was introduced only recently by Cop et al. [20]. This study compared ESL and native reading of a novel by native speakers of Dutch, observing longer sentence reading times, more fixations and shorter saccades in ESL reading. Differently from this study, our work focuses on comparison of reading patterns between different native languages. We also analyze a related, but different metric, namely speed normalized fixation durations on word sequences.

Eyetracking for NLP tasks Recent work in NLP has demonstrated that reading gaze can serve as a valuable supervision signal for standard NLP tasks. Prominent examples of such work include POS tagging [5, 4], syntactic parsing [6] and sentence compression [46]. Our work also tackles a traditional NLP task with free-form text, but differs from this line of research in that it addresses this task only in comprehension. Furthermore, while these studies use gaze recordings of native readers, our work focuses on non-native readers.

NLI in production As previously mentioned NLI has been drawing considerable attention in NLP [48, 101]. NLI has also been driving much of the work on identification of native language related features in writing [105, 43, 13, 103, 97, 98, 59, 14]. Chapters 2 and 3, and [70, 67] have also linked usage patterns and grammatical errors in production to linguistic properties of the writer's native language. This chapter departs from NLI in writing and introduces NLI and related feature analysis in reading.

4.6 Conclusion

We present a novel framework for studying cross-linguistic influence in multilingualism by measuring gaze fixations during reading of free-form English text. We demonstrate for the first time that this signal can be used to determine a reader's native language. The effectiveness of linguistically motivated criteria for fixation clustering and our subsequent analysis suggest that the ESL reading process is affected by linguistic factors. Specifically, we show that linguistic similarities between native languages are reflected in similarities in ESL reading. We also identify several key features that characterize reading in different native languages, and discuss their potential connection to structural and lexical properties of the native langauge. The presented results demonstrate that eyetracking data can be instrumental for developing predictive and explanatory models of second language reading.

While this work is focused on NLIR from fixations, our general framework can be used to address additional aspects of reading, such as analysis of saccades and gaze trajectories. In future work, we also plan to explore the role of native and second language writing system characteristics in second language reading. More broadly, our methodology introduces parallels with production studies in NLP, creating new opportunities for integration of data, methodologies and tasks between production and comprehension. Furthermore, it holds promise for formulating language learning theory that is supported by empirical findings in naturalistic setups across language processing domains.

4.7 Supplemental Material

Eyetracking Setup We use a 44.4x33.3cm screen with 1024x768px resolution to present the reading materials, and a desktop mount Eyelink 1000 eyetracker (1000Hz) to record gaze. The screen, eyetracker camera and chinrest are horizontally aligned on a table surface. The screen center (x=512, y=384) is 79cm away from the center of the forehead bar, and 13cm below it. The eyetracker camera knob is 65cm away from forehead bar. Throughout the experiment participants hold a joystick with a button for indicating sentence completion, and two buttons for answering yes/no questions. We record gaze of the

participant's dominant eye.

Text Parameters All the textual material in the experiment is presented using Times font, normal style, with font size 23. In our setup, this corresponds to 0.36 degrees (11.3px) average lower case letter width, and 0.49 degrees (15.7px) average upper case letter width. We chose a non-monospace font, as such fonts are generally more common in reading. They also contain less between-letter horizontal spacing compared to monospace fonts, allowing fitting more text on the screen.

Calibration We use 3H line calibration with point repetition on the central horizontal line (y=384), with 16px outer circle and 6px inner circle fixation points. At least three calibrations are performed during the experiment, one at the beginning of each experimental section. We also recalibrate upon failure to produce a 300ms fixation on any fixation trigger preceding a sentence or a question within 4 seconds after its appearance. The mean validation error for calibrations across subjects is 0.146 degrees (std 0.038).

Chapter 5

Syntactic Annotation of Learner Language

5.1 Introduction

Despite the ubiquity of non-native English, as of 2015 there has been no publicly available syntactic treebank for ESL. To address this shortcoming, we developed and publicly released the Treebank of Learner English (TLE), a first of its kind resource for non-native English, containing 5,124 sentences manually annotated with POS tags and dependency trees. The TLE sentences are drawn from the FCE dataset [113], and authored by English learners from 10 different native language backgrounds. The treebank uses the Universal Dependencies (UD) formalism [23, 74], which provides a unified annotation framework across different languages and is geared towards multilingual NLP [65]. This characteristic allows our treebank to support computational analysis of ESL using not only English based but also multilingual approaches which seek to relate ESL phenomena to native language syntax.

While the annotation inventory and guidelines are defined by the English UD formalism, we build on previous work in learner language analysis [26, 28] to formulate an additional set of annotation conventions aiming at a uniform treatment of ungrammatical learner language. Our annotation scheme uses a two-layer analysis, whereby a distinct syntactic

annotation is provided for the *original* and the *corrected* version of each sentence. This approach is enabled by a pre-existing error annotation of the FCE [73] which is used to generate an error corrected variant of the dataset. Our inter-annotator agreement results provide evidence for the ability of the annotation scheme to support consistent annotation of ungrammatical structures.

Finally, a corpus that is annotated with both grammatical errors and syntactic dependencies paves the way for empirical investigation of the relation between grammaticality and syntax. Understanding this relation is vital for improving tagging and parsing performance on learner language [37], syntax based grammatical error correction [102, 72], and many other fundamental challenges in NLP. In this work, we take the first step in this direction by benchmarking tagging and parsing accuracy on our dataset under different training regimes, and obtaining several estimates for the impact of grammatical errors on these tasks.

To summarize, this chapter presents three contributions. First, we introduce the first large scale syntactic treebank for ESL, manually annotated with POS tags and universal dependencies. Second, we describe a linguistically motivated annotation scheme for ungrammatical learner English and provide empirical support for its consistency via inter-annotator agreement analysis. Third, we benchmark a state of the art parser on our dataset and estimate the influence of grammatical errors on the accuracy of automatic POS tagging and dependency parsing.

5.2 Treebank Overview

The TLE currently contains 5,124 sentences (97,681 tokens) with POS tag and dependency annotations in the English Universal Dependencies (UD) formalism [23, 74]. The sentences were obtained from the FCE corpus [113], a collection of upper intermediate English learner essays, containing error annotations with 75 error categories [73]. Sentence level segmentation was performed using an adaptation of the NLTK sentence tokenizer[1]. Under-segmented sentences were split further manually. Word level tokenization was gen-

[1] http://www.nltk.org/api/nltk.tokenize.html

erated using the Stanford PTB word tokenizer[2].

The treebank represents learners with 10 different native language backgrounds: Chinese, French, German, Italian, Japanese, Korean, Portuguese, Spanish, Russian and Turkish. For every native language, we randomly sampled 500 automatically segmented sentences, under the constraint that selected sentences have to contain at least one grammatical error that is not punctuation or spelling.

The TLE annotations are provided in two versions. The first version is the *original sentence* authored by the learner, containing grammatical errors. The second, *corrected sentence* version, is a grammatical variant of the original sentence, generated by correcting all the grammatical errors in the sentence according to the manual error annotation provided in the FCE dataset. The resulting corrected sentences constitute a parallel corpus of standard English. Table 5.1 presents basic statistics of both versions of the annotated sentences.

	original	corrected
sentences	5,124	5,124
tokens	97,681	98,976
sentence length	19.06 (std 9.47)	19.32 (std 9.59)
errors per sentence	2.67 (std 1.9)	-
authors	924	
native languages	10	

Table 5.1: Statistics of the TLE. Standard deviations are denoted in parenthesis.

To avoid potential annotation biases, the annotations of the treebank were created manually *from scratch*, without utilizing any automatic annotation tools. To further assure annotation quality, each annotated sentence was reviewed by two additional annotators. To the best of our knowledge, TLE is the first large scale English treebank constructed in a completely manual fashion.

[2]http://nlp.stanford.edu/software/tokenizer.shtml

5.3 Annotator Training

The treebank was annotated by six students, five undergraduates and one graduate. Among the undergraduates, three are linguistics majors and two are engineering majors with a linguistic minor. The graduate student is a linguist specializing in syntax. An additional graduate student in NLP participated in the final debugging of the dataset.

Prior to annotating the treebank sentences, the annotators were trained for about 8 weeks. During the training, the annotators attended tutorials on dependency grammars, and learned the English UD guidelines[3], the Penn Treebank POS guidelines [91], the grammatical error annotation scheme of the FCE [73], as well as the ESL guidelines described in section 5.5 and in the annotation manual.

Furthermore, the annotators completed six annotation exercises, in which they were required to annotate POS tags and dependencies for practice sentences from scratch. The exercises were done individually, and were followed by group meetings in which annotation disagreements were discussed and resolved. Each of the first three exercises consisted of 20 sentences from the UD gold standard for English, the English Web Treebank (EWT) [94]. The remaining three exercises contained 20-30 ESL sentences from the FCE. Many of the ESL guidelines were introduced or refined based on the disagreements in the ESL practice exercises and the subsequent group discussions. Several additional guidelines were introduced in the course of the annotation process.

During the training period, the annotators also learned to use a search tool that enables formulating queries over word and POS tag sequences as regular expressions and obtaining their annotation statistics in the EWT. After experimenting with both textual and graphical interfaces for performing the annotations, we converged on a simple text based format described in section 5.4.1, where the annotations were filled in using a spreadsheet or a text editor, and tested with a script for detecting annotation typos. The annotators continued to meet and discuss annotation issues on a weekly basis throughout the entire duration of the project.

[3] http://universaldependencies.org/#en

5.4 Annotation Procedure

The formation of the treebank was carried out in four steps: annotation, review, disagreement resolution and targeted debugging.

5.4.1 Annotation

In the first stage, the annotators were given sentences for annotation from scratch. We use a CoNLL based textual template in which each word is annotated in a separate line. Each line contains 6 columns, the first of which has the word index (IND) and the second the word itself (WORD). The remaining four columns had to be filled in with a Universal POS tag (UPOS), a Penn Treebank POS tag (POS), a head word index (HIND) and a dependency relation (REL) according to version 1 of the English UD guidelines.

The annotation section of the sentence is preceded by a metadata header. The first field in this header, denoted with SENT, contains the FCE error coded version of the sentence. The annotators were instructed to verify the error annotation, and add new error annotations if needed. Corrections to the sentence segmentation are specified in the SEGMENT field[4]. Further down, the field TYPO is designated for literal annotation of spelling errors and ill formed words that happen to form valid words (see section 5.5.2).

The example below presents a pre-annotated original sentence given to an annotator.

```
#SENT=That time I had to sleep in <ns type= "MD"><c>a</c></ns> tent.
#SEGMENT=
#TYPO=

#IND  WORD   UPOS POS  HIND REL
1     That
2     time
3     I
4     had
5     to
```

[4]The released version of the treebank splits the sentences according to the markings in the SEGMENT field when those apply both to the original and corrected versions of the sentence. Resulting segments without grammatical errors in the original version are currently discarded.

```
6       sleep
7       in
8       tent
9       .
```

Upon completion of the original sentence, the annotators proceeded to annotate the corrected sentence version. To reduce annotation time, annotators used a script that copies over annotations from the original sentence and updates head indices of tokens that appear in both sentence versions. Head indices and relation labels were filled in only if the head word of the token appeared in both the original and corrected sentence versions. Tokens with automatically filled annotations included an additional # sign in a seventh column of each word's annotation. The # signs had to be removed, and the corresponding annotations either approved or changed as appropriate. Tokens that did not appear in the original sentence version were annotated from scratch.

5.4.2 Review

All annotated sentences were randomly assigned to a second annotator (henceforth *reviewer*), in a double blind manner. The reviewer's task was to mark all the annotations that they would have annotated differently. To assist the review process, we compiled a list of common annotation errors, available in the released annotation manual.

The annotations were reviewed using an *active* editing scheme in which an explicit action was required for all the existing annotations. The scheme was introduced to prevent reviewers from overlooking annotation issues due to passive approval. Specifically, an additional # sign was added at the seventh column of each token's annotation. The reviewer then had to either "sign off" on the existing annotation by erasing the # sign, or provide an alternative annotation following the # sign.

5.4.3 Disagreement Resolution

In the final stage of the annotation process all annotator-reviewer disagreements were resolved by a third annotator (henceforth *judge*), whose main task was to decide in favor of

the annotator or the reviewer. Similarly to the review process, the judging task was carried out in a double blind manner. Judges were allowed to resolve annotator-reviewer disagreements with a third alternative, as well as introduce new corrections for annotation issues overlooked by the reviewers.

Another task performed by the judges was to mark acceptable *alternative annotations* for ambiguous structures determined through review disagreements or otherwise present in the sentence. These annotations were specified in an additional metadata field called AMBIGUITY. The ambiguity markings are provided along with the resolved version of the annotations.

5.4.4 Final Debugging

After applying the resolutions produced by the judges, we queried the corpus with debugging tests for specific linguistics constructions. This additional testing phase further reduced the number of annotation errors and inconsistencies in the treebank. Including the training period, the treebank creation lasted over a year, with an aggregate of more than 2,000 annotation hours.

5.5 Annotation Scheme for ESL

Our annotations use the existing inventory of English UD POS tags and dependency relations, and follow the standard UD annotation guidelines for English. However, these guidelines were formulated with grammatical usage of English in mind and do not cover non canonical syntactic structures arising due to grammatical errors[5]. To encourage consistent and linguistically motivated annotation of such structures, we formulated a complementary set of ESL annotation guidelines.

Our ESL annotation guidelines follow the general principle of *literal reading*, which emphasizes syntactic analysis according to the observed language usage. This strategy continues a line of work in SLA which advocates for centering analysis of learner language

[5]The English UD guidelines do address several issues encountered in informal genres, such as the relation "goeswith", which is used for fragmented words resulting from typos.

around morpho-syntactic surface evidence [82, 28]. Similarly to our framework, which includes a parallel annotation of corrected sentences, such strategies are often presented in the context of multi-layer annotation schemes that also account for error corrected sentence forms [42, 26, 85].

Deploying a strategy of literal annotation within UD, a formalism which enforces cross-linguistic consistency of annotations, will enable meaningful comparisons between non-canonical structures in English and canonical structures in the author's native language. As a result, a key novel characteristic of our treebank is its ability to support cross-lingual studies of learner language.

5.5.1 Literal Annotation

With respect to POS tagging, literal annotation implies adhering as much as possible to the observed morphological forms of the words. Syntactically, argument structure is annotated according to the usage of the word rather than its typical distribution in the relevant context. The following list of conventions defines the notion of literal reading for some of the common non canonical structures associated with grammatical errors.

Argument Structure

Extraneous prepositions We annotate all nominal dependents introduced by extraneous prepositions as nominal modifiers. In the following sentence, "him" is marked as a nominal modifier (*nmod*) instead of an indirect object (*iobj*) of "give".

```
#SENT=...I had to give <ns type="UT"><i>to</i> </ns> him water...

...
21   I       PRON  PRP  22   nsubj
22   had     VERB  VBD  5    parataxis
23   to      PART  TO   24   mark
24   give    VERB  VB   22   xcomp
25   to      ADP   IN   26   case
26   him     PRON  PRP  24   nmod
27   water   NOUN  NN   24   dobj
```

...

Omitted prepositions We treat nominal dependents of a predicate that are lacking a preposition as arguments rather than nominal modifiers. In the example below, “money” is marked as a direct object (*dobj*) instead of a nominal modifier (*nmod*) of “ask”. As “you” functions in this context as a second argument of “ask”, it is annotated as an indirect object (*iobj*) instead of a direct object (*dobj*).

#SENT=...I have to ask you <ns type="MT"> <c>for</c></ns> the money <ns type="RT"> of<c>for</c></ns> the tickets back.

...

12	I	PRON	PRP	13	nsubj
13	have	VERB	VBP	2	conj
14	to	PART	TO	15	mark
15	**ask**	VERB	VB	13	xcomp
16	**you**	PRON	PRP	15	**iobj**
17	the	DET	DT	18	det
18	**money**	NOUN	NN	15	**dobj**
19	of	ADP	IN	21	case
20	the	DET	DT	21	det
21	tickets	NOUN	NNS	18	nmod
22	back	ADV	RB	15	advmod
23	.	PUNCT	.	2	punct

Tense

Cases of erroneous tense usage are annotated according to the morphological tense of the verb. For example, below we annotate “shopping” with present participle VBG, while the correction “shop” is annotated in the corrected version of the sentence as VBP.

#SENT=...when you <ns type="TV">shopping <c>shop</c></ns>...

...

4	when	ADV	WRB	6	advmod
5	you	PRON	PRP	6	nsubj

```
6    shopping   VERB   VBG  12   advcl
...
```

Word Formation

Erroneous word formations that are contextually plausible and can be assigned with a PTB tag are annotated literally. In the following example, “stuffs” is handled as a plural count noun.

```
#SENT=...into fashionable <ns type="CN"> <i>stuffs</i><c>stuff</c></ns>...

...
7    into          ADP    IN    9    case
8    fashionable   ADJ    JJ    9    amod
9    stuffs        NOUN   NNS   2    ccomp
...
```

Similarly, in the example below we annotate “necessaryiest” as a superlative.

```
#SENT=The necessaryiest things...

1    The          DET    DT    3    det
2    necessaryiest ADJ   JJS   3    amod
3    things       NOUN   NNS   0    root
...
```

5.5.2 Exceptions to Literal Annotation

Although our general annotation strategy for ESL follows literal sentence readings, several types of word formation errors make such readings uninformative or impossible, essentially forcing certain words to be annotated using some degree of interpretation [86]. We hence annotate the following cases in the original sentence according to an interpretation of an intended word meaning, obtained from the FCE error correction.

Spelling

Spelling errors are annotated according to the correctly spelled version of the word. To support error analysis of automatic annotation tools, misspelled words that happen to form valid words are annotated in the metadata field TYPO for POS tags with respect to the most common usage of the misspelled word form. In the example below, the TYPO field contains the typical POS annotation of "where", which is clearly unintended in the context of the sentence.

```
#SENT=...we <ns type="SX"><i>where</i> <c>were</c></ns> invited to visit...
#TYPO=5 ADV WRB

...
4    we        PRON  PRP   6    nsubjpass
5    where     AUX   VBD   6    auxpass
6    invited   VERB  VBN   0    root
7    to        PART  TO    8    mark
8    visit     VERB  VB    6    xcomp
...
```

Word Formation

Erroneous word formations that cannot be assigned with an existing PTB tag are annotated with respect to the correct word form.

```
#SENT=I am <ns type="IV"><i>writting</i> <c>writing</c></ns>...

1    I         PRON  PRP   3    nsubj
2    am        AUX   VBP   3    aux
3    writting  VERB  VBG   0    root
...
```

In particular, ill formed adjectives that have a plural suffix receive a standard adjectival POS tag. When applicable, such cases also receive an additional marking for unnecessary agreement in the error annotation using the attribute "ua".

```
#SENT=...<ns type="IJ" ua=true> <i>interestings</i><c>interesting</c></ns>
things...

...
6    interestings    ADJ   JJ    7    amod
7    things          NOUN  NNS   3 dobj
...
```

Wrong word formations that result in a valid, but contextually implausible word form are also annotated according to the word correction. In the example below, the nominal form "sale" is likely to be an unintended result of an ill formed verb. Similarly to spelling errors that result in valid words, we mark the typical literal POS annotation in the TYPO metadata field.

```
#SENT=...they do not <ns type="DV"><i>sale</i> <c>sell</c></ns> them...
#TYPO=15 NOUN NN

...
12   they      PRON  PRP   15   nsubj
13   do        AUX   VBP   15   aux
14   not       PART  RB    15   neg
15   sale      VERB  VB    0    root
16   them      PRON  PRP   15   dobj
...
```

Taken together, our ESL conventions cover many of the annotation challenges related to grammatical errors present in the TLE.

5.6 Editing Agreement

We utilize our two step review process to estimate agreement rates between annotators[6]. We measure agreement as the fraction of annotation tokens approved by the editor. Table 5.2 presents the agreement between annotators and reviewers, as well as the agreement between

[6]All experimental results on agreement and parsing exclude punctuation tokens.

reviewers and the judges. Cohen's Kappa scores [18] for POS tags and dependency labels in all evaluation conditions are above 0.96. Agreement measurements are provided for both the original the corrected versions of the dataset.

Annotator-Reviewer	UPOS	POS	HIND	REL
original	98.83	98.35	97.74	96.98
corrected	99.02	98.61	97.97	97.20
Reviewer-Judge				
original	99.72	99.68	99.37	99.15
corrected	99.80	99.77	99.45	99.28

Table 5.2: Inter-annotator agreement on the entire TLE corpus. Agreement is measured as the fraction of tokens that remain unchanged after an editing round. The four evaluation columns correspond to universal POS tags, PTB POS tags, unlabeled attachment, and dependency labels.

Overall, the results indicate a high agreement rate in the two editing tasks. Importantly, the gap between the agreement on the original and corrected sentences is small. Note that this result is obtained despite the introduction of several ESL annotation guidelines in the course of the annotation process, which inevitably increased the number of edits related to grammatical errors. We interpret this outcome as evidence for the effectiveness of the ESL annotation scheme in supporting consistent annotations of learner language.

5.7 Parsing Experiments

The TLE enables studying parsing for learner language and exploring relationships between grammatical errors and parsing performance. Here, we present parsing benchmarks on our dataset, and provide several estimates for the extent to which grammatical errors degrade the quality of automatic POS tagging and dependency parsing.

Our first experiment measures tagging and parsing accuracy on the TLE and approximates the global impact of grammatical errors on automatic annotation via performance comparison between the original and error corrected sentence versions. In this, and subsequent experiments, we utilize version 2.2 of the Turbo tagger and Turbo parser [63], state of the art tools for statistical POS tagging and dependency parsing.

Table 5.3 presents tagging and parsing results on a test set of 500 TLE sentences (9,591 original tokens, 9,700 corrected tokens). Results are provided for three different training regimes. The first regime uses the training portion of version 1.3 of the EWT, the UD English treebank, containing 12,543 sentences (204,586 tokens). The second training mode uses 4,124 training sentences (78,541 original tokens, 79,581 corrected tokens) from the TLE corpus. In the third setup we combine these two training corpora. The remaining 500 TLE sentences (9,549 original tokens, 9,695 corrected tokens) are allocated to a development set, not used in this experiment. Parsing of the test sentences was performed on predicted POS tags.

Test set	**Train Set**	**UPOS**	**POS**	**UAS**	**LA**	**LAS**
TLE_{orig}	EWT	91.87	94.28	86.51	88.07	81.44
TLE_{corr}	EWT	92.9	95.17	88.37	89.74	83.8
TLE_{orig}	TLE_{orig}	95.88	94.94	87.71	89.26	83.4
TLE_{corr}	TLE_{corr}	96.92	95.17	89.69	90.92	85.64
TLE_{orig}	EWT+TLE_{orig}	93.33	95.77	90.3	91.09	86.27
TLE_{corr}	EWT+TLE_{corr}	94.27	96.48	92.15	92.54	88.3

Table 5.3: Tagging and parsing results on a test set of 500 sentences from the TLE corpus. EWT is the English UD treebank. TLE_{orig} are original sentences from the TLE. TLE_{corr} are the corresponding error corrected sentences.

The EWT training regime, which uses out of domain texts written in standard English, provides the lowest performance on all the evaluation metrics. An additional factor which negatively affects performance in this regime are systematic differences in the EWT annotation of possessive pronouns, expletives and names compared to the UD guidelines, which are utilized in the TLE. In particular, the EWT annotates possessive pronoun UPOS as PRON rather than DET, which leads the UPOS results in this setup to be lower than the PTB POS results. Improved results are obtained using the TLE training data, which, despite its smaller size, is closer in genre and syntactic characteristics to the TLE test set. The strongest PTB POS tagging and parsing results are obtained by combining the EWT with the TLE training data, yielding 95.77 POS accuracy and a UAS of 90.3 on the original version of the TLE test set.

The dual annotation of sentences in their original and error corrected forms enables

estimating the impact of grammatical errors on tagging and parsing by examining the performance gaps between the two sentence versions. Averaged across the three training conditions, the POS tagging accuracy on the original sentences is lower than the accuracy on the sentence corrections by 1.0 UPOS and 0.61 POS. Parsing performance degrades by 1.9 UAS, 1.59 LA and 2.21 LAS.

To further elucidate the influence of grammatical errors on parsing quality, table 5.4 compares performance on tokens in the original sentences appearing inside grammatical error tags to those appearing outside such tags. Although grammatical errors may lead to tagging and parsing errors with respect to any element in the sentence, we expect erroneous tokens to be more challenging to analyze compared to grammatical tokens.

Tokens	**Train Set**	**UPOS**	**POS**	**UAS**	**LA**	**LAS**
Ungrammatical	EWT	87.97	88.61	82.66	82.66	74.93
Grammatical	EWT	92.62	95.37	87.26	89.11	82.7
Ungrammatical	TLE$_{orig}$	90.76	88.68	83.81	83.31	77.22
Grammatical	TLE$_{orig}$	96.86	96.14	88.46	90.41	84.59
Ungrammatical	EWT+TLE$_{orig}$	89.76	90.97	86.32	85.96	80.37
Grammatical	EWT+TLE$_{orig}$	94.02	96.7	91.07	92.08	87.41

Table 5.4: Tagging and parsing results on the original version of the TLE test set for tokens marked with grammatical errors (Ungrammatical) and tokens not marked for errors (Grammatical).

This comparison indeed reveals a substantial difference between the two types of tokens, with an average gap of 5.0 UPOS, 6.65 POS, 4.67 UAS, 6.56 LA and 7.39 LAS. Note that differently from the global measurements in the first experiment, this analysis, which focuses on the local impact of remove/replace errors, suggests a stronger effect of grammatical errors on the dependency labels than on the dependency structure.

Finally, we measure tagging and parsing performance relative to the fraction of sentence tokens marked with grammatical errors. Similarly to the previous experiment, this analysis focuses on remove/replace rather than insert errors.

Figure 5-1 presents the average sentential performance as a function of the percentage of tokens in the original sentence marked with grammatical errors. In this experiment, we train the parser on the EWT training set and test on the entire TLE corpus. Performance

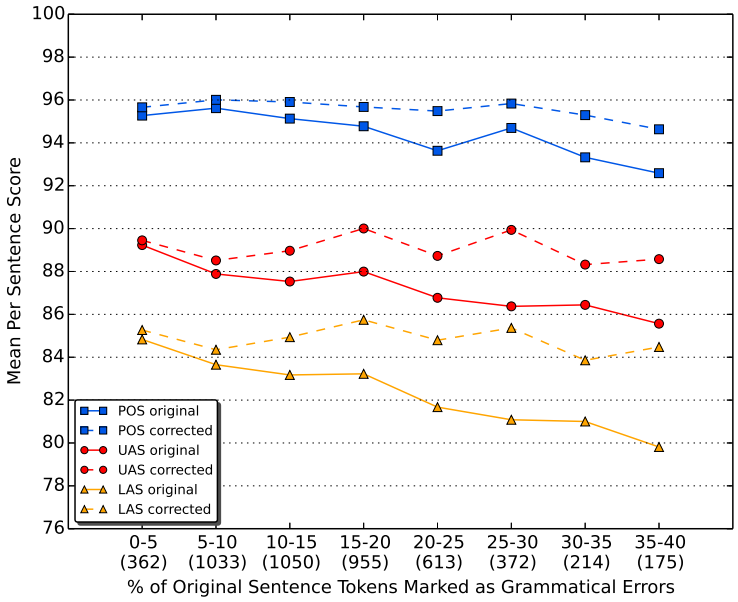


Figure 5-1: Mean per sentence POS accuracy, UAS and LAS of the Turbo tagger and Turbo parser, as a function of the percentage of original sentence tokens marked with grammatical errors. The tagger and the parser are trained on the EWT corpus, and tested on all 5,124 sentences of the TLE. Points connected by continuous lines denote performance on the original TLE sentences. Points connected by dashed lines denote performance on the corresponding error corrected sentences. The number of sentences whose errors fall within each percentage range appears in parenthesis.

curves are presented for POS, UAS and LAS on the original and error corrected versions of the annotations. We observe that while the performance on the corrected sentences is close to constant, original sentence performance is decreasing as the percentage of the erroneous tokens in the sentence grows.

Overall, our results suggest a negative, albeit limited effect of grammatical errors on parsing. This outcome contrasts a study by Geertzen et al. [37] which reported a larger performance gap of 7.6 UAS and 8.8 LAS between sentences with and without grammatical errors. We believe that our analysis provides a more accurate estimate of this impact, as it controls for both sentence content and sentence length. The latter factor is crucial, since it correlates positively with the number of grammatical errors in the sentence, and negatively with parsing accuracy.

5.8 Related Work

Previous studies on learner language proposed several annotation schemes for both POS tags and syntax [42, 26, 28, 85]. The unifying theme in these proposals is a multi-layered analysis aiming to decouple the observed language usage from conventional structures in the foreign language.

In the context of ESL, Dıaz et al. [26] propose three parallel POS tag annotations for the *lexical*, *morphological* and *distributional* forms of each word. In our work, we adopt the distinction between morphological word forms, which roughly correspond to our literal word readings, and distributional forms as the error corrected words. However, we account for morphological forms only when these constitute valid existing PTB POS tags and are contextually plausible. Furthermore, while the internal structure of invalid word forms is an interesting object of investigation, we believe that it is more suitable for annotation as word features rather than POS tags. Our treebank supports the addition of such features to the existing annotations.

The work of Ragheb and Dickinson [27, 82, 28] proposes ESL annotation guidelines for POS tags and syntactic dependencies based on the CHILDES annotation framework. This approach, called "morphosyntactic dependencies" is related to our annotation scheme in its focus on surface structures. Differently from this proposal, our annotations are grounded in a parallel annotation of grammatical errors and include an additional layer of analysis for the corrected forms. Moreover, we refrain from introducing new syntactic categories and dependency relations specific to ESL, thereby supporting computational treatment of ESL using existing resources for standard English. At the same time, we utilize a multilingual formalism which, in conjunction with our literal annotation strategy, facilitates linking the annotations to native language syntax.

While the above mentioned studies focus on annotation guidelines, attention has also been drawn to the topic of parsing in the learner language domain. However, due to the shortage of syntactic resources for ESL, much of the work in this area resorted to using surrogates for learner data. For example, in Foster [32] and Foster et al. [33] parsing experiments are carried out on synthetic learner-like data, that was created by automatic

insertion of grammatical errors to well formed English text. In Cahill et al. [16] a treebank of secondary level native students texts was used to approximate learner text in order to evaluate a parser that utilizes unlabeled learner data.

Syntactic annotations for ESL were previously developed by Nagata et al. [69], who annotate an English learner corpus with POS tags and shallow syntactic parses. Recently, this dataset was extended to additional sentences from the ICLE [40] and annotations of phrase structure trees [68] which also contain information about grammatical errors. Differently from this annotation effort, our treebank decouples annotations of syntax and grammatical errors, uses dependency trees, and covers a wide range of learner native languages. An additional syntactic dataset for ESL, currently not available publicly, are 1,000 sentences from the EFCamDat dataset [37], annotated with Stanford dependencies [25]. This dataset was used to measure the impact of grammatical errors on parsing by comparing performance on sentences with grammatical errors to error free sentences. The TLE enables a more direct way of estimating the magnitude of this performance gap by comparing performance on the same sentences in their original and error corrected versions. Our comparison suggests that the effect of grammatical errors on parsing is smaller than the one reported in this study.

5.9 Public Release

Official releases of the TLE are publicly available through the Universal Dependences repository `universaldependencies.org`. We also developed a web GUI which enables querying the treebank online, available at `esltreebank.org`. Figure 5-2 presents a screenshot of the interface and an example query. The complete manual of ESL guidelines used by the annotators is publicly available at `http://people.csail.mit.edu/berzak/tle_guidelines/guidelines.pdf`. The manual contains further details on our annotation scheme, additional annotation guidelines and a list of common annotation errors.

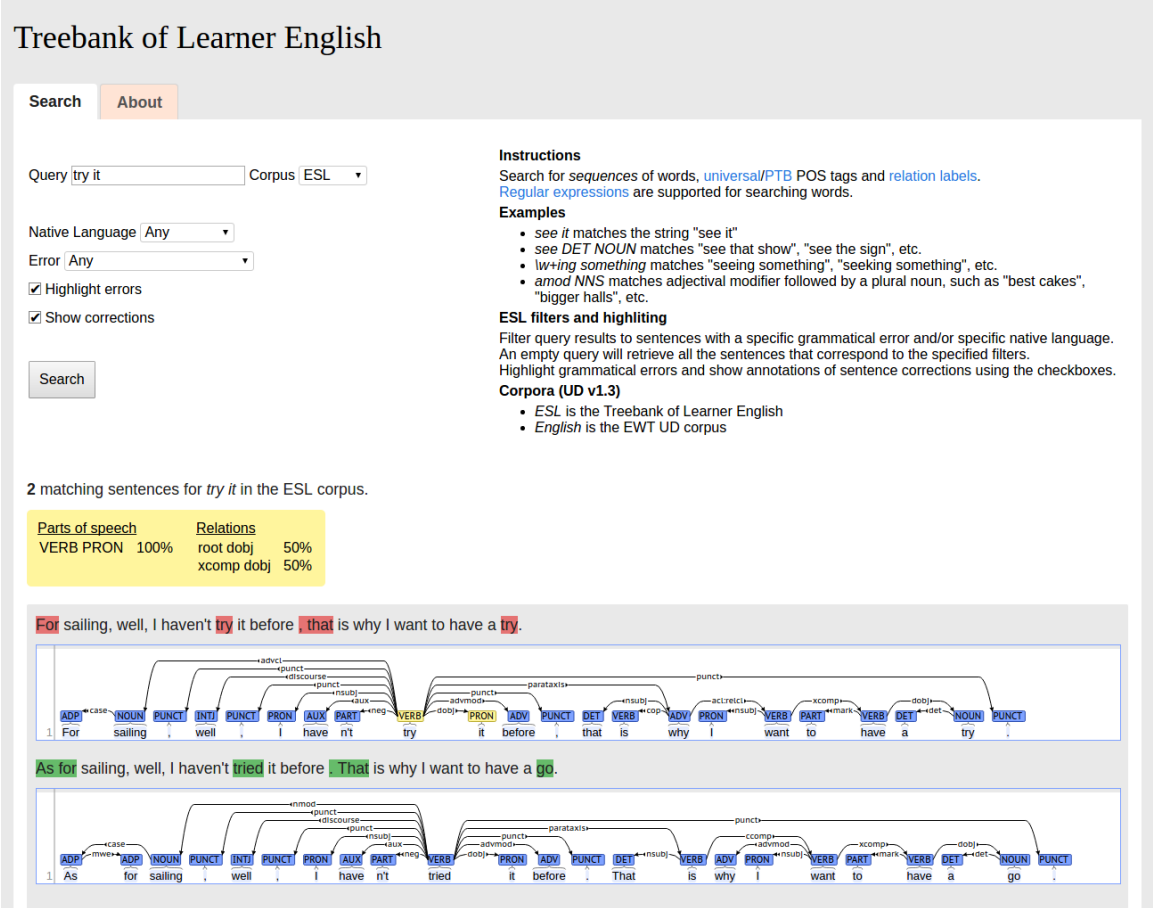


Figure 5-2: The web query engine of the TLE

5.10 Conclusion

We present the first large scale treebank of learner language, manually annotated and double-reviewed for POS tags and universal dependencies. The annotation is accompanied by a linguistically motivated framework for handling syntactic structures associated with grammatical errors. Finally, we benchmark automatic tagging and parsing on our corpus, and measure the effect of grammatical errors on tagging and parsing quality. The treebank will support empirical study of learner syntax in NLP, corpus linguistics and second language acquisition.

Chapter 6

Conclusion

We presented a body of work which takes an interdisciplinary computationally driven approach to the study of second language learning and multilingualism, and constitutes a radical departure from previous work on these topics in linguistics and NLP. In particular, we provided a novel conceptual and experimental framework that ties performance and behavioral psycholinguistic signal in second language processing with the linguistic characterization of languages of the world. Our experiments provide empirical evidence for the systematicity of cross-linguistic transfer, and demonstrate the ability to make both global and fine grained inference between typological properties of the first language on the one hand, and aspects of linguistic processing in a second language on the other.

This thesis also introduces an integrative approach for the study of multilingualism by combining language production and language comprehension in one framework. Our approach examines linguistic performance on various scales of granularity across reading and writing, and supports comparison of results for similar underlying linguistic phenomena in both tasks. The presented framework lays empirical and methodological foundations for developing cognitive and linguistic theories focusing on fundamental learning mechanisms and their manifestations across different language processing tasks. To support future research within this framework, we introduce the TLE, a novel dataset geared towards multilingual research of learner language production, and in the future will also publicly release our eyetracking data.

Our investigation highlights the importance of cross-linguistic influence in second language acquisition, and demonstrates the promise of interdisciplinary and cross-task approaches for studying core principles of multilingualism. The empirical findings, tools and data introduced in this thesis will be instrumental for computational and linguistic research on learner language. They will also support the development of novel NLP technologies which will address the unique characteristics of learner language.

6.1 Future Work

The work presented in this thesis opens multiple avenues for future research. In what follows we outline two such directions.

- **Scaling and Linguistic Diversity** Our current investigation provides proofs of concept on limited sets of native languages (14 in chapters 2 and 3, 10 in chapter 5 and 4 in chapter 4). A natural question that arises is the extent to which our framework will scale with the introduction of additional languages. Larger language diversity may also reveal additional linguistic phenomena that are important in language learning but hard to detect with a restricted set of languages. Moreover, in all cases we used English as the second language, which limits our ability to investigate the role of second language properties in language learning. Integration of new datasets for other second languages will be instrumental for generalizing and extending our findings on ESL.

- **Generative Models of Learner Language** This thesis uses discriminative learning approaches, making it challenging to provide a full predictive account of production and comprehension in a second language. For example, while we can identify and analyze reading times on important linguistic constructions via feature analysis, we currently do not have a computational model that will be able to predict the entire reading trajectory of an ESL learner. Developing such models will require integrating our results within a generative computational framework which will combine both native language specific and learner general aspects of language acquisition and

processing. This research direction will further improve our understanding of language learning and multilingualism.

Appendix A

Methodology of Syntactic Annotation

A.1 Introduction

Research in NLP relies heavily on the availability of human annotations for various linguistic prediction tasks. Such resources are commonly treated as de facto gold standards and are used for both training and evaluation of algorithms for automatic annotation. At the same time, human agreement on these annotations provides an indicator for the difficulty of the task, and can be instrumental for estimating upper limits for the performance obtainable by computational methods.

Linguistic gold standards are often constructed using pre-existing annotations, generated by automatic tools. The output of such tools is then manually corrected by human annotators to produce the gold standard. The justification for this annotation methodology was first introduced in a set of experiments on POS tag annotation conducted as part of the Penn Treebank project [62]. In this study, the authors concluded that tagger-based annotations are not only much faster to obtain, but also more consistent and of higher quality compared to annotations from scratch. Following the Penn Treebank, syntactic annotation projects for various languages, including German [12], French [1], Arabic [58] and many others, were annotated using automatic tools as a starting point. Despite the widespread use of this annotation pipeline, there is, to our knowledge, little prior work on syntactic annotation quality and on the reliability of system evaluations on such data.

In this appendix, we present a systematic study of the influence of automatic tool output on characteristics of annotations created for NLP purposes. Our investigation is motivated by the hypothesis that annotations obtained using such methodologies may be subject to the problem of *anchoring*, a well established and robust cognitive bias in which human decisions are affected by pre-existing values [106]. In the presence of anchors, participants reason relative to the existing values, and as a result may provide different solutions from those they would have reported otherwise. Most commonly, anchoring is manifested as an alignment *towards* the given values.

Focusing on the key NLP tasks of POS tagging and dependency parsing, we demonstrate that the standard approach of obtaining annotations via human correction of automatically generated POS tags and dependencies exhibits a clear anchoring effect – a phenomenon we refer to as *parser bias*. Given this evidence, we examine two potential adverse implications of this effect on parser-based gold standards.

First, we show that parser bias entails substantial overestimation of parser performance. In particular, we demonstrate that bias towards the output of a specific tagger-parser pair leads to over-estimation of the performance of these tools relative to other tools. Moreover, we observe general performance gains for automatic tools relative to their performance on human-based gold standards. Second, we study whether parser bias affects the quality of the resulting gold standards. Extending the experimental setup of Marcus et al. [62], we demonstrate that parser bias may lead to *lower* annotation quality for parser-based annotations compared to human-based annotations.

Furthermore, we conduct an experiment on inter-annotator agreement for POS tagging and dependency parsing which controls for parser bias. Our experiment on a subset of section 23 of the WSJ Penn Treebank yields agreement rates of 95.65 for POS tagging and 94.17 for dependency parsing. This result is significant in light of the state of the art tagging and parsing performance for English newswire. With parsing reaching the level of human agreement, and tagging surpassing it, a more thorough examination of evaluation resources and evaluation methodologies for these tasks is called for.

To summarize, we present the first study to measure and analyze anchoring in the standard parser-based approach to creation of gold standards for POS tagging and dependency

parsing in NLP. We conclude that gold standard annotations that are based on editing output of automatic tools can lead to inaccurate figures in system evaluations and lower annotation quality. Our human agreement experiment, which controls for parser bias, yields agreement rates that are comparable to state of the art automatic tagging and dependency parsing performance, highlighting the need for a more extensive investigation of tagger and parser evaluation in NLP.

A.2 Experimental Setup

A.2.1 Annotation Tasks

We examine two standard annotation tasks in NLP, POS tagging and dependency parsing. In the POS tagging task, each word in a sentence has to be categorized with a Penn Treebank POS tag [91] (henceforth POS). The dependency parsing task consists of providing a sentence with a labeled dependency tree using the Universal Dependencies (UD) formalism [23], according to version 1 of the UD English guidelines[1]. To perform this task, the annotator is required to specify the head word index (henceforth HIND) and relation label (henceforth REL) of each word in the sentence.

We distinguish between three variants of these tasks, *annotation*, *reviewing* and *ranking*. In the annotation variant, participants are asked to conduct annotation from scratch. In the reviewing variant, they are asked to provide alternative annotations for all annotation tokens with which they disagree. The participants are not informed about the source of the given annotation, which, depending on the experimental condition can be either parser output or human annotation. In the ranking task, the participants rank several annotation options with respect to their quality. Similarly to the review task, the participants are not given the sources of the different annotation options. Participants performing the annotation, reviewing and ranking tasks are referred to as annotators, reviewers and judges, respectively.

[1] http://universaldependencies.org/#en

A.2.2 Annotation Format

All annotation tasks are performed using a CoNLL style text-based template, also used in chapter 5, in which each word appears in a separate line. The first two columns of each line contain the word index and the word, respectively. The next three columns are designated for annotation of POS, HIND and REL.

In the annotation task, these values have to be specified by the annotator from scratch. In the review task, participants are required to edit pre-annotated values for a given sentence. The sixth column in the review template contains an additional # sign, whose goal is to prevent reviewers from overlooking and passively approving existing annotations. Corrections are specified following this sign in a space separated format, where each of the existing three annotation tokens is either corrected with an alternative annotation value or approved using a * sign. Approval of all three annotation tokens is marked by removing the # sign. The example below presents a fragment from a sentence used for the reviewing task, in which the reviewer approves the annotations of all the words, with the exception of "help", where the POS is corrected from VB to NN and the relation label *xcomp* is replaced with *dobj*.

```
...
5   you       PRP   6   nsubj
6   need      VBP   3   ccomp
7   help      VB    6   xcomp  # NN * dobj
...
```

The format of the ranking task is exemplified below. The annotation options are presented to the participants in a random order. Participants specify the rank of each annotation token following the vertical bar. In this sentence, the label *cop* is preferred over *aux* for the word "be" and *xcomp* is preferred over *advcl* for the word "Common".

```
...
8   it         PRP   10   nsubjpass
```

```
9   is          VBZ   10   auxpass
10  planed      VBN   0    root
11  to          TO    15   mark
12  be          VB    15   aux-cop | 2-1
13  in          IN    15   case
14  Wimbledon   NNP   15   compound
15  Common      NNP   10   advcl-xcomp | 2-1
...
```

The participants used basic validation scripts which checked for typos and proper formatting of the annotations, reviews and rankings.

A.2.3 Evaluation Metrics

We measure both parsing performance and inter-annotator agreement using tagging and parsing evaluation metrics. This choice allows for a direct comparison between parsing and agreement results. In this context, POS refers to tagging accuracy. We utilize the standard metrics Unlabeled Attachment Score (UAS) and Label Accuracy (LA) to measure accuracy of head attachment and dependency labels. We also utilize the standard parsing metric Labeled Attachment Score (LAS), which takes into account both dependency arcs and dependency labels. In all our parsing and agreement experiments, we exclude punctuation tokens from the evaluation.

A.2.4 Corpora

We use sentences from two publicly available datasets, covering two different genres. The first corpus, used in the experiments in sections A.3 and A.4, is the First Certificate in English (FCE) Cambridge Learner Corpus [113]. This dataset contains essays authored by upper-intermediate level English learners[2].

The second corpus is the WSJ part of the Penn Treebank (WSJ PTB) [62]. Since its

[2]The annotation bias and quality results reported in sections A.3 and A.4 use the original learner sentences, which contain grammatical errors. These results were replicated on the error corrected versions of the sentences.

release, this dataset has been the most commonly used resource for training and evaluation of English parsers. Our experiment on inter-annotator agreement in section A.5 uses a random subset of the sentences in section 23 of the WSJ PTB, which is traditionally reserved for tagging and parsing evaluation.

A.2.5 Annotators

For this experiment we recruited five of students who annotated the TLE as annotators. Three of the students are linguistics majors and two are engineering majors with linguistics minors. With respect to our experiments, the extensive experience of our annotators and their prior work as a group strengthen our results, as these characteristics reduce the effect of anchoring biases and increase inter-annotator agreement.

A.3 Parser Bias

Our first experiment is designed to test whether expert human annotators are biased towards POS tags and dependencies generated by automatic tools. We examine the common out-of-domain annotation scenario, where automatic tools are often trained on an existing treebank in one domain, and used to generate initial annotations to speed-up the creation of a gold standard for a new domain. We use the EWT UD corpus as the existing gold standard, and a sample of the FCE dataset as the new corpus.

Procedure

Our experimental procedure, illustrated in figure A-1(a) contains a set of 360 sentences (6,979 tokens) from the FCE, for which we generate three gold standards: one based on human annotations and two based on parser outputs. To this end, for each sentence, we assign *at random* four of the participants to the following annotation and review tasks. The fifth participant is left out to perform the quality ranking task described in section A.4.

The first participant annotates the sentence from scratch, and a second participant reviews this annotation. The overall agreement of the reviewers with the annotators is 98.24 POS, 97.16 UAS, 96.3 LA and 94.81 LAS. The next two participants review parser outputs.

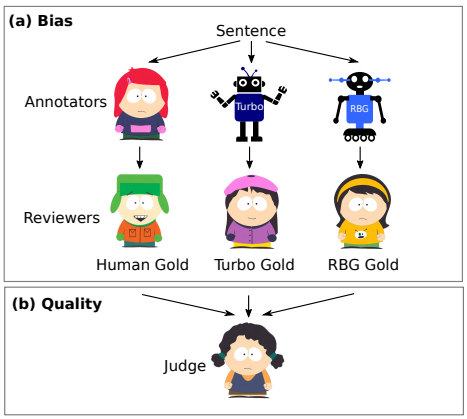


Figure A-1: Experimental setup for parser bias (a) and annotation quality (b) on 360 sentences (6,979 tokens) from the FCE. For each sentence, five human annotators are assigned at random to one of three roles: annotation, review or quality assessment. In the bias experiment, presented in section A.3, every sentence is annotated by a human, Turbo parser (based on Turbo tagger output) and RBG parser (based on Stanford tagger output). Each annotation is reviewed by a different human participant to produce three gold standards of each sentence: "Human Gold", "Turbo Gold" and "RBG Gold". The fifth annotator performs a quality assessment task described in section A.4, which requires to rank the three gold standards in cases of disagreement.

One participant reviews an annotation generated by the Turbo tagger and parser [63]. The other participant reviews the output of the Stanford tagger [104] and RBG parser [52]. The taggers and parsers were trained on the gold annotations of the EWT UD treebank, version 1.1. Both parsers use predicted POS tags for the FCE sentences.

Assigning the reviews to the human annotations yields a human based gold standard for each sentence called "Human Gold". Assigning the reviews to the tagger and parser outputs yields two parser-based gold standards, "Turbo Gold" and "RBG Gold". We chose the Turbo-Turbo and Stanford-RBG tagger-parser pairs as these tools obtain comparable performance on standard evaluation benchmarks, while yielding substantially different annotations due to different training algorithms and feature sets. For our sentences, the agreement between the Turbo tagger and Stanford tagger is 96.97 POS. The agreement between the Turbo parser and RBG parser based on the respective tagger outputs is 90.76 UAS, 91.6 LA and 87.34 LAS.

	Turbo				RBG			
	POS	UAS	LA	LAS	POS	UAS	LA	LAS
Human Gold	95.32	87.29	88.35	82.29	95.59	87.19	88.03	82.05
Turbo Gold	**97.62**	**91.86**	**92.54**	**89.16**	*96.64*	*89.16*	*89.75*	*84.86*
Error Reduction %	49.15	35.96	35.97	38.79	23.81	15.38	14.37	15.65
RBG Gold	96.43	*88.65*	*89.95*	*84.42*	**97.76**	**91.22**	**91.84**	**87.87**
Error Reduction %	23.72	10.7	13.73	12.03	49.21	31.46	31.83	32.42

Table A.1: Annotator bias towards taggers and parsers on 360 sentences (6,979 tokens) from the FCE. Tagging and parsing results are reported for the Turbo parser (based on the output of the turbo Tagger) and RBG parser (based on the output of the Stanford tagger) on three gold standards. Human Gold are manual corrections of human annotations. Turbo Gold are manual corrections of the output of Turbo tagger and Turbo parser. RBG Gold are manual corrections of the Stanford tagger and RBG parser. Error reduction rates are reported relative to the results obtained by the two tagger-parser pairs on the Human Gold annotations. Note that (1) The parsers perform equally well on Human Gold. (2) Each parser performs better than the other parser on its own reviews. (3) Each parser performs better on the reviews of the other parser compared to its performance on Human Gold. The differences in (2) and (3) are statistically significant with $p \ll 0.001$ using McNemar's test.

Parser Specific and Parser Shared Bias

In order to test for parser bias, in table A.1 we compare the performance of the Turbo-Turbo and Stanford-RBG tagger-parser pairs on our three gold standards. First, we observe that while these tools perform equally well on Human Gold, each tagger-parser pair performs better than the other on its own reviews. These *parser specific* performance gaps are substantial, with an average of 1.15 POS, 2.63 UAS, 2.34 LA and 3.88 LAS between the two conditions. This result suggests the presence of a bias towards the output of specific tagger-parser combinations. The practical implication of this outcome is that a gold standard created by editing an output of a parser is likely to boost the performance of that parser in evaluations and over-estimate its performance relative to other parsers.

Second, we note that the performance of each of the parsers on the gold standard of the other parser is still higher than its performance on the human gold standard. The average performance gap between these conditions is 1.08 POS, 1.66 UAS, 1.66 LA and 2.47 LAS. This difference suggests an annotation bias towards *shared* aspects in the predictions of taggers and parsers, which differ from the human based annotations. The consequence of this observation is that irrespective of the specific tool that was used to pre-annotate the

data, parser-based gold standards are likely to result in higher parsing performance relative to human-based gold standards.

Taken together, the parser specific and parser shared effects lead to a dramatic overall average error reduction of 49.18% POS, 33.71% UAS, 34.9% LA and 35.61% LAS on the parser-based gold standards compared to the human-based gold standard. To the best of our knowledge, these results are the first systematic demonstration of the tendency of the common approach of parser-based creation of gold standards to yield biased annotations and lead to overestimation of tagging and parsing performance.

A.4 Annotation Quality

In this section we extend our investigation to examine the impact of parser bias on the quality of parser-based gold standards. To this end, we perform a manual comparison between human-based and parser-based gold standards.

Our quality assessment experiment, depicted schematically in figure A-1(b), is a ranking task. For each sentence, a randomly chosen judge, who did not annotate or review the given sentence, ranks disagreements between the three gold standards Human Gold, Turbo Gold and RBG Gold, generated in the parser bias experiment in section A.3.

Human Gold Preference %	POS	HIND	REL
Turbo Gold	64.32*	63.96*	61.5*
# disagreements	199	444	439
RBG Gold	56.72	61.38*	57.73*
# disagreements	201	435	440

Table A.2: Human preference rates for a human-based gold standard Human Gold over the two parser-based gold standards Turbo Gold and RBG Gold. # disagreements denotes the number of tokens that differ between Human Gold and the respective parser-based gold standard. Statistically significant values for a two-tailed Z test with $p < 0.01$ are marked with *. Note that for both tagger-parser pairs, human judges tend to prefer human-based over parser-based annotations.

Table A.2 presents the preference rates of judges for the human-based gold standard over each of the two parser-based gold standards. In all three evaluation categories, human judges tend to prefer the human-based gold standard over both parser-based gold standards.

This result demonstrates that the initial reduced quality of the parser outputs compared to human annotations indeed percolates via anchoring to the resulting gold standards.

The analysis of the quality assessment experiment thus far did not distinguish between cases where the two parsers agree and where they disagree. In order to gain further insight into the relation between parser bias and annotation quality, we break down the results reported in table A.2 into two cases which relate directly to the *parser specific* and *parser shared* components of the tagging and parsing performance gaps observed in the parser bias results reported in section A.3. In the first case, called "parser specific approval", a reviewer approves a parser annotation which disagrees both with the output of the other parser and the Human Gold annotation. In the second case, called "parser shared approval", a reviewer approves a parser output which is shared by both parsers but differs with respect to Human Gold.

Human Gold Preference %	POS	HIND	REL
Turbo specific approval	85.42*	78.69*	80.73*
# disagreements	48	122	109
RBG specific approval	73.81*	77.98*	77.78*
# disagreements	42	109	108
Parser shared approval	51.85	58.49*	51.57
# disagreements	243	424	415

Table A.3: Breakdown of the Human preference rates for the human-based gold standard over the parser-based gold standards in table A.2, into cases of agreement and disagreement between the two parsers. Parser specific approval are cases in which a parser output approved by the reviewer differs from both the output of the other parser and the Human Gold annotation. Parser shared approval denotes cases where an approved parser output is identical to the output of the other parser but differs from the Human Gold annotation. Statistically significant values for a two-tailed Z test with $p < 0.01$ are marked with *. Note that parser specific approval is substantially more detrimental to the resulting annotation quality compared to parser shared approval.

Table A.3 presents the judge preference rates for the Human-Gold annotations in these two scenarios. We observe that cases in which the parsers disagree are of substantially worse quality compared to human-based annotations. However, in cases of agreement between the parsers, the resulting gold standards do not exhibit a clear disadvantage relative to the Human Gold annotations.

This result highlights the crucial role of parser specific approval in the overall preference of judges towards human-based annotations in table A.2. Furthermore, it suggests that annotations on which multiple state of the art parsers agree are of sufficiently high accuracy to be used to save annotation time without substantial impact on the quality of the resulting resource. In section A.7 we propose an annotation scheme which leverages this insight.

A.5 Inter-annotator Agreement

Agreement estimates in NLP are often obtained in annotation setups where both annotators edit the same automatically generated input. However, in such experimental conditions, anchoring can introduce cases of spurious disagreement as well as spurious agreement between annotators due to alignment of one or both participants towards the given input. The initial quality of the provided annotations in combination with the parser bias effect observed in section A.3 may influence the resulting agreement estimates. For example, in Marcus et al. [62] annotators were shown to produce POS tagging agreement of 92.8 on annotation from scratch, compared to 96.5 on reviews of tagger output.

Our goal in this section is to obtain estimates for inter-annotator agreement on POS tagging and dependency parsing that control for parser bias, and as a result, reflect more accurately human agreement on these tasks. We thus introduce a novel pipeline based on human annotation only, which eliminates parser bias from the agreement measurements. Our experiment extends the human-based annotation study of Marcus et al. [62] to include also syntactic trees. Importantly, we include an additional review step for the initial annotations, designed to increase the precision of the agreement measurements by reducing the number of errors in the original annotations.

For this experiment, we use 300 sentences (7,227 tokens) from section 23 of the PTB-WSJ, the standard test set for English parsing in NLP. The experimental setup, depicted graphically in figure 2, includes four participants randomly assigned for each sentence to annotation and review tasks. Two of the participants provide the sentence with annotations from scratch, while the remaining two participants provide reviews. Each reviewer edits one of the annotations independently, allowing for correction of annotation errors while

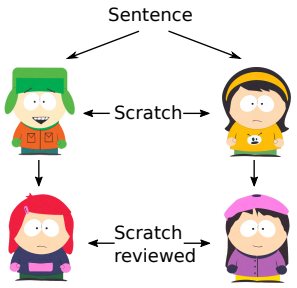


Figure A-2: Experimental setup for the inter-annotator agreement experiment. 300 sentences (7,227 tokens) from section 23 of the PTB-WSJ are annotated and reviewed by four participants. The participants are assigned to the following tasks *at random* for each sentence. Two participants annotate the sentence from scratch, and the remaining two participants review one of these annotations each. Agreement is measured on the annotations ("scratch") as well after assigning the review edits ("scratch reviewed").

maintaining the independence of the annotation sources. We measure agreement between the initial annotations ("scratch"), as well as the agreement between the reviewed versions of our sentences ("scratch reviewed").

The agreement results for the annotations and the reviews are presented in table A.4. The initial agreement rate on POS annotation from scratch is higher than in [62]. This difference is likely to arise, at least in part, due to the fact that their experiment was conducted at the beginning of the annotation project, when the annotators had a more limited annotation experience compared to our participants. Overall, we note that the agreement rates from scratch are relatively low. The review round raises the agreement on all the evaluation categories due to elimination of annotation errors present the original annotations.

	POS	UAS	LA	LAS
scratch	94.78	93.07	92.3	88.32
scratch reviewed	**95.65**	**94.17**	**94.04**	**90.33**

Table A.4: Inter-annotator agreement on 300 sentences (7,227 tokens) from the PTB-WSJ section 23. "scratch" is agreement on independent annotations from scratch. "scratch reviewed" is agreement on the same sentences after an additional independent review round of the annotations.

Our post-review agreement results are consequential in light of the current state of the art performance on tagging and parsing in NLP. For more than a decade, POS taggers have been achieving over 97% accuracy with the PTB POS tag set on the PTB-WSJ test set. For

example, the best model of the Stanford tagger reported in Toutanova et al. [104] produces an accuracy of 97.24 POS on sections 22-24 of the PTB-WSJ. These accuracies are above the human agreement in our experiment.

With respect to dependency parsing, recent parsers obtain results which are on par or higher than our inter-annotator agreement estimates. For example, Weiss et al. [109] report 94.26 UAS and Andor et al. [3] report 94.61 UAS on section 23 of the PTB-WSJ using an automatic conversion of the PTB phrase structure trees to Stanford dependencies [24]. These results are not fully comparable to ours due to differences in the utilized dependency formalism and the automatic conversion of the annotations. Nonetheless, we believe that the similarities in the tasks and evaluation data are sufficiently strong to indicate that dependency parsing for standard English newswire may be reaching human agreement levels.

A.6 Related Work

The term "anchoring" was coined in a seminal paper by Tversky and Kahneman [106], which demonstrated that numerical estimation can be biased by uninformative prior information. Subsequent work across various domains of decision making confirmed the robustness of anchoring using both informative and uninformative anchors [35]. Pertinent to our study, anchoring biases were also demonstrated when the participants were domain experts, although to a lesser degree than in the early anchoring experiments [111, 66].

Prior work in NLP examined the influence of pre-tagging [31] and pre-parsing [95] on human annotations. Our work introduces a systematic study of this topic using a novel experimental framework as well as substantially more sentences and annotators. Differently from these studies, our methodology enables characterizing annotation bias as anchoring and measuring its effect on tagger and parser evaluations.

Our study also extends the POS tagging experiments of Marcus et al. [62], which compared inter-annotator agreement and annotation quality on manual POS tagging in annotation from scratch and tagger-based review conditions. The first result reported in that study was that tagger-based editing increases inter-annotator agreement compared to annotation from scratch. Our work provides a novel agreement benchmark for POS tagging

which reduces annotation errors through a review process while controlling for tagger bias, and obtains agreement measurements for dependency parsing. The second result reported in Marcus et al. [62] was that tagger-based edits are of higher quality compared to annotations from scratch when evaluated against an additional independent annotation. We modify this experiment by introducing ranking as an alternative mechanism for quality assessment, and adding a review round for human annotations from scratch. Our experiment demonstrates that in this configuration, parser-based annotations are of *lower* quality compared to human-based annotations.

Several estimates of expert inter-annotator agreement for English parsing were previously reported. However, most such evaluations were conducted using annotation setups that can be affected by an anchoring bias [17, 83, 94]. A notable exception is the study of Sampson and Babarczy [90] who measure agreement on annotation from scratch for English parsing in the SUSANNE framework [89]. The reported results, however, are not directly comparable to ours, due to the use of a substantially different syntactic representation, as well as a different agreement metric. Their study further suggests that despite the high expertise of the annotators, the main source of annotation disagreements was annotation errors. Our work alleviates this issue by using annotation reviews, which reduce the number of erroneous annotations while maintaining the independence of the annotation sources. Experiments on non-expert dependency annotation from scratch were previously reported for French, suggesting low agreement rates (79%) with an expert annotation benchmark [39].

A.7 Discussion

We present a systematic study of the impact of anchoring on POS and dependency annotations used in NLP, demonstrating that annotators exhibit an anchoring bias effect towards the output of automatic annotation tools. This bias leads to an artificial boost of performance figures for the parsers in question and results in lower annotation quality as compared with human-based annotations.

Our analysis demonstrates that despite the adverse effects of parser bias, predictions

that are shared across different parsers do not significantly lower the quality of the annotations. This finding gives rise to the following *hybrid annotation* strategy as a potential future alternative to human-based as well as parser-based annotation pipelines. In a hybrid annotation setup, human annotators review annotations on which several parsers agree, and complete the remaining annotations from scratch. Such a strategy would largely maintain the annotation speed-ups of parser-based annotation schemes. At the same time, it is expected to achieve annotation quality comparable to human-based annotation by avoiding parser specific bias, which plays a pivotal role in the reduced quality of single-parser reviewing pipelines.

Further on, we obtain, to the best of our knowledge for the first time, syntactic inter-annotator agreement measurements on WSJ-PTB sentences. Our experimental procedure reduces annotation errors and controls for parser bias. Despite the detailed annotation guidelines, the extensive experience of our annotators, and their prior work as a group, our experiment indicates rather low agreement rates, which are below state of the art tagging performance and on par with state of the art parsing results on this dataset. We note that our results do not necessarily reflect an upper bound on the achievable syntactic inter-annotator agreement for English newswire. Higher agreement rates could in principle be obtained through further annotator training, refinement and revision of annotation guidelines, as well as additional automatic validation tests for the annotations. Nonetheless, we believe that our estimates reliably reflect a realistic scenario of expert syntactic annotation.

The obtained agreement rates call for a more extensive examination of annotator disagreements on parsing and tagging. Recent work in this area has already proposed an analysis of expert annotator disagreements for POS tagging in the absence of annotation guidelines [80]. Our annotations will enable conducting such studies for annotation with guidelines, and support extending this line of investigation to annotations of syntactic dependencies. As a first step towards this goal, we plan to carry out an in-depth analysis of disagreement in the collected data, characterize the main sources of inconsistent annotation and subsequently formulate further strategies for improving annotation accuracy. We believe that better understanding of human disagreements and their relation to disagreements between humans and parsers will also contribute to advancing evaluation methodologies

for POS tagging and syntactic parsing in NLP, an important topic that has received only limited attention thus far [92, 79].

Finally, since the release of the Penn Treebank in 1992, it has been serving as the standard benchmark for English parsing evaluation. Over the past few years, improvements in parsing performance on this dataset were obtained in small increments, and are commonly reported without a linguistic analysis of the improved predictions. As dependency parsing performance on English newswire may be reaching human expert agreement, not only new evaluation practices, but also more attention to noisier domains and other languages may be in place.

Bibliography

[1] Anne Abeillé, Lionel Clément, and François Toussenel. Building a treebank for french. In *Treebanks*, pages 165–187. Springer, 2003.

[2] Rosa Alonso Alonso. *Crosslinguistic Influence in Second Language Acquisition*, volume 95. Multilingual Matters, 2015.

[3] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. In *Proceedings of ACL*, pages 2442–2452, 2016.

[4] Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. Weakly supervised part-of-speech tagging using eye-tracking data. In *ACL*, volume 2, pages 579–584, 2016.

[5] Maria Barrett and Anders Søgaard. Reading behavior predicts syntactic categories. In *CoNLL*, pages 345–349, 2015.

[6] Maria Barrett and Anders Søgaard. Using reading behavior to predict grammatical functions. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, pages 1–5, 2015.

[7] Yevgeni Berzak, Yan Huang, Andrei Barbu, Anna Korhonen, and Boris Katz. Anchoring and agreement in syntactic annotations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2224, Austin, Texas, 2016. Association for Computational Linguistics.

[8] Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. Universal dependencies for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 737–746, Berlin, Germany, August 2016. Association for Computational Linguistics.

[9] Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. Predicting native language from gaze. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 541–551, Vancouver, Canada, 2017. Association for Computational Linguistics.

[10] Yevgeni Berzak, Roi Reichart, and Boris Katz. Reconstructing native language typology from foreign language usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 21–29, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics.

[11] Yevgeni Berzak, Roi Reichart, and Boris Katz. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in esl. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 94–102, Beijing, China, July 2015. Association for Computational Linguistics.

[12] Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, volume 168, 2002.

[13] Julian Brooke and Graeme Hirst. Measuring interlanguage: Native language identification with l1-influence metrics. In *LREC*, pages 779–784, 2012.

[14] Serhiy Bykh and Detmar Meurers. Advancing linguistic features and insights by label-informed feature grouping: An exploration in the context of native language identification. In *COLING*, 2016.

[15] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[16] Aoife Cahill, Binod Gyawali, and James V Bruno. Self-training for parsing learner text. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 66–73, 2014.

[17] John Carroll, Guido Minnen, and Ted Briscoe. Corpus annotation for parser evaluation. *arXiv preprint cs/9907013*, 1999.

[18] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[19] Chris Collins and Richard Kayne. *Syntactic Structures of the world's languages*. 2009.

[20] Uschi Cop, Denis Drieghe, and Wouter Duyck. Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PLOS ONE*, 10(8):1–38, 08 2015.

[21] David Crystal. *English as a global language*. Ernst Klett Sprachen, 2003.

[22] Hal Daumé III. Non-parametric Bayesian areal linguistics. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 593–601. Association for Computational Linguistics, 2009.

[23] Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of LREC*, pages 4585–4592, 2014.

[24] Marie-Catherine de Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

[25] Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. Technical report, Technical report, Stanford University, 2008.

[26] Ana Dıaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. Towards interlanguage pos annotation for effective learner corpora in sla and flt. *Language Forum*, 36(1–2):139–154, 2010.

[27] Markus Dickinson and Marwa Ragheb. Dependency annotation for learner corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, pages 59–70, 2009.

[28] Markus Dickinson and Marwa Ragheb. Annotation for learner English guidelines, v. 0.1. Technical report, Indiana University, Bloomington, IN, June 2013. June 9, 2013.

[29] Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

[30] Paola E Dussias. Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics*, 30:149–166, 2010.

[31] Karën Fort and Benoît Sagot. Influence of pre-annotation on pos-tagged corpus development. In *Proceedings of the fourth linguistic annotation workshop*, pages 56–63, 2010.

[32] Jennifer Foster. Treebanks gone bad. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(3-4):129–145, 2007.

[33] Jennifer Foster, Joachim Wagner, and Josef Van Genabith. Adapting a wsj-trained parser to grammatically noisy text. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 221–224. Association for Computational Linguistics, 2008.

[34] Charles C Fries. Teaching and learning english as a foreign language. 1945.

[35] Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1):35–42, 2011.

[36] Susan M Gass and Larry Selinker. *Language Transfer in Language Learning: Revised edition*, volume 5. John Benjamins Publishing, 1992.

[37] Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (ef-camdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, 2013.

[38] 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.

[39] Kim Gerdes. Collaborative dependency annotation. *DepLing 2013*, 88, 2013.

[40] Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. International corpus of learner english, 2009.

[41] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. Glottolog 2.6. *Leipzig: Max Planck Institute for Evolutionary Anthropology.*, 2015.

[42] Hagen Hirschmann, Seanna Doolittle, and Anke Lüdeling. Syntactic annotation of non-canonical linguistic structures. 2007.

[43] Scott Jarvis and Scott A Crossley. *Approaching language transfer through text classification: Explorations in the detection-based approach*, volume 64. Multilingual Matters, 2012.

[44] Scott Jarvis and Aneta Pavlenko. *Crosslinguistic influence in language and cognition*. Routledge, 2007.

[45] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938.

[46] Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. Improving sentence compression by learning to predict gaze. *NAACL-HLT*, 2016.

[47] Ekaterina Kochmar and Ekaterina Shutova. Cross-lingual lexico-semantic transfer in language learning. Association for Computational Linguistics, 2016.

[48] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author‘s native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628. ACM, 2005.

[49] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[50] Robert Lado. Linguistics across cultures: Applied linguistics for language teachers. 1957.

[51] Godfrey N Lance and William Thomas Williams. A general theory of classificatory sorting strategies ii. clustering systems. *The computer journal*, 10(3):271–277, 1967.

[52] Tao Lei, Yu Xin, Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. Low-rank tensors for scoring dependency structures. In *Proceedings of ACL*, volume 1, pages 1381–1391, 2014.

[53] Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, 2:278–292, 1960.

[54] M. Paul Lewis. Ethnologue: Languages of the world. `www.ethnologue.com`, 2014.

[55] Paul M. Lewis, Gary F. Simons, and Charles D. Fennig, editors. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 2015.

[56] Patrick Littel, David Mortensen, and Lori Levin, editors. *URIEL Typological Database*. Pittsburgh: Carnegie Mellon University, 2016.

[57] Patrick Littell, David Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. *EACL 2017*, page 8, 2017.

[58] Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467, 2004.

[59] Shervin Malmasi and Mark Dras. Language transfer hypotheses with linear svm weights. In *EMNLP*, pages 1385–1390, 2014.

[60] Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. A Report on the 2017 Native Language Identification Shared Task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[61] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[62] 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.

[63] André FT Martins, Miguel Almeida, and Noah A Smith. Turning on the turbo: Fast third-order non-projective turbo parsers. In *ACL (2)*, pages 617–622. Citeseer, 2013.

[64] Gita Martohardjono and Suzanne Flynn. Language transfer: what do we really mean. In L. Eubank, L. Selinker, and M. Sharwood Smith, editors, *The current state of Interlanguage: studies in honor of William E. Rutherford*, pages 205–219. John Benjamins: The Netherlands, 1995.

[65] Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. In *ACL (2)*, pages 92–97. Citeseer, 2013.

[66] Thomas Mussweiler and Fritz Strack. Numeric judgments under uncertainty: The role of knowledge in anchoring. *Journal of Experimental Social Psychology*, 36(5):495–518, 2000.

[67] Ryo Nagata. Language family relationship preserved in non-native english. In *COLING*, pages 1940–1949, 2014.

[68] Ryo Nagata and Keisuke Sakaguchi. Phrase structure annotation and parsing for learner english. In *ACL*, 2016.

[69] Ryo Nagata, Edward Whittaker, and Vera Sheinman. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1210–1219. Association for Computational Linguistics, 2011.

[70] Ryo Nagata and Edward W. D. Whittaker. Reconstructing an indo-european family tree from non-native english texts. In *ACL*, pages 1137–1147, 2013.

[71] Tahira Naseem, Regina Barzilay, and Amir Globerson. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637. Association for Computational Linguistics, 2012.

[72] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The conll-2014 shared task on grammatical error correction. In *CoNLL Shared Task*, pages 1–14, 2014.

[73] Diane Nicholls. The cambridge learner corpus: Error coding and analysis for lexicography and elt. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 572–581, 2003.

[74] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[75] Terence Odlin. *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press, 1989.

[76] Helen O'Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, and Anna Korhonen. Survey on the use of typological information in natural language processing. In

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, pages 1297–1308, Osaka, Japan, December 2016.

[77] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *LREC*, 2012.

[78] Steven T Piantadosi, Harry Tily, and Edward Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011.

[79] Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, and Anders Søgaard. Do dependency parsing metrics correlate with human judgments? In *Proceedings of CoNLL*, 2015.

[80] Barbara Plank, Dirk Hovy, and Anders Søgaard. Linguistically debatable or just plain wrong? In *Proceedings of ACL: Short Papers*, pages 507–511, 2014.

[81] Ella Rabinovich, Noam Ordan, and Shuly Wintner. Found in translation: Reconstructing phylogenetic language trees from translations. In *ACL*, pages 530–540, 2017.

[82] Marwa Ragheb and Markus Dickinson. Defining syntax for learner language annotation. In *COLING (Posters)*, pages 965–974, 2012.

[83] Owen Rambow, Cassandre Creswell, Rachel Szekely, Harriet Taber, and Marilyn A Walker. A dependency treebank for english. In *Proceedings of LREC*, 2002.

[84] Leah Roberts and Anna Siyanova-Chanturia. Using eye-tracking to investigate topics in l2 acquisition and l2 processing. *Studies in Second Language Acquisition*, 35(02):213–235, 2013.

[85] Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, 48(1):65–92, 2014.

[86] Victoria Rosén and Koenraad De Smedt. Syntactic annotation of learner corpora. *Systematisk, variert, men ikke tilfeldig*, pages 120–132, 2010.

[87] Alla Rozovskaya and Dan Roth. Algorithm selection and model adaptation for esl correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 924–933. Association for Computational Linguistics, 2011.

[88] Alla Rozovskaya and Dan Roth. Building a state-of-the-art grammatical error correction system. *Transactions of the Association for Computational Linguistics*, 2(10):419–434, 2014.

[89] Geoffrey Sampson. English for the computer: Susanne corpus and analytic scheme. 1995.

[90] Geoffrey Sampson and Anna Babarczy. Definitional and human constraints on structural annotation of english. *Natural Language Engineering*, 14(04):471–494, 2008.

[91] Beatrice Santorini. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). *Technical Reports (CIS)*, 1990.

[92] Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *Proceedings of ACL*, pages 663–672, 2011.

[93] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, pages 591–611, 1965.

[94] Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. A gold standard dependency corpus for english. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

[95] Arne Skjærholt. Influence of preprocessing on dependency syntax annotation: speed and agreement. *LAW VII & ID*, 2013.

[96] J.J. Song. *The Oxford Handbook of Linguistic Typology*. Oxford Handbooks in Linguistics. OUP Oxford, 2011.

[97] Ben Swanson and Eugene Charniak. Extracting the native language signal for second language acquisition. In *HLT-NAACL*, pages 85–94, 2013.

[98] Ben Swanson and Eugene Charniak. Data driven language transfer hypotheses. *EACL 2014*, page 169, 2014.

[99] Oscar Täckström, Ryan McDonald, and Joakim Nivre. Target language adaptation of discriminative transfer parsers. *Proceedings of NAACL-HLT*, 2013.

[100] Yee Whye Teh, Hal Daumé III, and Daniel M Roy. Bayesian agglomerative clustering with coalescents. In *NIPS*, 2007.

[101] Joel Tetreault, Daniel Blanchard, and Aoife Cahill. A report on the first native language identification shared task. *NAACL/HLT 2013*, page 48, 2013.

[102] Joel Tetreault, Jennifer Foster, and Martin Chodorow. Using parse features for preposition selection and error detection. In *Proceedings of the acl 2010 conference short papers*, pages 353–358. Association for Computational Linguistics, 2010.

[103] Joel R Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *COLING*, pages 2585–2602, 2012.

[104] 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.

[105] Oren Tsur and Ari Rappoport. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16. Association for Computational Linguistics, 2007.

[106] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[107] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[108] Ronald Wardhaugh. The contrastive analysis hypothesis. *TESOL quarterly*, pages 123–130, 1970.

[109] David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. Structured training for neural network transition-based parsing. In *Proceedings of ACL*, pages 323–333, 2015.

[110] Randal L Whitman and Kenneth L Jackson. The unpredictability of contrastive analysis. *Language learning*, 22(1):29–41, 1972.

[111] Timothy D Wilson, Christopher E Houston, Kathryn M Etling, and Nancy Brekke. A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4):387, 1996.

[112] Sze-Meng Jojo Wong and Mark Dras. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 53–61. Citeseer, 2009.

[113] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In *ACL*, pages 180–189, 2011.