Automated Gloss Mapping for Inferring Grammatical Properties

Michael Lockwood

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2016

Committee:

Emily M. Bender

Fei Xia

Program Authorized to Offer Degree:
Department of Linguistics

©Copyright 2016
Michael Lockwood

University of Washington

Abstract

Automated Gloss Mapping
for Inferring Grammatical Properties

Michael Lockwood

Chair of the Supervisory Committee:
Emily M. Bender
Department of Linguistics

This thesis describes a software system that maps glosses from interlinear glossed text (IGT) to an internally consistent set. This study hypothesizes that mapping glosses supports better inference of grammatical properties. Inference refers to analyzing information from IGT to automatically determine grammatical properties of a language for which a computational grammar can be constructed. The IGT will likely contain unknown or non-standard glosses. By mapping all glosses to an internally consistent set the non-standard rate should decrease which would provide more precise information for infering grammatical properties. The inference procedure matches standard grams from a language to tense, aspect, and mood categories. These inferred gram and category pairs called choices are used to create computational grammars. The final results demonstrate that the methodology successfully reduces the non-standard gloss rate.

ACKNOWLEDGMENTS

I want to thank Emily M. Bender for her direction and guidance, Fei Xia for sharing her machine learning expertise, Olga Zamaraeva for her assistance with MOM, Michael Wyane Goodman for his work on Xigt, the students of LING 575 in Spring 2015 for their feedback during the initial development of this project, and the students of previous LING 567 classes for their IGT and choices files.

TABLE OF CONTENTS

Page

List of Tables iv

List of Figures vi

Glossary vii

Chapter 1: Introduction 1

Chapter 2: Background 4
2.1 AGGREGATION 4
2.2 LinGO Grammar Matrix 4
2.3 Xigt and Language CoLLAGE 5
2.4 Inference 6
2.5 Gloss Conventions 7
2.6 Tense, Aspect, and Mood 8
2.7 Summary 9

Chapter 3: Methodology 10
3.1 Gloss Classification 11
3.1.1 Data Extraction 11
3.1.2 Feature Development 11
3.1.3 Classification Model 17
3.2 Merge Unique Gloss Classifications 18
3.3 Produce Language Gram Sets 20
3.3.1 Post-Processing 20
3.3.2 Referencing 21
3.4 Inference of Tense, Aspect, and Mood Choices 22

3.4.1 Category Assignment . . . 22
3.4.2 Choices File Creation . . . 23
3.5 Summary . . . 24

Chapter 4: Evaluation . . . 25
4.1 Original Datasets . . . 25
4.2 Gold Standard . . . 27
4.2.1 Classification Types . . . 27
4.2.2 Annotation Procedure . . . 31
4.2.3 Reporting . . . 33
4.3 Evaluation Methodology . . . 35
4.3.1 Merged Classifications . . . 35
4.3.2 Language Gram Sets . . . 35
4.3.3 Inference . . . 38
4.4 Summary . . . 40

Chapter 5: Results . . . 41
5.1 Merged Classifications . . . 41
5.2 Language Gram Sets . . . 42
5.3 Inference . . . 45
5.4 Result Visualization . . . 47
5.5 Summary . . . 49

Chapter 6: Discussion . . . 50
6.1 Methodology . . . 50
6.2 Evaluation . . . 52
6.2.1 Data Assumptions . . . 52
6.2.2 Error Analysis . . . 53
6.3 Summary . . . 54

Chapter 7: Conclusion . . . 55

Appendix A: Standard Gloss Set . . . 56

Appendix B: Standard Predetermined Set . . . 62

Appendix C: Matrix-ODIN Morphology Enhanced Inference 66

LIST OF TABLES

Table Number Page

1.1 Example 1 Gloss Descriptions 2

3.1 Example Vector from fra, DEV2: IPFV 15
3.1 Example Vector from fra, DEV2: IPFV 16
3.2 Naïve Bayes Interpolation of IMP 19
3.3 Classification Interpolation of IMP 20

4.1 Language CoLLAGE Datasets and Languages 25
4.1 Language CoLLAGE Datasets and Languages 26
4.1 Language CoLLAGE Datasets and Languages 27
4.2 Gold Classification Type Frequencies 33
4.3 Gold Classification Type Frequencies Alternate 34
4.4 Language Gram Set Confusion Matrices Class Explanations 36
4.5 Language Gram Set Comparison Class Explanations 37

5.1 Merged Classification Accuracies 42
5.2 Language Gram Sets Comparative Evaluation 43
5.3 Language Gram Sets Leave-One-Out Cross Validation 44
5.3 Language Gram Sets Leave-One-Out Cross Validation 45
5.4 Inference Comparative Evaluation 46

A.1 Standard Gloss Set 56
A.1 Standard Gloss Set 57
A.1 Standard Gloss Set 58
A.1 Standard Gloss Set 59
A.1 Standard Gloss Set 60
A.1 Standard Gloss Set 61

B.1 Standard Predetermined Set 62
B.1 Standard Predetermined Set 63

B.1 Standard Predetermined Set 64
B.1 Standard Predetermined Set 65

LIST OF FIGURES

Figure Number Page

5.1 Language Gram Sets Comparative Evaluation Scatter Plot 47
5.2 Language Gram Sets Leave-One-Out-Cross-Validation Scatter Plot 48
5.3 Inference Comparative Evaluation Scatter Plot 48

GLOSSARY

CATEGORY: A group of linguistic attributes. Examples of categories are case, tense, aspect, and mood.

CHOICE: A linguistic attribute assigned to a language. For this project choices are stored in the form of a `choices file` which organizes these linguistic attributes for computational purposes.

CLASSIFICATION LABEL: Refers to the label of a vector for machine learning. The classification label set for gloss mapping includes all of the standard grams in addition to the classification types except misspelled and confused.

CLASSIFICATION TYPE: Archetypes that describe categories of unprocessed glosses. These types are used for evaluation and classification. Classification types include standard, misspelled, confused, incomplete, combined, user-identified, unrecovered, part-of-speech, and lexical entries.

FEATURE: A field of data which acts as a parameter for machine learning by informing the model.

FEATURE VALUE: A specific value for a feature held by a particular machine learning vector.

FINAL: Describes the state of a gloss after post-processing from the gloss mapping stage.

GLOSS: All tokens in the gloss line are considered glosses.

GLOSS MAPPING: Describes the methodology that includes both classifying glosses and the post-processing that occurs to convert select non-standard type labels to standard grams.

GOLD STANDARD: A gold standard contains a gold classification type, gold classification label, and gold gram(s). Gold means that is considered the standard for evaluation purposes. Developed by annotating all unique glosses.

GRAM: The tokens of a gloss line which contain linguistic information and have been identified in the standard set.

IGT: Interlinear Glossed Text, collections of phrases containing the source language line, a morpheme line, a gloss line, and a translation line typically to English.

INPUT: Describes IGT and choices files before any processing has occurred.

NON-STANDARD: Describes glosses which are not found in the standard set.

NON-STANDARD TYPE: Refers to all of the classification types except for *standard*.

OUTPUT: Describes the state of a gloss after post-processing from the gloss mapping stage.

POST-PROCESSED: Describes glosses which have been post-processed after machine learning. This involves combining individual glosses to a unique gloss for the language and processing refinements if necessary. Refinements convert some non-standard type labels to grams if the input gloss is in the predetermined set.

PREDETERMINED: A set of glosses which may appear as incomplete or part-of-speech classification types. They should be automatically identified and if incomplete converted to the appropriate standard grams.

UNPROCESSED: Describes IGT and choices files before any processing has occurred.

STANDARD: Describes glosses which are in the standard set.

STANDARDIZED: Describes glosses that have been output by the machine learning process which maps glosses to the standard set.

UNIQUE GLOSS: Refers to all of the unprocessed glosses in a language which exactly match. Collectively they form a unique gloss. This is used to disseminate information from the unique gloss to each gloss instance and vice-versa.

VECTOR: A collection of all feature and feature value pairs for a gloss instance. Machine learning algorithms train and test models by using vectors.

Chapter 1

INTRODUCTION

This paper presents a method for mapping glosses found in Interlinear Gloss Text (IGT) to an internally consistent set for the purpose of inferring the grammatical properties of a language. An example of IGT is found in Example 1.

(1) Ivan chital knigu
Ivan chita-l knig-u
Ivan.MASC.3SG.NOM read-IMP.PST.3SG book-FEM.3SG.ACC
Ivan was reading a book. [rus] (Gracheva & Suskic, 2010:42)

The first line of IGT contains the text in the source language. The second line separates morphemes in the language. The third line contains the glosses attached to each morpheme in addition to lexical entries. This gloss line is the primary line of focus for mapping glosses. The final line is the translation line of the text to English.[1]

The terminology for glosses is critical for precisely referring to different classes of glosses. Any value in the gloss line is considered a gloss. This includes the lexical entries such as *Ivan* in the above example. Grams are those that contain linguistic information such as case, person, number, tense, aspect, and mood among other categories. Therefore all grams are glosses but not all glosses are grams. A standard gram is one which is contained within the internally consistent set of grams. A non-standard gram is one which is not included in the internally consistent set or is confused with another gram. For the interest of this study the internally consistent set is also called the standard set. This does not imply that the set has been adopted or recognized by any

[1]IGT often only has three of these four lines and will omit either the first line (source) or the second line (morphemes).

organization. Table 1.1 illustrates the different gloss classifications for the glosses in Example 1.

Gloss	**Type**
Ivan	Gloss (lexical item)
MASC	Non-Standard Gram (misspelled)
3SG	Standard Gram
NOM	Standard Gram
read	Gloss (lexical item)
IMP	Non-Standard Gram (confused)
PST	Standard Gram
3SG	Standard Gram
book	Gloss (lexical item)
FEM	Non-Standard Gram (misspelled)
3SG	Standard Gram
ACC	Standard Gram

Table 1.1: Example 1 Gloss Descriptions

The gloss mapping system takes IGT from a language as input and will produce a set of standard grams. Mapping glosses to a standard set is hypothesized to improve the performance of tools that infer grammatical properties about the language.

Another set of terminology applies to the stages of glosses. Glosses directly from IGT, before any processing or learning has happened, are known as input or unprocessed glosses. Once machine learning has classified them they are considered classified or standardized glosses. Due to the inclusion of both grams and non-standard types §4.2.1 in the standard classification label set some post-processing occurs which converts select non-standard types to grams. Once the post-processing is finished the grams are identified by the terms final, output, and post-processed.

The thesis asks two research questions:

I What methodology will map glosses to a standard set such that the resulting mapped grams reduce the rate of non-standard grams?

II Will increasing the rate of standard grams improve inference procedures that utilize grams?

This thesis will approach the second question via a case study of tense, aspect, and mood. It is plausible that grams could be used to infer other grammatical properties.

Chapter 2 provides context for the existing research and software upon which this study builds. Chapter 3 describes the process for mapping glosses to the standard set and the inference procedure for tense, aspect, and mood. Chapter 4 explains baseline and metrics for both the gloss mapping and inference. Chapter 5 displays the results of all methods in comparison to their respective baselines. Chapter 6 presents an error analysis and overview of potential improvements and future work.

Chapter 2
BACKGROUND

This chapter provides some background on the previous work upon which this study builds including AGGREGATION, the LinGO Grammar Matrix, Xigt, inference frameworks, gloss conventions, and a library for tense, aspect, and mood.

2.1 AGGREGATION

This thesis project belongs within the broader research of the AGGREGATION project [Bender et al., 2013].[1] The AGGREGATION project, Automatic Generation of Grammars for Endangered Languages from Glosses and Typological Information, pursues the automatic generation of computational resources for endangered languages by using sources of IGT. The software products produced by AGGREGATION will allow a user to input IGT data for a language of their choice and receive a computational grammar as output. AGGREGATION exists within the DELPH-IN consortium, a collection of computational linguistic research that utilizes Head-Driven Phrase Structure Grammars (HPSG) [Pollard and Sag, 1994] and Minimal Recursion Semantics (MRS) [Copestake et al., 2005]. Every AGGREGATION generated grammar is modeled within the HPSG and MRS framework.

2.2 LinGO Grammar Matrix

The LinGO Grammar Matrix customization system [Bender et al., 2002][Bender et al., 2010] is an online resource for generating computational grammars. These computational grammars contain a core grammar that is hypothesized to be cross-linguistically useful. Additional rules and properties adapt the Grammar Matrix core grammar to a particular language which adds both coverage and

[1] `http://http://depts.washington.edu/uwcl/aggregation/`

constraints. A user of the customization system can answer preset questions about the linguistic systems and properties of a language. Such systems include but are not limited to word order, case system, tense, aspect, mood, morphology, and information structure. The information the user enters into the customization system is stored in a data format called a `choices file`. The user may export the choices file and then load the file at a later time; the Grammar Matrix customization system will know how to read the choices file. Once the user is satisfied with the information they have provided to the customization system they can request to download a computational grammar. The user can directly utilize this computational grammar with existing DELPH-IN tools such as the Linguistic Knowledge Builder (LKB) [Copestake, 2002]. The difficulty with the Grammar Matrix as it stands now is that it requires a user to spend copious amounts of time entering linguistic data into a web browser. AGGREGATION intends to automate this process.

2.3 Xigt and Language CoLLAGE

The AGGREGATION project team developed the Xigt [Goodman et al., 2015] data format which stores collections of IGT. The Xigt format organizes all of the word, gloss, morpheme, and translation data into eXtensible Mark-up Language (XML) tiers. These tiers are then aligned with IDs, values, and string positioning. Xigt delivers both the individual layers such as words, glosses, and morphemes and the alignments between tiers. The alignment between words, glosses, and morphemes is critical for the methodology of this thesis project. A program called INTENT [Xia et al., 2014] automates the enrichment of Xigt. Enrichment produces the alignments between each of the tiers as well as POS tags and syntactic parses which are not used in this study. The Xigt data available for this project comes from Language CoLLAGE [Bender, 2014b]. Language CoLLAGE consists of IGT and choices files from previous term projects in the University of Washington LING 567 class. This is a substantial resource for evaluation because it links both IGT and choices files developed synchronously by the students.

2.4 Inference

Inference tools automate the creation of choices files. The first inference procedures that AGGREGATION developed were for word order and case systems [Bender et al., 2013]. The word order inference procedure utilizes a very specific methodology for its inference which would not be useful for tense, aspect, and mood.

Case inference was tested by two different methods. One called GRAM assumed standard gloss inputs and inferred case systems from the glosses found in the IGT for the language. The other method called SAO (Subject Agent, and Object) utilized the English translation line to determine whether the verb had subject, agent, and/or object attributes and which gloss referred to each. SAO was developed to avoid resolving the unclean gloss data by instead focusing on theoretical properties of languages [Bender et al., 2013]. This approach was motivated because of the high rate of non-standard glosses found in IGT. Datasets outside Language CoLLAGE may contain even greater rates of non-standard glosses. ODIN [Lewis and Xia, 2010][Xia et al., 2014] is one such dataset exemplifying a high non-standard rate of glosses. ODIN is a collection of IGT instances that were automatically extracted from linguistic documents on the web. ODIN was featured in Lewis and Xia's research product RiPLes, a set of tools for resource poor languages [Lewis and Xia, 2008]. As of the writing of this thesis, ODIN has collected 130,351 instances of IGT from 2017 documents representing 1274 languages.[2]

I explored two methods for inferring tense, aspect, and mood. I scoped both a modified GRAM method and a theoretical method. Early on in project development I determined such a theoretical method for tense, aspect, and mood would be infeasible because they are not structural properties of languages. Furthermore, English is not rich in aspect and mood indicators and thus the translation line would not provide as much benefit as it did for case inference [Bender et al., 2013]. Tense, aspect, and mood choices vary greatly among languages with which category they represent. An example is habitual which appears as both a tense and an aspect in Language CoLLAGE. Because the GRAM method was selected the project focused on mapping glosses to a standard set.

---

[2] `http://odin.languistlist.org`. Accessed December 1, 2015.

2.5 Gloss Conventions

This study proposes its own internally consistent standard set of grams. Many common glosses have a vast array of spellings. Present helps to illustrate this; it occurs frequently as Present, PRS, and PRES. A standard set with a referencing system can map Present and PRS as equivalent, PRES as a misspelling of PRS, and therefore determine that Present, PRS, and PRES are all equivalent grams. This is critically motivated for two reasons. The first is that inference evaluation scores would dramatically increase. Students who developed data in Language CoLLAGE often named choice values separately from the glosses in the IGT. They both refer to the same linguistic property but have different names. It is very common to observe GOLD Ontology [Indiana University, 2010] names in the choices file and Leipzig [Bickel et al., 2008] grams in the IGT. Without a standard that maps equivalent grams together the evaluation scores of the inference procedure would be artificially low because the GOLD and Leipzig grams would not match. Secondly, mapping glosses reduces spelling errors and finds standard grams that the author failed to gloss from lexical items such as pronouns and demonstratives. This reduces false positives and increases true positives.

Most grams in the standard set include pairs of equivalent Leipzig [Bickel et al., 2008] and GOLD ontology [Indiana University, 2010] glosses. The Leipzig glossing standard proposed rules for gloss conventions and has a lexicon of mapped grams [Bickel et al., 2008]. The GOLD ontology produced a list of linguistic properties which tends to be more complete than Leipzig [Indiana University, 2010]. While Leipzig glossing rules have been propositioned as a potential standard, neither Leipzig nor GOLD are universally recognized as glossing standards.

Often times there is a gloss that represents a well-known linguistic property for which an official Leipzig and/or GOLD gram does not exist. The standard set remedies this by containing a list of pseudo-grams which are hypothesized to be useful for the purpose of this study. A pseudo-gram indicates that a gram belongs within the standard set although it is not found with Leipzig and/or GOLD conventions. An example is the Leipzig pseudo-gram of HAB for habitual. The Leipzig standards do not have a representation for habitual. The GOLD Ontology has a gram of Habitual which becomes paired with the Leipzig pseudo-gram of HAB.

The standard set took quite some time to develop. It contains a full accounting of all Leipzig suggested grams and an additional set of about thirty grams. It continues to expand as time is available to add glosses. Besides matching pairs and pseudo-grams it also contains categorical suggestions to explain which linguistic properties the gram might entail. For example, HAB and Habitual may be either tense or aspect and these properties are suggested in the standard set. Last, the standard set contains six classification types which are used for machine learning; refer to Chapter 4 which documents the classification types. At the time of writing there are 114 pairs in the standard set. The standard set is available in its entirety in Appendix A.

2.6 Tense, Aspect, and Mood

In 2011 Poulson developed the tense and aspect library for the Grammar Matrix customization system [Poulson, 2011]. Few choices files before 2011 include these categories.

Tenses, aspects, and moods differ from other systems in that certain choices vary in whether they are a tense, aspect, or mood by language and linguists have varying interpretations of their meaning. Poulson's approach purposefully designed the tense, aspect, and mood library to handle this variation [Poulson, 2011]. Case provides an example of the opposite. Variation is not allowed with case because it is assumed that there are a finite number of case systems, excluding oblique cases. Tense, aspect, and mood function not like major case systems but more like oblique cases which have general meaning but their precise meaning and usage varies by language. They also do not carry structural properties [Poulson, 2011].

Tense and aspect are impacted by a concept called hierarchy. An example of a hierarchy is that recent and remote are both subtypes of past if they are indeed past tense. It would not be unreasonable to suggest that a language used remote to denote some remote future event. This illustrates one manner in which tense and aspect can vary in their interpretation. To take the examples Poulson provided, there are languages which make varying kinds of distinctions in their hierarchies. It is common to assume a past, future, present, or some form of non-past or non-present distinction. But as Poulson notes there could be recent and non-recent distinction [Poulson, 2011]. The implication is that the tense and aspect library does not assume any pre-built hierarchies

but allows the user to define their own hierarchy entirely from scratch. Radio button options are allowed which specify past, present, and/or future. This was a design choice to delineate the likely hierarchies but again the user can avoid these options and define their own hierarchies. With aspect Poulson suggests a perfective and imperfective distinction but still allows the user to create their own hierarchy [Poulson, 2011].

Poulson's work also addresses situation aspect. Situation aspect includes stative and dynamic properties. The user can define situation aspects choices if they desire [Poulson, 2011]. This project does not attempt to model situation aspect and instead focuses on viewpoint aspect.

2.7 Summary

Mapping glosses for the purpose of inference is only possible because an extensive amount of work has occurred to build customizable computational grammars. The LinGO Grammar Matrix customization system provides the framework for converting choices files into computational grammars. The thorough work of previous LING 567 students has provided a wealth of data in Language CoLLAGE which includes both choices files and Xigt data. Using grams for inference has already been proven successful by the AGGREGATION project [Bender et al., 2013]. Mapping the glosses to a compliant set should improve those results which can extend to other linguistic systems beyond case, tense, aspect, and mood. Much credit is due to Poulson for creating the tense and aspect library of the customization system which interprets the tense, aspect, and mood section of choices files for the Grammar Matrix [Poulson, 2011].

Chapter 3

METHODOLOGY

The methodology includes four stages to map glosses and infer tense, aspect, and mood. The first stage involves classifying each gloss instance to a final standard gram or intermediary handling type. Then the classification results are interpolated as unique glosses by input string and language. Some post-processing occurs to further refine classified outputs and output a final standard gram set for the language. The system can use the standard gram sets for each language to infer tense, aspect, and mood properties; these properties will then be written as choices files.

The methodology will be illustrated with reference to the following Russian IGT:

(2) Ty spishj
Ty sp-ishj
2SG.NOM sleep-2SG.PRS
You are sleeping. [rus] (Gracheva & Suskic, 2010:15)

(3) Spishj ty
Sp-ishj ty
Sleep-2SG.PRS 2SG.NOM
You are sleeping. [rus] (Gracheva & Suskic, 2010:16)

(4) Ty spishj
Ty sp-ishj
2SG.NOM sleep-IMP.PRS.2SG
You are sleeping. [rus] (Gracheva & Suskic, 2010:58)

(5) Ty chitaeshj
Ty chita-eshj
2SG.NOM read-2SG
You are reading. [rus] (Gracheva & Suskic, 2010:85)

These IGT were selected because they contain all references to 2SG. This will demonstrate the process of vectorizing each occurrence of 2SG and the interpolation of these glosses to a single

classification. Note that like many languages these IGT examples contain inconsistencies. In some contexts the author chose to add the tense and aspect, sometimes it was only one, and in one case it was neither. In all four of these examples both the tense and the aspect, Imperfective and Present, should have been identified in the gloss line. One last final note, observe how (2) and (4) are the exact same sentence in Russian yet they have differing gloss lines.

3.1 Gloss Classification

The gloss classification procedure comprises the bulk of the gloss mapping methodology. Raw input glosses are extracted from Xigt data. Each gloss instance is converted to a feature vector. There is significant discussion about feature development because the selection of features informs the final model and that selection impacts the model performance. The model is trained and tested using these feature vectors to produce classifications for each input gloss. These classifications are either standard grams or intermediary labels that will be resolved during post-processing.

3.1.1 Data Extraction

The data extraction provides context for the gloss classification task. The loading function accepts datasets with languages which have a pair of an ISO 639-3 value and a path to the location of an enriched-Xigt file for the language. The first time a script runs for that exact match of language and dataset the gloss, morpheme, and translation information is extracted from the enriched-Xigt file.

Each gloss becomes a vector and information pertaining to that gloss is stored along with it as features. In the Russian example, there are eight vectors representing each occurrence of 2SG. There are two instances of 2SG in each of the four IGT examples.

3.1.2 Feature Development

During the vector development stage mentioned in §3.1.1, a vector for each gloss instance is created and information about the gloss is stored along with it. This information is then encoded as features

within the vector which includes the gold standard for the gloss. These features inform the machine learning algorithms to create a model that will intake glosses and output the corresponding standard gram for the gloss. The gold standard is obtained from the gold standard annotation; refer to Chapter 4 for more information about gold standard development.

This section presents all of the features that were explored for this study and then identifies which features were actually used in the final analysis. The first instance of 2SG and IMP of Example 4 illustrate how each of these features are created. It is important to recognize that when numerical values are possible the value of feature is not set to that numerical value. Instead the feature specifies the feature class and the attributed value of the feature class. The actual value of the feature is always set to 1 to denote presence of the value. Absence of a feature means it is not included in the set or is marked with a 0. This design decision occurred because multivalued features created very low performance results.

The feature set includes ancillary features such as the gloss itself, the ISO code, nearby glosses within a window size of two glosses, whether the gloss is a standard gram, and whether the gloss is contained within the predetermined set. The predetermined set contains pronouns, demonstratives, and part-of-speech glosses that should be automatically identified for post-process corrections. Refer to Appendix B for further explanation and a table of all of the predetermined set glosses.

(6) **2SG**

gloss_2sg = 1

iso_rus = 1

gloss_next_nom = 1

gloss_next2_sleep = 1

standard_2sg = 1

conversion_2sg = 0

(7) **IMP**

gloss_imp = 1

iso_rus = 1

gloss_prev2_nom = 1

gloss_prev_sleep = 1

gloss_next_prs = 1

gloss_next2_2sg = 1

standard_imp = 1

predetermined_imp = 0

A feature is added for each gloss that shares the same morpheme as the current gloss to the vector of the current gloss. Another feature sets the count of the number of shared morphemes. Examples 8 and 9 demonstrate shared morphemes. In Example 8 the morpheme *ty* aligns with both 2SG and NOM; this indicates that 2SG shares a morpheme with NOM. In Example 9 IMP shares the morpheme *ishj* with PRS and 2SG. Another set of features is developed called segments. Segmentation is a recursive process that searches for standard gram string matches within the input gloss. This includes the remaining letters that were not matched. The matched grams are added under a segmentation matched feature value and the non-matched strings are added with a segmentation non-matched feature value.

(8) **2SG**

shared_morpheme_NOM = 1

shared_morpheme_count_1 = 1

segment_match_2 = 1

segment_match_sg = 1

(9) **IMP**

shared_morpheme_prs = 1

shared_morpheme_2sg = 1

shared_morpheme_count_2 = 1

segment_match_m = 1

segment_nonmatch_i = 1

segment_nonmatch_p = 1

The script then calculates the string edit distance between the current gloss and every gram in the standard set. It stores the result for each as a feature. The function uses the Levenshtein string distance measure [Levenshtein, 1966]. There are too many to document but a partial list is included below in each example.

(10) **2SG**

distance_2_2 = 1

distance_nom_3 = 1

distance_imp_3 = 1

distance_ipfv_4 = 1

distance_prs_2 = 1

(11) **IMP**

distance_2_3 = 1

distance_nom_2 = 1

distance_imp_0 = 1

distance_ipfv_2 = 1

distance_prs_2 = 1

To disambiguate lexical entries from glosses the script then assesses how often the gloss was observed in the dataset, whether the gloss lacks vowels, a count of the vowels the gloss contains, the length of the gloss, whether the gloss is a lexical entry, and the use of case (all lower, all upper, or mixed). The count feature can only be determined after all glosses have been read from the input data. It requires knowledge of how often the gloss occurred in the particular language. This means that all vectors representing that unique gloss will have the same count number. The count feature is meant to disambiguate lexical entries from grams.

(12) **2SG**

count_8 = 1

length_3 = 1

vowels_0 = 1

upper_case = 1

(13) **IMP**

count_12 = 1

length_3 = 1

vowels_1 = 1

upper_case = 1

An example of a vector's features is found in Table 3.1. The vector represents the actual features of the gloss IPFV from an example IGT of French in DEV2. Remember that even though the feature values in the table specify non-1 values, the features are specified as feature class and feature value equal to 1. Computational this is represented as `(feature class)_(feature value) = 1`.

Table 3.1: Example Vector from fra, DEV2: IPFV

Feature Class	**Feature Value(s)**	**Explanation**
gloss	ipfv	The lower case gloss
is standard	1	Boolean feature, denotes that IPFV is a standard gloss
is predetermined	0	Boolean feature, denotes that IPFV is not a predetermined gloss
standard	ipfv	IPFV marked as a standard itself
morphemes	3sg, pst, ipfv	IPFV has co-occurred with 3SG and PST for this particular vector instance
segment matches	f, ipfv, p, pfv	IPFV contains the grams of F, IPFV, P, and PFV
segment non-matches	i, v	IPFV had the strings I and V left over after standard segments were matched

Table 3.1: Example Vector from fra, DEV2: IPFV

Feature Class	**Feature Value(s)**	**Explanation**
segment count	6	IPFV's count of both segment matches and non-matches
distances	2=4, nom=3, imp=2, ipfv=0, prs=3, ...	IPFV's Levenshtein edit distances to all glosses in the standard set
count	25	IPFV had 25 occurrences in DEV2, fra
length	4	IPFV has a length of 4 characters
vowels	1	IPFV has one vowel, the letter I
lexical match	0	IPFV does not a lexical entry in the lexicon
upper case	1	IPFV on input occurred in all upper case

A small selection of features from Table 3.1 are displayed in Example 14:

(14)
```
gloss_ipfv = 1
is_standard = 1
is_predetermined = 0
shared_morpheme_3sg = 1
distance_nom_3 = 1
count_25 = 1
```

The Chi-square distribution test was applied to the features presented above in attempt to determine which features should be selected for the final model. This was an important endeavor because model performance varied widely depending on which features were included. Excluding many features led to better performance. Unfortunately the sheer number of distance features rendered the results of the Chi-square less than helpful. Even with removing distance features the

Chi-square distribution did not help determine statistically significant features. Instead multiple runs were performed that tested different combinations of features in the training set. The features that were chosen resulted in the highest observed accuracy on the training data. The final set of features includes the gloss itself, nearby glosses within a window size of two glosses, whether the gloss is a standard gram, whether the gloss is contained within the predetermined set, morphemes, segments, count, lexical entry/predetermined match, and case.

3.1.3 Classification Model

The model uses machine learning algorithms to classify each of the feature vectors. The standard classification label set includes all of the grams in the standard set of Appendix A and the classification types of §4.2.1 except standard, misspelled, and confused. This is not a sequence labeling task. Languages do not share a common structural order and thus sequence models are not hypothesized to be cross-linguistically useful.

The classification model will intake each feature vector which represents a gloss instance and assign one of the classification labels to the feature vector as output. The standard classification label set mixes both standard grams and classification types. The following explains why this is motivated. For glosses that are indeed grams the model will classify them by the actual gram. There is no classification label of *standard*. Misspelled and confused glosses similarly do not have a classification label. The model will instead try to assign one of the standard grams to misspelled or confused glosses. For standard, misspelled, and confused glosses, the classification model can classify their vectors with their standard gram. The classification model itself is not the appropriate tool to convert lexical entries to their correct grams, delimit combined grams, standardize a user's gram, recover difficult grams, or standardize part-of-speech and lexical items. These scenarios represent the non-standard types of incomplete, combined, user-identified, unrecovered, part-of-speech, and lexical entry found in §4.2.1. Instead, the model will assign classification labels to glosses it believes belong within each of these types. A post-processing system will then review all feature vectors who have one of these non-standard type labels for further refinement. Post-processing is described in the following section §3.3.1.

This study considered four existing machine learning methods; Naïve Bayes, kNN (k-th Nearest Neighbors), Maximum Entropy [Berger et al., 1996] , and Transformation Based Learning (TBL) [Brill, 1995]. The Naïve Bayes algorithm includes two implemented methods: binary and multinomial. kNN includes two implemented methods: euclidean distance and cosine. The k-th neighbor threshold for this project was set at ten. Tie-breaking is accomplished recursively. If a tie arises, the system will continuously pick a next neighbor until the tie is resolved. The Maximum Entropy method was applied with both general iterative scaling and improved iterative scaling. I developed the code for the Naïve Bayes, kNN, and TBL algorithms but utilized the Natural Language Toolkit (NLTK) [Loper and Bird, 2002] for producing the Maximum Entropy model.

The classification stage optionally produces model and system files so that the developer can identify how the learning method made its decisions. Accuracy files are produced to document how each individual machine learning algorithm performs. The evidence from these files indicated that TBL outperforms the other models by a wide margin. In early development an interpolation of models produced a better result than any method on its own. As development matured interpolation only hindered the results. The final model only utilizes TBL.

The actual output and system goal is not the classification itself but the final standard gram set for each language. To achieve this outcome the focus shifts from individual IGT gloss instances to unique glosses for each language.

3.2 Merge Unique Gloss Classifications

Returning to the example of Russian 2SG there were eight instances of 2SG occurring as a gloss and consequently there were eight 2SG feature vectors. These eight instances are considered one unique gloss. The unique gloss will receive the classification label which occurs most often among the classification results of its feature vectors. The classification label for unique gloss, after post-processing for non-standard types, will become a part of the standard gram set for the language.

The following text documents the theoretically ideal process which would interpolate the results of each of the four learning algorithms. For each of the learning algorithms which have more than one interpretation, their classification results are averaged first before being interpolated with

the other learning methods. Table 3.2 illustrates an example of how Naïve Bayes results are averaged. The sum of the classification results for each classification label are averaged and then ordered. The maximum weighted classification is the most probable.

Classification Label	**Method 1**	**Method 2**	**Distribution Weight**
imp	7	6	0.6500
ipfv	2	1	0.1500
combined	1	0	0.0500
incomplete	0	1	0.0500
lexical entry	0	1	0.0500
pfv	0	1	0.0500
Naïve Bayes incorrectly assigns the IMP classification to IMP (it is actually IPFV)			

Table 3.2: Naïve Bayes Interpolation of IMP

Once results for all feature vectors have been averaged within each learning algorithm they are combined by weighted interpolation. This produces a final distribution for the unique gloss for the entire language. The distribution that each learning algorithm produces showcases which potential standard gloss or non-standard type it assigned for all occurrences of the gloss input. The learning algorithm weights were at first evenly distributed as Table 3.3 exemplifies. But due to the superior performance of TBL interpolation weights were shifted so that TBL has a weight of 1.0 and the other methods have a weight of 0.0.

The classification label with the highest score for the unique gloss is then selected. If the final classification is a standard gram then it will automatically be added to the language's standard gram set. If the final classification is a non-standard type the gloss will be post-processed.

Classification Label	**kNN**	**Maximum Entropy**	**Naïve Bayes**	**TBL**	**Total**
Weight	0.25	0.25	0.25	0.25	1.00
ipfv	0.6000	0.8000	0.1333	1.0000	0.6333
imp	0.1000	0.1000	0.6667	0.0000	0.2167
lexical entry	0.1000	0.1000	0.1000	0.0000	0.0667
combined	0.1000	0.0000	0.0333	0.0000	0.0333
incomplete	0.1000	0.0000	0.0333	0.0000	0.0333
pfv	0.0000	0.0000	0.0333	0.0000	0.0083
Interpolation correctly assigns the IPFV classification to IMP					

Table 3.3: Classification Interpolation of IMP

3.3 Produce Language Gram Sets

The final steps in the gloss mapping procedure are to post-process the merged unique gloss outputs and to reference equivalent glosses. The system will output standard gram sets and reference data structures for each language.

3.3.1 Post-Processing

The post-processing stage takes classification outputs from the machine learning stage and assembles a final set of grams for each language. Post-processing procedures vary by non-standard type. If the classification is *incomplete* or *combined* the script will assign the appropriate standard grams for the gloss. A lookup table of predetermined glosses is used to convert glosses with the incomplete classification to grams. Combined glosses are split based on the most likely boundaries. These boundaries are determined from the best matches of the segmentation stage which assigned feature values. With the Language CoLLAGE data this approach has correctly identified standard grams for all incomplete and combined glosses. Glosses which have been classified by the model

as *user-identified* will automatically be added as grams to the language set as they were spelled in the input. Classifications of *unrecovered*, *part-of-speech*, and *lexical entry* are stored as they are spelled at input. Their storage involves separate data structures so the end user may retrieve them but they will not be added to the standard gram set for the language. They are stored in case other projects would benefit from having an observed list of unrecovered glosses, part-of-speech glosses, or lexical entries.

3.3.2 Referencing

Referencing is an important procedural output of the gloss mapping. The misspelling and confusion errors may appear in other documents important for evaluation or system interoperability such as choices files. Referencing maps gloss inputs to standardized post-process outputs. When other sources of data contain non-standard glosses the referencing can identify that the non-standard gloss matches a standard gram. This only functions for non-standard glosses which have been observed by the model. Referenced glosses are stored in containers called centroids. A centroid contains a pair of equivalent Leipzig and GOLD grams from the standard gram set and all of the input glosses which were classified as one those two grams. This is a language-specific process to prevent pairs of separate grams from being considered equivalent with each other. In the Russian example *IMP* was incorrectly annotated. The IPFV centroid will then show the following information in Example 15.

(15) Leipzig: ipfv
GOLD: imperfective
References: {rus: imp, ...}

This could cause potential problems if Russian had true IMP grams in its IGT data from the Language CoLLAGE. Chapter 6 analyzes this issue and suggests future improvements to the *confused* classification to resolve this.

3.4 Inference of Tense, Aspect, and Mood Choices

The inference procedure loads the reference centroids which were prepared during the post-processing stage §3.3.1 and the list of standard grams with their suggested categories. For the entire list please consult Appendix A.

3.4.1 Category Assignment

The main methodology for inference is to update and rank the suggested categories for each gram. Recall that a category is a linguistic system such as tense, aspect, or mood. Based on the training data and model, frequencies are built which weight how often each category is assigned to each gram. These are Laplace smoothed by adding a count of 1 to each of the suggested categories that were loaded for each gram. This allows categories that could represent the gram to still be possible even if the category did not appear with the gram in the training data. Additional conditional probabilities assess how often a category is assigned to a gram based on other category and gram pairs for the language. This attempts to address how likely gram X is category Y given a distribution of other gram and category pairs. A global search algorithm then assigns categories to grams based on what gram and category pairs are the most probable given the conditional probabilities between pairs and the likelihood that gram X is category Y. This is done at the language level. In practice this has only improved one example of Habitual changing it from aspect to tense. Otherwise the most common category for a gram is typically the only observed category for that gram. Simple assignment of the most common category for the gram would therefore by sufficient at assigning categories to grams. A previous methodology attempted this process with a more theoretical motivation; refer to Appendix C for information about this other inference approach.

As a default, the system will select the GOLD gram for each gloss and pair that with the category output of the position class inference. The user may toggle the output to the corresponding Leipzig gram or the most common IGT gloss based on observations in the DEV1 and DEV2 datasets.

At this point all grams have been assigned to categories such as tense, aspect, and mood. Infer-

ence itself is complete but in order to be useful for the AGGREGATION tools it must be written to a choices file. Recall from §2.2 that a choices file contains the linguistic properties of a language. This format is used by the Grammar Matrix customization system to produce a computational grammar. Writing choices files then enables the creation of computational grammars.

3.4.2 Choices File Creation

The choices file procedure will first write generic information to note the presence of the tense, aspect, and mood section. For tenses the script uses radio button equivalences. If non-past, past, present, or future is found they will be written with `gram=on`. This is important for tenses due to the suggested hierarchies for them which are predominantly past, present, and future focused [Poulson, 2011]. The script will write each remaining tense, aspect, and mood choice with the corresponding category and gram. This will match the choices file style with incrementing digits for each successive choice in the category. Example 16 provides an example choices file system output for Shona's tense, aspect, and mood.

(16)
```
section=tense-aspect-mood
tense-definition=choose
past=on
 past-subtype1_name=REC
 past-subtype2_name=REM
present=on
future=on
 aspect1_name=HAB
  aspect1_supertype1_name=aspect
     aspect2_name=CONT
  aspect2_supertype1_name=aspect
```

The Shona choices file in Example 16 has been reordered to assist the human readability of the choices file. The outputs do not necessarily keep subtypes with their supertype. The script

as programmed will keep tense, aspect, and mood choices together within their category but the hierarchy may not be in order. With this file the subtypes originally appeared after future. This does not affect the Grammar Matrix customization system's interpretation of the choices file. The hierarchy may not be complete. This study did not scope a method for extracting and building tense, aspect, and mood hierarchies.

3.5 Summary

This methodology is an approach for mapping input glosses to a standardized set and using outputs to infer grammatical properties. The data extraction process collects the data needed to build vectors. Features attributes from these vectors are identified and encoded to allow for machine learning input. I first programmed the model to interpolate many well-recognized natural language processing algorithms to classify input glosses. TBL was the most effective machine learning algorithm and I decided to make it the only algorithm informing the model. The classification results are both standard grams and non-standard types. Post-processing combines individual classification results and refines them depending on the final classification label. Reference centroids are created for evaluation purposes. Then inferencing uses a global search algorithm to match tense, aspect, and mood categories to mapped grams. Once these grams and categories are inferred choices files are generated. These choices files are compatible with the LinGO Grammar Matrix and can produce computational grammars with tense, aspect, and mood properties.

Chapter 4

EVALUATION

I performed an evaluation to assess both the performance of the gloss mapping and of the inference of tense, aspect, and mood. These are two separate tasks. The first one determines how successful the gloss mapping is at standardizing glosses. The second evaluation is used to determine if gloss mapping has an impact on inference procedures that rely on grams from the gloss line.

4.1 Original Datasets

The Language CoLLAGE [Bender, 2014b] DEV1 and DEV2 datasets were used as training data for the machine learning algorithms. The final model was thus trained on DEV1 and DEV2. Table 4.1 contains a list of all languages within each dataset. The IGT column refers to the number of IGT instances for that language. The TAM Choices column indicates whether that language has a corresponding choices files with tense, aspect, and mood choices. Languages without tense, aspect, and mood will not be evaluated for inference but they can still be evaluated for gloss mapping.

Table 4.1: Language CoLLAGE Datasets and Languages

Dataset	ISO 639-3	Language	IGT	TAM Choices
dev1	ang	English	106	False
dev1	hau	Hausa	20	False
dev1	isl	Icelandic	16	False
dev1	jpn	Japanese	68	True
dev1	jup	Hupdë	46	True

Table 4.1: Language CoLLAGE Datasets and Languages

Dataset	ISO 639-3	Language	IGT	TAM Choices
dev1	mnk	Mandinka	109	False
dev1	nan	Nan Chinese	61	False
dev1	pbt	Pashto	251	True
dev1	rus	Russian	166	True
dev1	sna	Shona	91	True
dev2	bre	Breton	75	True
dev2	cym	Welsh	97	False
dev2	fra	French	228	True
dev2	lut	Lushootseed	86	True
dev2	mal	Malayalam	11	False
dev2	ojg	Ojibwa	97	True
dev2	qub	Quechua	96	False
dev2	sci	Sri Lankan Malay	87	True
dev2	tam	Tamil	65	True
dev2	zul	Zulu	49	False
test	ain	Ainu	76	True
test	ary	Moroccan Spoken Arabic	97	False
test	ces	Czech	75	False
test	hbs	Serbian	96	True
test	hix	Hixkaryána	41	True
test	inh	Ingush	127	True
test	jaa	Jamamadí	216	True
test	kat	Georgian	62	True
test	tgl	Tagalog	33	False

Table 4.1: Language CoLLAGE Datasets and Languages

Dataset	ISO 639-3	Language	IGT	TAM Choices
test	tha	Thai	14	True
test	vie	Vietnamese	128	True

4.2 Gold Standard

The input IGT glosses from the Language CoLLAGE datasets contain many non-standardized glosses. In order to evaluate the effectiveness of mapping glosses a gold standard needed to be developed for every gloss in the input IGT. This gold standard includes a few components. The first is an identification of the classification type which categorizes different archetypes of standard and non-standard glosses. The second is the gold classification label which is the label that the model should predict. The standard classification label set includes all 113 standard grams in Appendix A and 6 of the non-standard classification types. The third gold standard parameter is a list of standard grams that represent the gloss. There will only be one standard gram in this list unless the gloss is a combined type. This provides a standard to measure post-processing which is needed to assign grams to certain non-standard types after model classification. Annotation of the data was facilitated by scripts that enabled more efficient review of the input IGT glosses. These scripts generated reports which provide the evaluation baselines.

4.2.1 Classification Types

This project established classification types to precisely define the different contexts of when a gloss is not in the standard set. All classification types except standard may also be called non-standard types. The identified gloss non-standard types include misspelled, confused, incomplete, combined, user-identified, unrecoverable, part-of-speech, and lexical entry. A gloss that is already

standard will be annotated as standard.

1. Misspelled grams occur where a gloss, typically Leipzig, is correctly identified in context but has been misspelled in that it does not match the exact spelling of a standard gram.

   (17) Input Gloss: IMPF
   Gold Classification Type: Misspelled
   Gold Classification Label: IPFV
   Gold Grams: IPFV
   Explanation: Imperfect (IPFV) was incorrectly spelled as IMPF. Its classification label is the gram IPFV. Misspelled is not a classification label and thus does not have a post-process stage. Its gold gram is the same as the classification label. Misspelled glosses do not trigger post-processing

2. Confused grams occur when one gram was substituted for another. Usually these are Leipzig grams.

   (18) Input: IMP
   Gold Classification Type: Confused
   Gold Classification Label: IPFV
   Gold Grams: IPFV
   Explanation: The IGT author meant IPFV (Imperfect) but glossed IMP (Imperative). Its classification label is the gram IPFV not IMP because the model should target IPFV. Confused is not a classification label and thus does not have a post-process stage. Its gold gram is the same as the classification label. Confused glosses do not trigger post-processing.

3. Incomplete refers to lexical entries that appeared in the gloss line that should have been glossed or lexical entries that were only partially glossed. Most often these are pronouns or demonstratives.

(19) Input: I
Gold Classification Type: Incomplete
Gold Classification Label: Incomplete
Gold Grams: 1SG
Explanation: The IGT author did not convert the pronoun *I* to its standard gram(s). Its classification label is Incomplete which is how the model should classify the gloss *I*. The gold gram for *I* is 1SG and post-processing should map *I* to the 1SG gram.

4. Combined refers to multiple glosses that should have been separated by a period delimiter.

(20) Input: 1SGPRF
Gold Classification Type: Combined
Gold Classification Label: Combined
Gold Grams: 1SG, PRF
Explanation: The IGT author did not delimit the glosses. Its classification label is Combined which is how the model should classify *1SGPRF*. The gold grams for *1SGPRF* are 1SG and PRF. Post-processing takes combined classified glosses and uses segmentation to divide them. It should split *1SGPRF* to *1SG* and *PRF*.

5. User-identified refers to correct glosses for the language that do not occur in the standardized set because they are specific to that language.

(21) Input: CL1
Gold Classification Type: User-Identified
Gold Classification Label: User-Identified
Gold Grams: CL1
Explanation: The IGT author created a classifier gloss specific to the language. Its classification label is User-Identified which is how the model should classify CL1. The gold gram for CL1 is itself. Post-processing will accept user-identified glosses

as-is into the standard gram set for the language. Part of the post-processing is to build data structures that identify which standard grams are user-identified.

6. Unrecoverable pertains to glosses that are not readily identified as a standard gram (regardless of whether it is standard, misspelled, confused, incomplete, combined, or user), part-of-speech, or lexical entry.

   (22) Input: PN
   Gold Classification Type: Unrecovered
   Gold Classification Label: Unrecovered
   Gold Grams: None
   Explanation: The gloss is neither recognized as a standard or any other non-standard type including part-of-speech glosses or lexical entries. Its classification label is Unrecovered which is how the model should classify PN. The gold gram is None which represents a null empty value. Post-processing will not include unrecovered glosses in the standard gram set for the language.

7. Part-of-speech glosses and lexical entries are documented in order to improve the machine learning process. One of the difficulties with the gloss line is that part-of-speech glosses and lexical entries are included and thus must be classified to avoid attributing them as standard grams.

   (23) Input: ADV
   Gold Classification Type: Part-of-Speech
   Gold Classification Label: Part-of-Speech
   Gold Grams: None
   Explanation: ADV is a part-of-speech and not a gram of interest. Its classification label is Part-of-Speech which is how the model should classify ADV. The gold gram is None which represents a null empty value. Post-processing will not include a

part-of-speech gloss in the standard set of grams for language but place it into a part-of-speech map for the language.

(24) Input: glass
Gold Classification Type: Lexical Entry
Gold Classification Label: Lexical Entry
Gold Grams: None
Explanation: Glass is a lexical entry and not a gram interest. Its classification label is Lexical Entry which is how the model should classify glass. The gold gram is None which represents a null empty value. Post-processing will not include a lexical entry in the standard set of grams for language but place it into a lexicon for the language.

This section hopefully clarifies the discussion in §3.1.3 and §3.3.1 about the difference between a standard gram and a non-standard type with respect to classification labels. A non-standard type is applied as an intermediate label to define the type of non-compliance for post-processing purposes. Additionally these classification types provide a basis for evaluation as §4.2.3 explains.

4.2.2 Annotation Procedure

The annotation procedure facilitates the creation of gold standard classification types, classification labels, and gold grams. Collectively these three gold standard components are called gold labels. The process includes a series of functions that ask the annotator guided questions in order to efficiently standardize the glosses. I personally annotated all of the data from Language CoLLAGE datasets DEV1, DEV2, and TEST.

If the gloss is a standard gram the functions will automatically set the gold classification type to *Standard* and the gold classification label and gold gram to the gloss itself. If the gloss is misspelled or confused the system will prompt the annotator to identify the actual standard gram which will become both the gold classification label and the gold gram. The gold non-standard type will be *Misspelled* or *Confused* respectively.

For incomplete and combined the functions set the gold classification type and gold classification label to *Incomplete* and *Combined* respectively. Then they prompt the annotator to identify what the actual grams should be for the gloss. These become the gold grams. This is important. The gold classification label trains the model to distinguish incomplete and combined types from other non-standard types and standard grams during classification. The gold grams then train the model to convert incomplete and combined classified glosses to their actual gold grams during the post-processing procedure. For a review of post-processing refer to §3.3.1. If the gloss is user-identified the script will set the gold classification type and the gold classification label to *User-Identified* and set the gold gram equal to the input gloss.

For the remaining non-standard types of unrecovered, part-of-speech, and lexical entry the system will assign their gold classification type and their gold classification label to their non-standard type such as *Unrecovered*. They all receive a gold gram of the null value *None*. The *None* value will prevent them from becoming standard grams in any sense.

Part of the efficiency arises from asking the annotator to review each unique gloss for the language once instead of every single gloss instance. Other efficiency measures include automatically assuming a gloss is standard if it matches a gram in the standard set, a gloss is a lexical entry if it has been labeled as a lexical entry in another language already, and a gloss is incomplete if it is a member of the predetermined pronoun and demonstrative list. This is nonetheless a tedious process. Despite the guidance of the scripting, an annotator wanting accurate results will still have to consult the IGT directly to resolve ambiguities or gain greater context for the observed gloss. The annotator will additionally have to review the gold standard output to ensure that all assumptions about standard glosses and lexical entries were indeed accurate. When DEV1 was annotated many of these measures were not in place which resulted in an annotation time of about two hours for the 984 IGT instances. DEV2 and TEST only required about ten minutes to annotate their combined 1866 IGT instances.

This entire process is only important for the evaluation of the methodology. A user trying to obtain a computational grammar will not need to annotate their IGT. That would defeat the purpose of mapping glosses because they would be hand-annotating the results that the map gloss learning

methods should achieve automatically.

An export and load format called CLSO (CLaSs and Object representation) was created specifically for the frequent exporting and loading of objects in the gloss mapping project. A CLSO file for gold labels is produced once the annotator has finished developing the gold standard files for the entire Language CoLLAGE. This preserves the answers that the annotator provided for each gloss so when the dataset is loaded at a different time the user will not have to recreate the gold standard.

4.2.3 Reporting

The following reporting was produced to document the rate of standard glosses for Language CoLLAGE datasets. It includes an aggregate of frequencies by classification type for the dataset. Table 4.2 presents the aggregate results for DEV1, DEV2, and TEST.

	DEV1		**DEV2**		**TEST**		**Total**	
Classification Type	Count	Freq	Count	Freq	Count	Freq	Count	Freq
Standard Gloss	164	16%	181	19%	191	18%	536	18%
Lexical Entry	653	65%	591	62%	739	68%	1983	65%
Incomplete Gloss	83	8%	83	9%	45	4%	211	7%
Misspelled Gloss	54	5%	37	4%	45	4%	136	4%
User-Identified Gloss	28	3%	37	4%	42	4%	107	4%
Combined Gloss	4	0%	10	1%	12	1%	26	1%
Part-of-Speech	13	1%	7	1%	6	1%	26	1%
Unrecovered Gloss	7	1%	2	0%	1	0%	10	0%
Confused Gloss	1	0%	1	0%	0	0%	2	0%
Total Glosses	1007	100%	949	100%	1081	100%	3037	100%

Table 4.2: Gold Classification Type Frequencies

	DEV1		DEV2		TEST		Total	
Classification Type	Count	Freq	Count	Freq	Count	Freq	Count	Freq
Standard Gloss	164	48%	181	52%	191	57%	536	52%
Incomplete Gloss	83	24%	83	24%	45	13%	211	21%
Misspelled Gloss	54	16%	37	11%	45	13%	136	13%
User-Identified Gloss	28	8%	37	11%	42	12%	107	10%
Combined Gloss	4	1%	10	3%	12	4%	26	3%
Unrecovered Gloss	7	2%	2	1%	1	0%	10	1%
Confused Gloss	1	0%	1	0%	0	0%	2	0%
Total Glosses	341	100%	351	100%	336	100%	1028	100%

Table 4.3: Gold Classification Type Frequencies Alternate

Table 4.3 provides an alternate interpretation by removing part-of-speech and lexical entries for configuring the frequencies. The main analysis including part-of-speech and lexical entries is preferred over the alternate frequencies because the model must classify part-of-speech and lexical entries to prohibit them from becoming grams. Lexical items and part-of-speech glosses interpreted as grams would reduce the overall precision of the system's predicted standard grams. An example is the lexical entry *all* which could be confused for *allative* case which has a Leipzig form of *ALL*. There is no perfect marker in the IGT that would indicate whether a gloss is a lexical item or a gram and thus the model must make such identifications.

The most common non-standard classification types are misspellings or incomplete glosses. The Language CoLLAGE [Bender, 2014b] data should be more compliant than other sources such as ODIN [Lewis and Xia, 2010] [Xia et al., 2014]. The students who documented IGT for the Language CoLLAGE were instructed to follow Leipzig [Bickel et al., 2008] as closely as possible which would not be the case necessarily for other linguists developing or annotating IGT. This means that ODIN and other sources likely have higher non-standard rates.

4.3 Evaluation Methodology

Chapter 3 presented four sections of the map glossing and inference procedure. The first section of classifying each gloss is not evaluated but the second section where the classifications by unique gloss are evaluated. This is due to the applicability of the merged classification to the final sets whereas individual classifications that are incorrect and not selected during the merge do not impact the final results. The final language sets and the inference procedure are additionally evaluated.

4.3.1 Merged Classifications

Merged classifications are measured by comparing the system outputs for each unique gloss to its gold standard annotation. This is measured in accuracy as this occurs over all instances; ergo precision, recall, and accuracy are all equal.

Baseline 1 is set at the observed standard rate and lexical entry rate for the TEST dataset. The TEST dataset has a standard rate of 18% and a lexical entry rate of 68% as documented in Table 4.3 which accounts for a total baseline of 86%. Actually, the standard rate across the entire Language CoLLAGE is also 18%. The combined standard and lexical entry rate is used because it makes the assumption that all standard grams should be accepted as-is and all others should be assigned as lexical entries.

Baseline 2 assumed a unigram model where glosses seen in the training set were assigned the most common gold value and unseen glosses were either marked as a lexical entry or if they were in the standard set their input value was assumed to be their gold value. The unigram model performs very well at 91%. This is 5% higher than the initial assumptions of **Baseline 1** because some of the non-standard inputs in DEV1 and DEV2 better inform the TEST model to make correct assignments.

4.3.2 Language Gram Sets

Language gram sets represent the set of grams the model believes a language contains based on its IGT. An assumption is that output glosses from map gloss would be useful in other applications

such as inference. The inference section will directly test this assumption. However it is important to first determine how well gloss mapping performs at developing language gram sets which contain only the glosses which occurred for the language in its IGT.

For each dataset a report is produced of map gloss classification and post-process performance which contains an overall accuracy figure, a breakdown of machine learning performance by unique vector, and a series of files called CPRF (Comparative Precision, Recall, and F-score). A CPRF compares two confusion matrices. To understand what each class of the confusion matrix means for gloss mapping consult Table 4.4. The first confusion matrix is the baseline. This is **Baseline 3** which is a language gram set created by assuming gloss inputs as-is. The second confusion matrix is the final language gram set after map glossing and post-processing.

Class	**Explanation**
True Positive	A gloss that the system declares is a standard gram and is indeed in the set of standard grams for the language.
False Positive	A gloss that the system declares is a standard gram but it is not in the set of standard grams for the language.
False Negative	A gloss that the system does not accept as a standard gram but is actually in the set of standard grams for the language.
True Negative	A gloss that the system does not detect and is indeed not in the set of standard grams for the language. This is ignored in analysis because it does not impact precision or recall.

Table 4.4: Language Gram Set Confusion Matrices Class Explanations

After obtaining confusion matrices for both **Baseline 3** and map glossing they are then compared. Table 4.5 provides an explanation of the four classes contained in a CPRF file and how to interpret them in the context of gloss mapping. True Positives to False Negatives and Added False Positives display what the baseline correctly achieved that the gloss mapping did not. Conversely, False Negative to True Positives and Removed False Positives illustrate how gloss mapping improved over the baseline. To view the actual results consult Chapter 5 and more specifically Tables 5.2 and 5.4.

Class	**Explanation**
True Positive to False Negative	Baseline True Positives are standard grams as input glosses and in this class they were not classified to have the same standard gram and become False Negatives.
False Negatives to True Positives	Baseline False Negatives are non-standard grams input glosses and in this class they were classified correctly to the appropriate standard gram and become True Positives.
Added False Positives	Mapped False Positives are standard grams that have been added to the final language set that were not in the original baseline set and are not actually members of the language's standard gram set.
Removed False Positives	Baseline False Positives are confused non-standard glosses at input which the classification correctly removes to prevent them from appearing in the language's final set.

Table 4.5: Language Gram Set Comparison Class Explanations

4.3.3 Inference

Inference refers to reviewing data to automatically generate choices files with learned grammatical properties for a language. Inference focuses on mapping standard grams for a language to categories such as tense, aspect, and mood.

Several baseline options were explored. A random method would need to have some probabilistic distribution of glosses or at the very least a limited number of gloss probabilities for tense, aspect, and mood categories. This is not feasible due to the large amount of tense, aspect, and mood grams and limited data in the Language CoLLAGE which would produce very sparse statistics. Two baseline options were developed that do not rely on such stochastic or statistical distributions.

Baseline 4 uses the most frequent tense, aspect, and mood gram and category pairs. A gram and category pair is a unique combination of a standard gram and a category such as *Habitual & aspect* which is different than *Habitual & tense*. Frequency was determined by sorting observed gram and category pairs from the Language CoLLAGE DEV1 and DEV2 dataset in descending frequency. The five most frequent tense, aspect, and mood grams are in order: *Past & tense*, *Future & tense*, *Present & tense*, *Perfective & aspect*, and *Imperfective & aspect*. **Baseline 4** sets all languages as having these five and only these five choices for their tense, aspect, and mood.

Baseline 5 helps assess the comparative efficacy of both the inference method and the gloss mapping. In this baseline, the inference procedure uses all standard grams from the language's input gloss line which is similar to **Baseline 3**. The tense, aspect, and mood categories are then determined by frequency in Language CoLLAGE. The expectation is that the script will only infer exact gloss matches and perhaps false positives from lexical entries that resemble glosses.

Evaluation uses the reference centroid mapping developed during the post-processing stage §3.3.1 to match equivalent choices values together. It is quite common that the gold choices file will describe a choices value under one name and the IGT will have an equivalent gloss with a different name. When the inferred choices files are produced their outputs will be based upon the GOLD version of the gloss in the IGT which may not match the convention of the gold choices files.

The centroids provide a mapping between a standard gram Leipzig and GOLD pair and all equivalent unprocessed glosses. This enables matching choices values that represent the same linguistic property but are not exact string matches. Example 15 of §3.3.1 shows a centroid for Russian where *IPFV*, *Imperfective*, and *IMP* are equivalent. This allows the inferred choices file containing the choice pair Imperfective & aspect to match the gold choices file regardless of whether it contains imperfective as either of those three references. It is important to note that a choice is only considered equivalent if both the choices value and category match the gold standard. So any of the Imperfective strings detailed in Example 15 would be accepted as equivalent to the gold standard as long as their category, in this case aspect, also matches.

Both **Baseline 4** and **Baseline 5** are compared against the final inferred outputs which were produced after the implementation of the gloss mapping. The inferred outputs are rated according to how well they match the gold standard choices file for the language based solely the alignment of choices. The AGGREGATION [Bender, 2014a] ChoicesReader script extracts tense, aspect, and mood choices for the gold, baseline, and inferred choices files. Each choice in the gold choices file is then compared against each choice in the baseline and inference sets. For each inference choice all of its reference centroids are checked. Baseline choices are not checked against the centroids because centroids are a part of the gloss mapping implementation. Consequently baseline choices will only match choices values that are standard grams and in the IGT. If a match results a True Positive is counted. If no match is found then a False Negative is reported. Once all of the gold standard glosses have been read, all of the baseline and inference choices that have not been matched are reported as False Positives. A confusion matrix is built for these results.

Precision, recall, and F-scores are then built for both the inference and the baseline confusion matrices for each language. The results are then passed on to CPRF files which show a comparative evaluation of how the inference procedure compares against both baseline versions. For a review of the CPRF file classes consult Table 4.5.

4.4 Summary

Evaluation uses a series of classification types to report the rate of non-standard input glosses and to compare baselines to inferred outputs. These are the same classification types that the machine learning model utilized to classify results. The initial gold standard revealed that both the TEST dataset and the entire Language CoLLAGE data have a standard rate of 18%. The TEST dataset has a combined rate of standard grams and lexical entries of 86% which sets one of baselines for the merged classification. The other classification baseline is produced from a unigram model which mapped glosses at a rate of 91%. A third baseline is set for comparing the final language sets of standard grams produced by the map glossing model. Two inference baselines were developed. One reflects frequent gram and category pairs and the other assumes input glosses as-is. A framework was presented for comparing language gram sets and inference choices against their baselines which involves comparing precision, recall, and F-scores against each of their gold standards. Chapter 5 presents the actual performance statistics for the data used in this study.

Chapter 5

RESULTS

The following results reveal that the gloss mapping methodology performed very well. The inference results suggest that while mapping glosses does improve performance, it appears using glosses from IGT to infer linguistic properties may not consistently produce desirable results.

5.1 Merged Classifications

The dataset of interest for all results is TEST. DEV1 and DEV2 were used to train the model. TEST was a held-out set that was used to evaluate the model and is considered the performance metric for gloss mapping. DEV1 and DEV2 have only been included for discussion but they do not represent model performance.

Table 5.1 displays the merged classification accuracies produced by the model for each of the datasets. DEV1 and DEV2 have near perfect accuracies because they were used to train the model. The only instances where DEV1 and DEV2 failed to produce the correct classification was for instances of *IMP* in which cases the model predicted that *IMP* was confused for *IPFV*. The model accuracy for the TEST dataset is 94%. This is higher than **Baseline 1** of 86% and **Baseline 2** of 91%.

The system does quite well at identifying standard grams, lexical entries, and glosses classified as incomplete. The model often misclassifies misspelled glosses as lexical entries, user-identified, or as the wrong standard gram. The model classified combined glosses most often as words or user-identified. The only part-of-speech misclassifications were lexical entries and user-identified as well. User-identified performs surprisingly well. This category would be expected to perform very poorly because the model is very unlikely to observe any user-identified grams before. But yet it somehow correctly classifies user-identified for 75% of the occurrences. The other 25% were

classified as lexical entries. This altogether suggests that further feature development should focus on precisely identifying lexical entries and user-identified so that the model does not frequently classify other glosses as one of these two classification types.

	DEV1			DEV2			TEST		
Classification Type	Correct	Total	%	Correct	Total	%	Correct	Total	%
Standard Gloss	160	164	98%	179	181	99%	186	191	97%
Lexical Entry	617	617	100%	570	570	100%	706	719	98 %
Incomplete Gloss	83	83	100%	81	81	100%	42	44	95%
Misspelled Gloss	52	52	100%	37	37	100%	15	41	37%
User-Identified Gloss	24	24	100%	31	31	100%	30	40	75 %
Combined Gloss	4	4	100%	10	10	100%	3	12	25 %
Part-of-Speech	13	13	100%	7	7	100%	4	6	67%
Unrecovered Gloss	7	7	100%	2	2	100%	0	1	0%
Confused Gloss	1	1	100%	1	1	100%	0	0	-
Total	961	965	100%	918	920	100%	986	1054	94%

Table 5.1: Merged Classification Accuracies

5.2 Language Gram Sets

Language gram sets are measured using precision and recall. Precision reveals the rate of how many grams in the final set actually belong in the final set. Recall indicates the rate of how well the model identifies all of the grams that should be in the final set. Precision and recall results for language gram sets are presented in two tables.

The first results in Table 5.2 provide a comprehensive overview about how gloss mapping performed by dataset. This is the preferred performance measure for the map gloss system because it was trained on DEV1 and DEV2 and was tested on TEST which was the true held-out set. This

also means that these results are not averages of from the above data because the cross-validation rotates which languages act as training and testing.

The results for TEST demonstrate that mapping glosses has significantly improved their precision and recall over the baseline. Though the TEST results still could benefit from refinement in future studies they do illustrate that mapping glosses to a standard set is achievable at high rates of precision and recall.

Metric	**DEV1**	**DEV2**	**TEST**
Baseline 3 Precision	17%	20%	18%
Baseline 3 Recall	71%	73%	76%
Baseline 3 F-Score	27%	31%	29%
Final Precision	98%	100%	97%
Final Recall	98%	99%	92%
Final F-Score	98%	99%	94%

Table 5.2: Language Gram Sets Comparative Evaluation

The second precision and recall results are in Table 5.3 which displays leave-one-out-cross-validation scores of the final TBL model. Leave-one-out-cross-validation for this study means that one language was left out while the glosses from all of the other languages were used to train the model. Then the glosses from the language that was left out are tested on the model. This is repeated for each language. This was motivated because there are not many languages in the Language CoLLAGE. Thus it presents a better representation about how well the model would perform if it could use all of the data in the Language CoLLAGE to train the model.

The first columns called Baseline 3 follow the **Baseline 3** measure in §4.3.3 which is produced by comparing input glosses against the gold standard. The baseline precision approximates the language's standard gloss rate. This is intuitive because the precision reflects the rate of how frequent correct grams occur in all of the outputs.

The TEST column shows the leave-one-out-cross-validation scores. Each row represents a language. To produce results for each row, all of the other languages trained the model and once the model was trained that language was tested on that model. All languages in the DEV1 and DEV2 datasets have perfect recall. This means that all glosses in the DEV datasets occurred in at least one other language in a similar enough context for the model to correctly classify them or they were already standard and the model classified them correctly by default. Many TEST dataset languages contain glosses that only apply to the specific language. This is why TEST does not have perfect recall. Upon inspecting data, the false negatives which decreased recall were due to misspellings with a lack of occurrence in any other language.

Table 5.3: Language Gram Sets Leave-One-Out Cross Validation

Language		**Baseline 3**			**TEST**		
Dataset	ISO	Prec.	Recall	F-Score	Prec.	Recall	F-Score
dev1	ang	10%	73%	17%	92%	100%	96%
dev1	hau	15%	47%	23%	94%	100%	97%
dev1	isl	28%	71%	40%	88%	100%	93%
dev1	jpn	24%	77%	37%	92%	100%	96%
dev1	jup	19%	67%	30%	96%	100%	98%
dev1	mnk	10%	67%	17%	88%	100%	93%
dev1	nan	6%	62%	11%	100%	100%	100%
dev1	pbt	25%	81%	39%	89%	100%	94%
dev1	rus	28%	68%	40%	92%	100%	96%
dev1	sna	19%	79%	31%	89%	100%	94%
dev2	bre	31%	81%	45%	100%	100%	100%
dev2	cym	12%	60%	20%	96%	100%	98%
dev2	fra	23%	78%	35%	96%	100%	98%
dev2	lut	29%	70%	41%	96%	100%	98%

Table 5.3: Language Gram Sets Leave-One-Out Cross Validation

Language		**Baseline 3**			**TEST**		
Dataset	ISO	Prec.	Recall	F-Score	Prec.	Recall	F-Score
dev2	mal	16%	55%	25%	92%	100%	96%
dev2	ojg	16%	67%	26%	96%	100%	98%
dev2	qub	22%	79%	34%	100%	100%	100%
dev2	sci	14%	69%	23%	94%	100%	97%
dev2	tam	30%	80%	44%	94%	100%	97%
dev2	zul	17%	79%	29%	82%	100%	90%
test	ain	22%	73%	34%	100%	100%	100%
test	ary	7%	62%	13%	83%	96%	89%
test	ces	36%	92%	52%	92%	100%	96%
test	hbs	35%	90%	50%	100%	100%	100%
test	hix	16%	79%	26%	100%	95%	97%
test	inh	10%	67%	17%	97%	100%	99%
test	jaa	21%	79%	33%	100%	100%	100%
test	kat	21%	67%	31%	96%	100%	98%
test	tgl	46%	75%	57%	100%	100%	100%
test	tha	0%	0%	0%	100%	100%	100%
test	vie	24%	83%	38%	94%	100%	97%

5.3 Inference

The inference procedures do not highlight as much change in performance. The first baseline results in Table 5.4 refers to **Baseline 4** which assumes the five most frequent choices in the Lan-

guage CoLLAGE. These choices of past, present, future, perfective, and imperfective inherently perform well because they frequently appeared in the DEV1 and DEV2 datasets. From this baseline only precision improved while recall significantly declined. The second baseline of **Baseline 5** is comprised of all input glosses which match a standard gram.

The results do suggest mapping glosses improves inference results. **Baseline 4** represented TEST well whereas **Baseline 5** shows that the TEST glosses on input produce poor inference results. The final map gloss system does improve on precision over both baselines and recall on the second baseline. The TEST has lower recall than the baseline which is mostly a product of choices files specifying grammatical properties that were not glossed in the IGT.

These results are not produced using leave-one-out-cross-validation. All inference results were produced using the DEV1 and DEV2 sets as training data for the model. TEST was applied only after the system was developed and is used for testing the model only.

Metric	**DEV1**	**DEV2**	**TEST**
Baseline 4 Precision	72%	53%	65%
Baseline 4 Recall	21%	41%	30%
Baseline 4 F-Score	32%	46%	41%
Baseline 5 Precision	74%	59%	55%
Baseline 5 Recall	16%	41%	13%
Baseline 5 F-Score	26%	48%	21%
Final Precision	88%	66%	82%
Final Recall	24%	49%	21%
Final F-Score	38%	56%	33%

Table 5.4: Inference Comparative Evaluation

5.4 Result Visualization

The following scatter plots present the same information as the above tables in a graphic form. They compare precision and recall scores. Figure 5.1 displays the strong gap between baseline for map gloss and the final map gloss results. Figure 5.2 shows how inferior the baseline is against the results of cross-validation map glossing. Figure 5.3 illustrates how varied the data is for inference results although it is still clear that final results are a marginal improvement over the baselines albeit not by much.

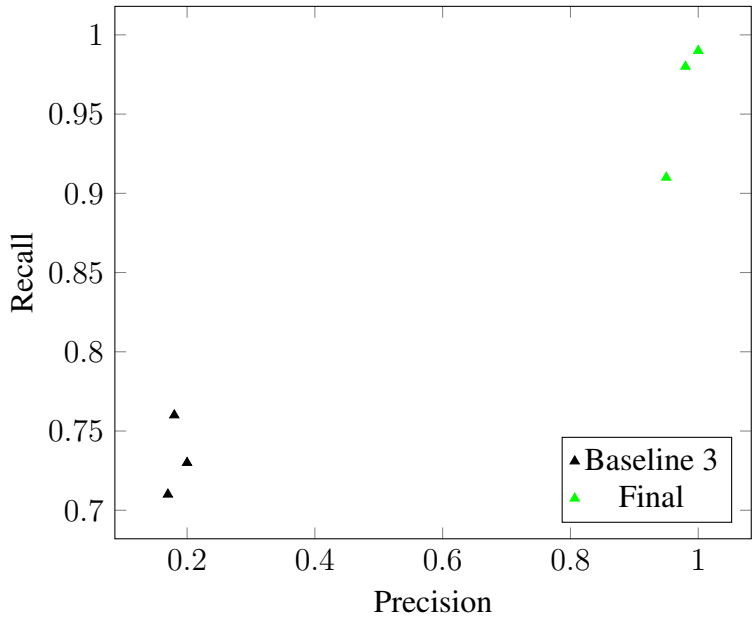

Figure 5.1: Language Gram Sets Comparative Evaluation Scatter Plot

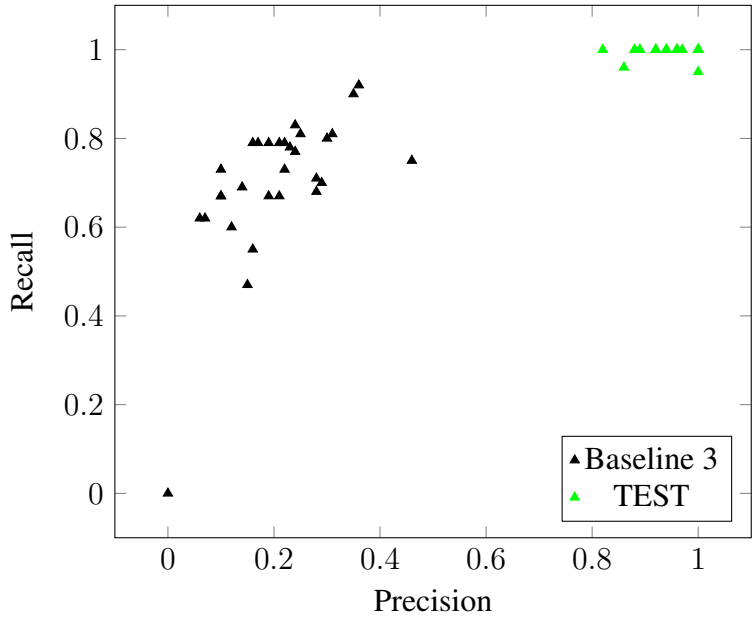

Figure 5.2: Language Gram Sets Leave-One-Out-Cross-Validation Scatter Plot

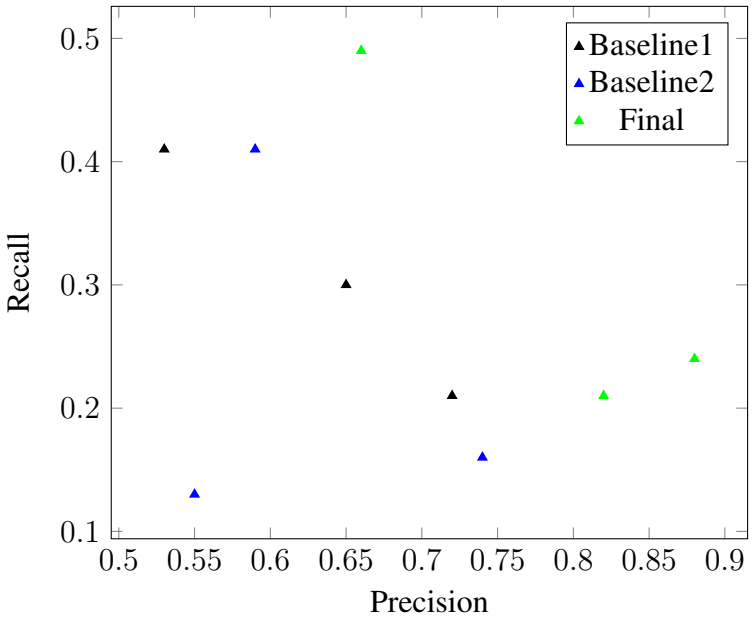

Figure 5.3: Inference Comparative Evaluation Scatter Plot

5.5 Summary

The results presented herein document the performance of the model in mapping glosses and inference tense, aspect, and mood choices from mapped grams. The machine learning component which classifies grams achieves a higher accuracy than the baseline figure. The final map gloss results by language perform better than their baselines regardless of whether leave-one-out-cross-validation or a traditional held-out evaluation framework is applied to model training and testing. The inference has some mixed results. The first baseline which makes assumptions based on observed choices performs better at recall than the actual map glossing method. However map glossing does successfully improve inference over a baseline of accepting glosses as-is.

Chapter 6

DISCUSSION

The map glossing aspect of the project produced very positive results. It was not only possible to achieve a high precision and recall improvement for gloss mapping but those improvements led to greater inference scores. Some errors still persist though that merit explanation. There are additional deficiencies in the methodology that may have diluted the final precision and recall scores.

6.1 Methodology

During the development of the features and machine learning algorithms it became clear that only TBL performed well. Naïve Bayes and kNN were not even producing half of the accuracy that TBL provides. Maximum Entropy was wildly inconsistent in its results based on what features were used. Its results were consistently inferior to TBL and did not interpolate well with the other methods.

Fortunately, TBL does perform very well even in testing. Unfortunately, TBL uniquely suffers from inherent overfitting in small datasets with limited features. The best example of this is with *IMP* and *IPFV*. The author of the Russian IGT documented all instances of imperfective as *IMP*. In the gold standard these vectors are appropriately labeled as confused because imperfective is actually *IPFV* in the Leipzig standard [Bickel et al., 2008]. There are several other languages with *IMP* in their datasets that actually mean imperative. However, Russian has substantially more IGT instances of *IMP* meaning imperfective than the other languages do of *IMP* meaning imperative combined. The final results then are to misclassify all *IMP* instances as *IPFV* which reduces the total vector misclassifications. While this does reduce the number of vector misclassifications it increases the rate of errors when vectors are aggregated by language.

Another problem with TBL occurs with *M* and *F* which appear very frequently. They make early key decision rules in the TBL model and if other vectors share common features with them they may receive a classification of *M* or *F*. Because the TBL does not have enough data it will not have the opportunity to build further rules to reclassify these instances. This is different from the aforementioned problem with *IPFV* and *IMP* because in this case the actual standard gram is neither *M* nor *F* but shares a feature value with *M* or *F* that forces its temporary classification to remain *M* or *F*. An early example is that *FP* would be classified as *F* and with few occurrences in DEV1 the model did not have enough evidence to convert its classification from *F* to *FP*. The *FP* instances that did not share the same feature value with *F* were classified correctly because they did not have that intermediate misclassification.

The methodology is limited in that it cannot assign multiple different standard values to the same unique gloss. If an author of IGT used *NOM* for both nominative and nominalizer the current map gloss system would only be able to select one of these grams. As of this writing, there is not a single language in Language CoLLAGE which has one of these ambiguous instances. However it is unreasonable to expect that other datasets outside of the Language CoLLAGE would not contain ambiguities. The Langauge CoLLAGE contains single author IGT whereas many datasets like ODIN [Lewis and Xia, 2010][Xia et al., 2014] include multiple sources of IGT for the same language which originate from different authors. These languages are likely to have ambiguities in their IGT where one input gloss could map to multiple standard grams. Modeling ambiguities has a trade-off. Currently if there are different gloss instance classifications for a unique gloss the most frequent label will be selected and all other labels will be ignored. By allowing multiple labels for each unique gloss many additional grams will be added to the final set. Recall may increase if some of these additional grams are True Positives but precision will certainly decline because many of them will be False Positives.

The feature development also presented some challenges. During the initial development of the gloss mapping, edit distances were strongly influential in improving results. The Levenshtein edit distance seems intuitive as a measure to help correct misspelled glosses. In practice the distance feature harms results. Trying to manipulate the feature as a count or equate its value with the

edit distance does not impact results. Attempts to only select close distances actually significantly decrease final results. Some of the lexical entry associated features like length and vowel count proved to be more harmful than beneficial. As mentioned in Chapter 3 the Chi Square technique of feature selection was tried. It did not produce beneficial results. Exploring regression testing for feature selection may have helped to select better features. Due to the lack of such analysis on the features, the final feature selections may not be optimal.

6.2 Evaluation

The evaluation could be improved with certain adjustments. The volume and annotation of the data could have impacted results. An error analysis addresses which design decisions led to common errors for both the gloss mapping and inference stages.

6.2.1 Data Assumptions

The most important adjustment would be the input of more training data which would significantly enhance the results of the TBL algorithm. This means more languages at which point it would be helpful to set a threshold for the minimum number of IGT instances per language. Some of the TEST dataset languages did not have many IGT instances which severely impacted the recall of the inference.

There are 2665 unique vectors and 22718 vectors for all Language CoLLAGE datasets. The gold standard annotation assumptions only ask the annotator to review the non-standard glosses within the 2665 unique vectors. While the gold standard annotations have been reviewed multiple times and errors from the DEV1 and DEV2 have been systematically recorded and resolved, there may still possibly be misinterpretations in the data. Most often this would be among incomplete, user-identified, and unrecovered non-standard types. It is also highly possible that individual vectors that share the same unique vector may require different handling. For example, if habitual occurred as both a gloss and a lexical instance in the IGT it should be split into two unique vectors. This likely would not impact inference methods but it is worth noting.

6.2.2 Error Analysis

The gloss mapping outputs have very few errors. Most often the errors are misspellings that were not discovered in the training data. In theory distance features should help correct these errors but actual testing revealed that precision and recall drop significantly when the edit distance features are included. The second most common errors seem to include a range of glosses that have three to five letters and have a gold standard of either combined, user-identified, unrecovered or lexical entry when the lexical entry is misspelled. The commonality between these glosses is that the training dataset has no record of their occurrence and at a length of three to five letters their feature values appear very similar. The model will often classify them as user-identified or lexical entries.

The inference evaluation struggles with how to evaluate tense, aspect, and mood chocies. Hierarchy is ignored but category is not. So if habitual is correctly identified for a language it is not actually counted unless it is matched to the appropriate category such as aspect. These errors however are very infrequent and seem to impact about ten choices across all of the DEV1, DEV2, and TEST languages.

There is some concern with the actual inference results. The gold standard for inference is the student authored choices files from the Language CoLLAGE. This file may be incomplete and lack choices that were found in the IGT and belong to the language. An example of this is French (fra) which contains *IPFV* and *PFV* in its IGT from Language CoLLAGE [Bender, 2014b] but its choices file does not contain any references to imperfective or perfective. At the extreme, a choices file may contain choices for which there is no corresponding IGT instance. The Hudpë (jup) gold standard choices file has over 60 tense, aspect, and mood combinations that do not in any form occur in the IGT. These errors cannot be automatically recovered and substantially affect results. The DEV1 dataset had near perfect precision and recall besides these cases with Hupdë which significantly lowered them. What this does suggest is that while theoretical methods do not work for tense, aspect, and mood the inference using glosses may not produce the intended results simply because IGT do not contain a complete representation of a language's choices. Note that authors of the choices files had access to additional resources beyond IGT.

6.3 Summary

TBL proved to be a very powerful learning model for mapping gloss although it does possess some limitations. Further data may help correct many of these limitations and improve model performance for mapping glosses. Inference results suffer when gold choices files contain choices for which no corresponding gloss exists.

Chapter 7

CONCLUSION

The methods presented in this paper successfully map glosses to a standard set. The evaluation highlights a significant improvement over baseline. The tense, aspect, and mood inference did not benefit as greatly from the improvements to map gloss. This is largely due to how datasets were selected and differences between gold standard IGT and choices files from Language CoLLAGE datasets.

The map glossing project will still continue to develop. Currently the script does not evaluate the post-processing of incomplete and combined grams separate from the general gloss mapping evaluation. Their glosses are generally common such as person, number, person-number, and gender and it follows that their glosses were likely already contained as individual glosses. The methodology for how inference informs category assignments of tense, aspect, and mood could improve to produce better results. These results would then hopefully transfer to user-identified glosses. Currently these user-identified grams are ignored for inference because it is difficult to identify computationally what they represent.

Gloss mapping did improve inference although there is some questions as to whether a gloss-based method would produce the best inference results although no known alternative exists. Due to the substantial investment in and success of gloss mapping other projects that utilize glosses should incorporate gloss mapping to increase their performance.

Appendix A

STANDARD GLOSS SET

Refer to Gloss Conventions §2.5 for a discussion of the standard gloss set. Note that each gram pair includes a list of potential categorical systems for that gram. Some grams have more than one category delimited by a hyphen. Most categories are readily identified such as per, num, pernum, case, tense, aspect, mood, gender, verb, and noun. Dirinv means direct and inverse. Info refers to information structure. Q is reserved for the question particle. Structure is the category for grams that specify sentence or phrasal structure. Coord represents conjunctions and coordination. Psuedo-gram values of !L and !G correspond to pseudo-Leipzig and pseudo-GOLD respectively.

Table A.1: Standard Gloss Set

Leipzig	**GOLD**	**Category**	**Pseudo**
1	FirstPerson	per	
2	SecondPerson	per	
3	ThirdPerson	per	
SG	SingularNumber	num	
DU	DualNumber	num	
PL	PluralNumber	num	
1SG	FirstPerson-SingularNumber	pernum	
2SG	SecondPerson-SingularNumber	pernum	
3SG	ThirdPerson-SingualNumber	pernum	
1DU	FirstPerson-DualNumber	pernum	
2DU	SecondPerson-DualNumber	pernum	

Table A.1: Standard Gloss Set

Leipzig	**GOLD**	**Category**	**Pseudo**
3DU	ThirdPerson-DualNumber	pernum	
1PL	FirstPerson-PluralNumber	pernum	
2PL	SecondPerson-PluralNumber	pernum	
3PL	ThirdPerson-PluralNumber	pernum	
A	Agent	case	
ABL	Ablative	case	
ABS	Absolutive	case	
ACC	Accusative	case	
ALL	Allative	case	
ANIM	Animate	gender	!L
ANTIP	Antipassive	verb	
APPL	Applicative	verb	
AUX	Auxiliary	verb	
BEN	Benefactive	case	
CAUS	Causative	verb	
COM	Comitative	case	
COMP	Complementizer	structure	
COMPL	Completive	aspect-mood	
COND	Conditional	verb	
CONJ	Conjunction	coord	!L
CONT	Continuous	aspect-tense	!L
COP	Copula	verb	
DAT	Dative	case	
DECL	Declarative	mood	

Table A.1: Standard Gloss Set

Leipzig	**GOLD**	**Category**	**Pseudo**
DEF	Definite	noun	
DEM	Demonstrative	dem	
DIR	Direct	dirinv	!L
DIST	Distal	tense-aspect	
DISTR	Distributive	aspect-pernum	
DUB	Dubitive	mood	
DUR	Durative	aspect	
EMPH	Emphatic	info	!L !G
ERG	Ergative	case	
EXCL	Exclusive	person	
F	Feminine	gender	
FP	FocusPoint	info	!L !G
FV	FinalVowel	other	!L !G
FAM	Familiar	person	!L !G
FOC	Focus	info	
FORM	Formal	person	!L !G
FUT	Future	tense	
GEN	Genitive	case	
HAB	Habitual	aspect-tense	
IMP	Imperative	verb	
INANIM	Inanimate	gender	!L
INCL	Inclusive	person	
IND	Indicative	mood	
INDF	Indefinite	noun	

Table A.1: Standard Gloss Set

Leipzig	**GOLD**	**Category**	**Pseudo**
INF	Infinitive	verb	
INS	Instrumental	case	
INTR	Intransitive	noun	
INV	Inverse	dirinv	!L
IPFV	Imperfective	aspect	
IRR	Irrealis	mood	
LOC	Locative	case	
M	Masculine	gender	
N	Neuter	gender	
N-	Non	non	
NEG	Negation	neg	
NFUT	NonFuture	tense	!L
NMLZ	Nominalizer	noun	
NOM	Nominative	case	
NPRS	NonPresent	tense	!L
NPST	NonPast	tense	!L
NRFUT	NearFuture	tense	!L
NRPRS	NearPresent	tense	!L
NRPST	NearPast	tense	!L
OBJ	Object	case	
OBL	Oblique	case	
P	Patient	case	
PASS	Passive	verb	
PERS	Personal	aspect-mood	!L

Table A.1: Standard Gloss Set

Leipzig	**GOLD**	**Category**	**Pseudo**
PFV	Perfective	aspect	
PL	PluralPerson	per	
POSS	Possessive	noun	
PRED	Predicative	structure	
PRET	Preterite	tense-aspect	!L !G
PRF	Perfect	tense-aspect	
PROG	Progressive	aspect-test	
PROH	Prohibitive	mood	
PROX	Proximal	tense-aspect	!G
PRS	Present	tense	
PST	Past	tense	
PTCP	Participle	verb	
PURP	Purposive	mood	!G
Q	Question	q	
QUOT	Quotative	mood	!G
REC	Recent	tense-aspect	!L
RECP	Reciprocal	noun	
REFL	Reflexive	noun	
REL	Relative	coord	
REM	Remote	tense-aspect	!L
REP	Repetitive	verb	
RES	Resultative	aspect	
S	Single	case	
SBJ	Subject	case	

Table A.1: Standard Gloss Set

Leipzig	**GOLD**	**Category**	**Pseudo**
SBJV	Subjunctive	mood	
SG	SingularPerson	per	
STAT	Stative	verb	
TOP	Topic	info	
TR	Transitive	verb	
VOC	Vocative	case	

Appendix B

STANDARD PREDETERMINED SET

The standard predetermined set contains glosses which are recognized to be within a closed set of pronouns, demonstratives, and part-of-speech (pos) entries. The conversion set helps to recognize when a lexical entry bearing linguistic information was not converted to a gram and when part-of-speech values were entered as glosses. This allows for improved classification by identifying and removing part-of-speech entries and improved post-processing by attributing grams to incomplete classified glosses. Other types of glosses beyond pronouns, demonstratives, and part-of-speech may benefit from inclusion on this list. Based on the datasets examined herein, DEV1, DEV2, and TEST, there were no other known glosses that warranted an automatic determination to convert them to standard grams.

Table B.1: Standard Predetermined Set

Gloss	**Type**	**Conversion**
I	pronoun	1sg
you	pronoun	2
he	pronoun	3sg
she	pronoun	3sg
we	pronoun	1pl
yall	pronoun	2pl
y'all	pronoun	2pl
they	pronoun	3pl
me	pronoun	1sg

Table B.1: Standard Predetermined Set

Gloss	**Type**	**Conversion**
him	pronoun	3sg
her	pronoun	3sg
us	pronoun	1pl
them	pronoun	3pl
my	pronoun	1sg
mine	pronoun	1sg
your	pronoun	2
yours	pronoun	2
his	pronoun	3sg
her	pronoun	3sg
hers	pronoun	3sg
our	pronoun	1pl
ours	pronoun	1pl
their	pronoun	3pl
theirs	pronoun	3pl
this	demonstrative	dem, sg
these	demonstrative	dem, pl
that	demonstrative	dem, sg
those	demonstrative	dem, pl
adj	pos	
adv	pos	
prep	pos	
pro	pos	
cc	pos	

Table B.1: Standard Predetermined Set

Gloss	**Type**	**Conversion**
cd	pos	
dt	pos	
ex	pos	
fw	pos	
in	pos	
jj	pos	
jjr	pos	
jjs	pos	
ls	pos	
md	pos	
nn	pos	
nnp	pos	
nnps	pos	
nns	pos	
pdt	pos	
pos	pos	
prp	pos	
prp$	pos	
rb	pos	
rbr	pos	
rbs	pos	
rp	pos	
sym	pos	
to	pos	

Table B.1: Standard Predetermined Set

Gloss	**Type**	**Conversion**
uh	pos	
vb	pos	
vdb	pos	
vbg	pos	
vbn	pos	
vbp	pos	
vbz	pos	
wdt	pos	
wp	pos	
wp$	pos	
wrb	pos	

Appendix C

MATRIX-ODIN MORPHOLOGY ENHANCED INFERENCE

In previous iterations of the inference methodology I attempted to use position classes to improve category assignment. A position class groups morphemes together based on their occurrence in the same position in reference to a root lexeme. Example 25 displays a partial position case in Russian for the present tense. This comes from the Russian choices file in Language CoLLAGE [Bender, 2014b]. This position class occurs after a verb of type `verb4` and imparts the meaning of present tense. The exact value of the position class will then determine person and number.

(25)
```
verb-slot1_name=present-1-conj-1
  verb-slot1_order=after
    verb-slot1_input1_type=verb4
    verb-slot1_morph1_name=1-conj-1-1st-sg
    verb-slot1_morph1_orth=ju
      verb-slot1_morph1_feat1_name=tense
      verb-slot1_morph1_feat1_value=present
      verb-slot1_morph1_feat1_head=verb
      verb-slot1_morph1_feat2_name=number
      verb-slot1_morph1_feat2_value=sg
      verb-slot1_morph1_feat2_head=subj
      verb-slot1_morph1_feat3_name=person
      verb-slot1_morph1_feat3_value=1st
      verb-slot1_morph1_feat3_head=subj
    verb-slot1_morph2_name=1-conj-1-2nd-sg
    verb-slot1_morph2_orth=eshj
```

```
verb-slot1_morph2_feat1_name=tense
verb-slot1_morph2_feat1_value=present
verb-slot1_morph2_feat1_head=verb
verb-slot1_morph2_feat2_name=number
verb-slot1_morph2_feat2_value=sg
verb-slot1_morph2_feat2_head=subj
verb-slot1_morph2_feat3_name=person
verb-slot1_morph2_feat3_value=2nd
verb-slot1_morph2_feat3_head=subj
```

Morphemes and glosses are closely related in this context. I hypothesized that position classes based on morphemes could help disambiguate whether a particular gram for a language is tense, aspect, or mood when that gram can appear in multiple categories. However, at this project's current status position classes mislead results more than they benefit. Russian partially demonstrates why this might be the case. The position classes each represent one tense or aspect choice such as present. The choices within the class distinguish person and number. Using position classes in this context would only help determine categories for person and number rather than tense and aspect.

To obtain position classes for a language this study runs a script called MOM (Matrix-ODIN Morphology) [Wax, 2014] which automates building morphological position classes. MOM first constructs independent position classes from the morphology it observes. It then utilizes an algorithm that combines position classes by comparing their overlap based on similar attachments.

Once the inference procedure has run MOM it will use MOM's resulting position classes to classify each centroid as either tense, aspect, or mood. It achieves this by scoring how likely certain grams will be tense, aspect, or mood. Likelihood estimates are obtained from frequencies in DEV1 and DEV2. The maximum global tense, aspect, mood score will be selected. This will assign one of those three categories to each of the position classes containing a tense, aspect, or mood gram and labeling all grams within that class by that category.

MOM's position classes do not seem to consistently group tense, aspect, and mood choices into exclusive and distinct categories. Further work could be done to revise the analysis so that MOM's outputs would be beneficial but as an initial proof of concept MOM did not improve the inference results.

BIBLIOGRAPHY

Emily M. Bender. Aggregation phase ii: Automatic generation of grammars for endangered languages from glosses and typological information, 2014a.

Emily M. Bender. Language collage: Grammatical description with the lingo grammar matrix. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2447–2451, Reykjavik, Iceland, May 2014b. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/639_Paper.pdf. ACL Anthology Identifier: L14-1508.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan, 2002.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. Grammar customization. *Research on Language & Computation*, 8(1), 2010. ISSN 1570-7075.

Emily M. Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 74–83, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2710.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

Balthasar Bickel, Bernard Comrie, and Martin Haspelmath. The leipzig glossing rules. conventions for interlinear morpheme by morpheme glosses, 2008.

Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565, 1995.

Ann Copestake. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Standford, California, 2002.

Ann Copestake, Dan Flickinger, Carl Pollard, and I. A. Sag. Minimal recursion semantics: An introduc- tion. *Research on Language & Computation*, 3(4), 2005.

Michael Wayne Goodman, Joshua Crowgey, Fei Xia, and Emily M. Bender. Xigt: extensible interlinear glossed text for natural language processing. *Language Resources and Evaluation*, 49(2):455–485, 2015. URL `http://dx.doi.org/10.1007/s10579-014-9276-1`.

Department of Linguistics (The LINGUIST List) Indiana University. General ontology for linguistic description (gold). http://linguistics-ontology.org/gold/2010, 2010.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10.8, pages 707–710, 1966.

William D Lewis and Fei Xia. Automatically identifying computationally relevant typological features. In *IJCNLP*, pages 685–690, 2008.

William D Lewis and Fei Xia. Developing odin: A multilingual repository of annotated language data for hundreds of the world's languages. *Journal of Literary and Linguistic Computing (LLC)*, 25(3):303, 2010.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. Chicago, IL and Stanford, CA: The University of Chicago Press and CSLI Publications, 1994.

Laurie Poulson. *Meta-modeling of Tense and Aspect in a Cross-linguistic Grammar Engineering Platform*. PhD thesis, University of Washington, Seattle, WA, 2011.

David Wax. *Automated Grammar Engineering for Verbal Morphology*. PhD thesis, University of Washington, Seattle, WA, 2014.

Fei Xia, William Lewis, Michael Wayne Goodman, Joshua Crowgey, and Emily M. Bender. Enriching odin. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3151–3157, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/1072_Paper.pdf`. ACL Anthology Identifier: L14-1055.