CHAPTER 1

Introduction/motivation

#0 Knowing about linguistic structure is important for feature design and error analysis in NLP.

The field of linguistics includes subfields that concern themselves with different levels or aspects of the structure of language, as well as subfields dedicated to studying how linguistic structure interacts with human cognition and society. A sample of subfields is briefly described in Table 1.1. At each of those levels of linguistic structure, linguists find systematic patterns over enumerable units where both the units and the patterns have both similarities and differences across languages.

Table 1.1: A non-exhaustive sample of structural subfields of linguistics

Subfield	Description
Phonetics	The study of the sounds of human language
Phonology	The study of sound systems in human languages
Morphology	The study of the formation and internal structure of words
Syntax	The study of the formation and internal structure of sentences
Semantics	The study of the meaning of sentences
Pragmatics	The study of the way sentences with their semantic meanings are used for particular communicative goals

Machine learning approaches to NLP require features which can describe and generalize across particular instances of language use such that the machine learner can find correlations between language use and its target set of labels. It is thus beneficial to NLP that natural language strings have implicit structure and that the field of linguistics has been studying and elucidating that structure. It follows that knowledge about linguistic structures can inform the design of features for machine learning approaches to NLP. Put more strongly: knowledge of linguistic structure will lead to the design of better features for machine learning.

Conversely, knowledge of linguistic structure can also inform error analysis for NLP systems. Specifically, system errors should be checked for linguistic generalizations which can suggest kinds of linguistic knowledge to add to the system.[1] For example, if expletive pronouns (non-

[1] Such error analysis is an excellent opportunity for collaboration between NLP researchers and linguists.

E. M. Bender, *Linguistic Fundamentals for Natural Language Processing* referring pronouns, see #89) are tripping up a coreference resolution system, system performance might be improved by adding a step that detects such pronouns first.

The goal of this book is to present information about linguistic structures that is immediately relevant to the design of NLP systems, in a fashion approachable to NLP researchers with little or no background in linguistics. The focus of this book will be on morphology and syntax (collectively known as morphosyntax) as structures at this level can be particularly relevant to text-based NLP systems. Similar books could (and should) be written concerning phonetics/phonology and semantics/pragmatics. The reader is encouraged to approach the book with particular NLP tasks in mind, and ask, for each aspect of linguistic structure described here, how it could be useful to those tasks.

#1 Morphosyntax is the difference between a sentence and a bag of words.

Morphosyntax is especially relevant to text-based NLP because so many NLP tasks are related to or rely on solutions to the problem of extracting from natural language a representation of who did what to whom. For example: machine translation seeks to represent the same information (including, at its core, who did what to whom) given in the source language in the target language; information extraction and question answering rely on extracting relations between entities, where both the relations and the entities are expressed in words; sentiment analysis is interested in who feels what about whom (or what); etc.[2] To attempt these tasks by treating each sentence (or paragraph or document) as a bag of words is to miss out on a lot of information encoded in the sentence. Consider the contrasts in meaning between the following sets of sentences (from English and Japanese):[3]

(1) a. Kim sent Pat Chris.

b. Kim sent Pat to Chris.

c. Kim was sent to Pat by Chris.

d. Kim was sent Pat by Chris.

[2] Even tasks that aren't concerned with the meaning expressed in the strings they process (e.g., the construction of language models) are impacted by morphosyntax in as much as they care about word order and/or identifying inflected forms as belonging to the same lemma.

[3] All examples from languages other than English in this book are presented in the format of interlinear glossed text (IGT), which consists of three or four lines: The first two lines represent the example in the source language, with one giving source language orthography and the second (optionally, for non-roman orthographies) a transliteration. At least one of these will indicate morpheme boundaries. The remaining two lines give a morpheme-by-morpheme gloss and a free translation into English. The morpheme-by-morpheme glosses use abbreviations for 'grams' (elements like PST for past tense). In general, these should conform to the Leipzig glossing rules [Bickel *et al.*, 2008], but may differ when the original source was using different conventions. The grams used in the IGT in this book are listed in Appendix A. When a gram is relevant to the discussion at hand, its meaning will be explained. The last line includes the ISO 639-3 language code indicating the language of the example.