MORGAN & CLAYPOOL PUBLISHERS

Linguistic Fundamentals for Natural Language Processing

100 Essentials from Morphology and Syntax

Emily M. Bender

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Graeme Hirst, *Series Editor*

Linguistic Fundamentals for Natural Language Processing

100 Essentials from Morphology and Syntax

Synthesis Lectures on Human Language Technologies

Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax
Emily M. Bender
2013

Semi-Supervised Learning and Domain Adaptation in Natural Language Processing
Anders Søgaard
2013

Semantic Relations Between Nominals
Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz
2013

Computational Modeling of Narrative
Inderjeet Mani
2012

Natural Language Processing for Historical Texts
Michael Piotrowski
2012

Sentiment Analysis and Opinion Mining
Bing Liu
2012

Discourse Processing
Manfred Stede
2011

Bitext Alignment
Jörg Tiedemann
2011

Linguistic Structure Prediction
Noah A. Smith
2011

Learning to Rank for Information Retrieval and Natural Language Processing
Hang Li
2011

Computational Modeling of Human Language Acquisition
Afra Alishahi
2010

Introduction to Arabic Natural Language Processing
Nizar Y. Habash
2010

Cross-Language Information Retrieval
Jian-Yun Nie
2010

Automated Grammatical Error Detection for Language Learners
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2010

Data-Intensive Text Processing with MapReduce
Jimmy Lin and Chris Dyer
2010

Semantic Role Labeling
Martha Palmer, Daniel Gildea, and Nianwen Xue
2010

Spoken Dialogue Systems
Kristiina Jokinen and Michael McTear
2009

Introduction to Chinese Natural Language Processing
Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang
2009

Introduction to Linguistic Annotation and Text Analytics
Graham Wilcock
2009

Dependency Parsing
Sandra Kübler, Ryan McDonald, and Joakim Nivre
2009

Statistical Language Models for Information Retrieval
ChengXiang Zhai
2008

Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax

Emily M. Bender

www.morganclaypool.com

ISBN: 9781627050111 paperback
ISBN: 9781627050128 ebook

DOI 10.2200/S00493ED1V01Y201303HLT020

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #20
Series Editor: Graeme Hirst, *University of Toronto*
Series ISSN
Synthesis Lectures on Human Language Technologies
Print 1947-4040 Electronic 1947-4059

Linguistic Fundamentals for Natural Language Processing

100 Essentials from Morphology and Syntax

Emily M. Bender
University of Washington

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #20

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Many NLP tasks have at their core a subtask of extracting the dependencies—who did what to whom—from natural language sentences. This task can be understood as the inverse of the problem solved in different ways by diverse human languages, namely, how to indicate the relationship between different parts of a sentence. Understanding how languages solve the problem can be extremely useful in both feature design and error analysis in the application of machine learning to NLP. Likewise, understanding cross-linguistic variation can be important for the design of MT systems and other multilingual applications. The purpose of this book is to present in a succinct and accessible fashion information about the morphological and syntactic structure of human languages that can be useful in creating more linguistically sophisticated, more language-independent, and thus more successful NLP systems.

KEYWORDS

NLP, morphology, syntax, linguistic typology, language variation

For Laurie Poulson

Contents

Acknowledgments . . . . . **xvii**

1 Introduction/motivation **1**

#0 Knowing about linguistic structure is important for feature design and error analysis in NLP. 1
#1 Morphosyntax is the difference between a sentence and a bag of words. 2
#2 The morphosyntax of a language is the constraints that it places on how words can be combined both in form and in the resulting meaning. 3
#3 Languages use morphology and syntax to indicate who did what to whom, and make use of a range of strategies to do so. 5
#4 Languages can be classified 'genetically', areally, or typologically. 5
#5 There are approximately 7,000 known living languages distributed across 128 language families. 7
#6 Incorporating information about linguistic structure and variation can make for more cross-linguistically portable NLP systems. 8

2 Morphology: Introduction **11**

#7 Morphemes are the smallest meaningful units of language, usually consisting of a sequence of phones paired with concrete meaning. 11
#8 The phones making up a morpheme don't have to be contiguous. 11
#9 The form of a morpheme doesn't have to consist of phones. 13
#10 The form of a morpheme can be null. 13
#11 Root morphemes convey core lexical meaning. 14
#12 Derivational affixes can change lexical meaning. 16
#13 Root+derivational affix combinations can have idiosyncratic meanings. 17
#14 Inflectional affixes add syntactically or semantically relevant features. 18
#15 Morphemes can be ambiguous and/or underspecified in their meaning. 19
#16 The notion 'word' can be contentious in many languages. 20
#17 Constraints on order operate differently between words than they do between morphemes. 21
#18 The distinction between words and morphemes is blurred by processes of language change. 22

#19 A clitic is a linguistic element which is syntactically independent but phonologically dependent. 23
#20 Languages vary in how many morphemes they have per word (on average and maximally). 24
#21 Languages vary in whether they are primarily prefixing or suffixing in their morphology. 25
#22 Languages vary in how easy it is to find the boundaries between morphemes within a word. 26

3 Morphophonology 29

#23 The morphophonology of a language describes the way in which surface forms are related to underlying, abstract sequences of morphemes. 29
#24 The form of a morpheme (root or affix) can be sensitive to its phonological context. 29
#25 The form of a morpheme (root or affix) can be sensitive to its morphological context. 31
#26 Suppletive forms replace a stem+affix combination with a wholly different word. 32
#27 Alphabetic and syllabic writing systems tend to reflect some but not all phonological processes. 33

4 Morphosyntax 35

#28 The morphosyntax of a language describes how the morphemes in a word affect its combinatoric potential. 35
#29 Morphological features associated with verbs and adjectives (and sometimes nouns) can include information about tense, aspect and mood. 36
#30 Morphological features associated with nouns can contribute information about person, number and gender. 38
#31 Morphological features associated with nouns can contribute information about case. 40
#32 Negation can be marked morphologically. 41
#33 Evidentiality can be marked morphologically. 42
#34 Definiteness can be marked morphologically. 43
#35 Honorifics can be marked morphologically. 43
#36 Possessives can be marked morphologically. 44
#37 Yet more grammatical notions can be marked morphologically. 46

#38 When an inflectional category is marked on multiple elements of sentence or phrase, it is usually considered to belong to one element and to express agreement on the others. 46
#39 Verbs commonly agree in person/number/gender with one or more arguments. 47
#40 Determiners and adjectives commonly agree with nouns in number, gender and case. 48
#41 Agreement can be with a feature that is not overtly marked on the controller. 49
#42 Languages vary in which kinds of information they mark morphologically. . . 50
#43 Languages vary in how many distinctions they draw within each morphologically marked category. 51

5 Syntax: Introduction 53

#44 Syntax places constraints on possible sentences. 53
#45 Syntax provides scaffolding for semantic composition. 54
#46 Constraints ruling out some strings as ungrammatical usually also constrain the range of possible semantic interpretations of other strings. 54

6 Parts of speech 57

#47 Parts of speech can be defined distributionally (in terms of morphology and syntax). 57
#48 Parts of speech can also be defined functionally (but not metaphysically). . . . 58
#49 There is no one universal set of parts of speech, even among the major categories. 59
#50 Part of speech extends to phrasal constituents. 60

7 Heads, arguments and adjuncts 61

#51 Words within sentences form intermediate groupings called constituents. . . . 61
#52 A syntactic head determines the internal structure and external distribution of the constituent it projects. 63
#53 Syntactic dependents can be classified as arguments and adjuncts. 65
#54 The number of semantic arguments provided for by a head is a fundamental lexical property. 65
#55 In many (perhaps all) languages, (some) arguments can be left unexpressed. . 66
#56 Words from different parts of speech can serve as heads selecting arguments . 67
#57 Adjuncts are not required by heads and generally can iterate. 69

#58 Adjuncts are syntactically dependents but semantically introduce predicates with take the syntactic head as an argument. 69
#59 Obligatoriness can be used as a test to distinguish arguments from adjuncts. . 71
#60 Entailment can be used as a test to distinguish arguments from adjuncts. 71
#61 Adjuncts can be single words, phrases, or clauses. 72
#62 Adjuncts can modify nominal constituents. 73
#63 Adjuncts can modify verbal constituents. 73
#64 Adjuncts can modify other types of constituents. 74
#65 Adjuncts express a wide range of meanings. 74
#66 The potential to be a modifier is inherent to the syntax of a constituent. 74
#67 Just about anything can be an argument, for some head. 75

8 Argument types and grammatical functions 79

#68 There is no agreed upon universal set of semantic roles, even for one language; nonetheless, arguments can be roughly categorized semantically. 79
#69 Arguments can also be categorized syntactically, though again there may not be universal syntactic argument types. 80
#70 A subject is the distinguished argument of a predicate and may be the only one to display certain grammatical properties. 83
#71 Arguments can generally be arranged in order of obliqueness. 84
#72 Clauses, finite or non-finite, open or closed, can also be arguments. 85
#73 Syntactic and semantic arguments aren't the same, though they often stand in regular relations to each other. 86
#74 For many applications, it is not the surface (syntactic) relations, but the deep (semantic) dependencies that matter. 88
#75 Lexical items map semantic roles to grammatical functions. 88
#76 Syntactic phenomena are sensitive to grammatical functions. 90
#77 Identifying the grammatical function of a constituent can help us understand its semantic role with respect to the head. 91
#78 Some languages identify grammatical functions primarily through word order. 91
#79 Some languages identify grammatical functions through agreement. 93
#80 Some languages identify grammatical functions through case marking. 95
#81 Marking of dependencies on heads is more common cross-linguistically than marking on dependents. 97
#82 Some morphosyntactic phenomena rearrange the lexical mapping. 97

9 Mismatches between syntactic position and semantic roles 101

#83 There are a variety of syntactic phenomena which obscure the relationship between syntactic and semantic arguments 101

#84 Passive is a grammatical process which demotes the subject to oblique status, making room for the next most prominent argument to appear as the subject. 101

#85 Related constructions include anti-passives, impersonal passives, and middles. 103

#86 English dative shift also affects the mapping between syntactic and semantic arguments. 104

#87 Morphological causatives add an argument and change the expression of at least one other. 106

#88 Many (all?) languages have semantically empty words which serve as syntactic glue. 107

#89 Expletives are constituents that can fill syntactic argument positions that don't have any associated semantic role. 109

#90 Raising verbs provide a syntactic argument position with no (local) semantic role, and relate it to a syntactic argument position of another predicate. 110

#91 Control verbs provide a syntactic and semantic argument which is related to a syntactic argument position of another predicate. 112

#92 In complex predicate constructions the arguments of a clause are licensed by multiple predicates working together. 113

#93 Coordinated structures can lead to one-to-many and many-to-one dependency relations. 115

#94 Long-distance dependencies separate arguments/adjuncts from their associated heads 116

#95 Some languages allow adnominal adjuncts to be separated from their head nouns. 118

#96 Many (all?) languages can drop arguments, but permissible argument drop varies by word class and by language. 119

#97 The referent of a dropped argument can be definite or indefinite, depending on the lexical item or construction licensing the argument drop. 121

10 Resources 123

#98 Morphological analyzers map surface strings (words in standard orthography) to regularized strings of morphemes or morphological features. . . . 123

#99 'Deep' syntactic parsers map surface strings (sentences) to semantic structures, including semantic dependencies. 124

#100 Typological databases summarize properties of languages at a high level. . . 125
Summary 125

A **Grams used in IGT** **127**

Bibliography **131**

Author's Biography **153**

General Index **155**

Index of Languages **165**

Acknowledgments

This book grew out of a tutorial I presented at the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, entitled "100 Things You Always Wanted to Know about Linguistics But Were Afraid to Ask*". I am grateful to Graeme Hirst for suggesting expanding that tutorial into a book. This book has also benefited from the thoughtful comments of Rebecca Dridan and two anonymous reviewers. Finally, I would like to thank the various people who answered my queries about the examples used in the text and other loose ends I was chasing down: Adam Albright, Sophia Ananiadou, Yoav Artzi, Tim Baldwin, Miriam Butt, Joshua Crowgey, Anita de Waard, Dan Flickinger, Antske Fokkens, Jeff Good, Varya Gracheva, Angelina Ivanova, Seppo Kittilä, Mayo Kudo, Thibaut Labarre, Peter Lippman, Lutz Marten, Steven Moran, Petya Osenova, Zina Pozen, Susanne Riehemann, Glenn Slayden, Lisa Tittle, Tom Wasow, Fei Xia, and Meliha Yetisgen-Yildiz.

Emily M. Bender
May 2013

*…for fear of being told 1000 more

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.

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.

(2) a. 田中　が　ライオン を　食べた。
Tanaka ga raion wo tabe-ta
Tanaka NOM lion ACC eat-PST
'Tanaka ate the lion.' [jpn]

b. 田中　を　ライオン が　食べた。
Tanaka wo raion ga tabe-ta
Tanaka ACC lion NOM eat-PST
'The lion ate Tanaka.' [jpn]

c. 田中　が　ライオン に　食べられた。
Tanaka ga raion ni tabe-rare-ta
Tanaka NOM lion DAT eat-PASS-PST
'Tanaka was eaten by the lion.' [jpn]

d. 田中　が　ライオン に　ケーキ を　食べられた。
Tanaka ga raion ni keeki wo tabe-rare-ta
Tanaka NOM lion DAT cake ACC eat-PASS-PST
'The lion ate the cake (to Tanaka's detriment).' [jpn]

Conversely, ignoring morphosyntax can obscure the connection between strings which in fact mean the same thing or have closely related meanings. This can be illustrated with the set of examples in (3), which all involve the same fundamental 'giving' situation.

(3) a. Kim gave Sandy a book.
b. Kim gave a book to Sandy.
c. A book was given to Sandy by Kim.
d. This is the book that Kim gave to Sandy.
e. Which book do you think Kim gave to Sandy?
f. It's a book that Kim gave to Sandy.
g. This book is difficult to imagine that Kim could give to Sandy.

#2 The morphosyntax of a language is the constraints that it places on how words can be combined both in form and in the resulting meaning.

Formal linguists typically study morphosyntax from the point of view of grammaticality, describing sets of rules (or alternatively, sets of constraints) which delimit the set of grammatical sentences in a language. Thus pairs of examples like the following are interesting because they potentially illuminate rules (or constraints) that might not be apparent from more run-of-the-mill constructions:

(4) a. Which articles did John file _ without reading _?

b. *John filed a bunch of articles without reading _.

Example (4a) illustrates a phenomenon called 'parasitic gaps'.[4] The * indicates that (4b) is judged to be ungrammatical; _ indicates a position in the sentence where something is 'missing', compared to other related sentences.[5]

Other linguists, including typologists (linguists who study cross-linguistic variation), field linguists (linguists who do primary descriptive work on little-known languages), and grammar engineers (computational linguists who build machine readable hand-crafted grammars), also look at languages in terms of sets of rules or constraints, but tend to put more emphasis on how those constraints relate form to meaning.

For example, Nichols observes in her grammar of Ingush (a Nakh-Daghestanian language of the Caucasus) that "[t]he verb agrees with its nominative argument," and illustrates the point with several examples including the following (2011:432):[6]

(5) a. jett aara-b.ealar
cow out-B.go.WP

'The cow went out.' [inh]

b. zhwalii aara-d.ealar
dog out-D.go.WP

'The dog went out.' [inh]

The difference in the verb forms between these two examples (*b* vs. *d*) reflects the noun class (or 'gender') of the subject. This can be seen as a constraint on well-formedness (if the verb doesn't agree with the gender of the noun bearing nominative case, the sentence is ill-formed) but also as a constraint on possible interpretations: If the verb does not agree with the noun, there may well be some other structure which could be assigned but not one in which the noun is functioning as the subject.

While the notion of grammaticality isn't always of interest in NLP (though it is useful in generation), the view of grammars as constraints on possible structures or possible relationships between words in given sentences is highly relevant.

[4] These examples and their judgments are from Engdahl 1983.

[5] While most syntactic theories make a binary distinction between grammatical and ungrammatical strings, human acceptability judgments are famously more gradient than that [Schütze, 1996, Ch. 3]. Linguists will sometimes use ?, ??, and ?* to indicate degrees of (un)acceptability between fully acceptable and fully unacceptable strings.

[6] WP stands for 'witnessed past tense', which contrasts in the tense system of Ingush with present, future and non-witnessed past forms. For explanation of the symbols used in glosses, see Appendix A.

#3 Languages use morphology and syntax to indicate who did what to whom, and make use of a range of strategies to do so.

Morphosyntax differentiates a sentence from a bag of words (see #1) by adding non-linear structure. That structure encodes information about the relationships between words. Individual words (specifically open class words) denote properties or situations. The structures (and function words) connecting those words build referring expressions out of properties and link the referring expressions to participant roles in the situations. This includes the bare-bones 'who did what to whom' as well as elaborations ("what kind of who did what kind of thing to what kind of whom, where, why, when and how").

Many of the topics covered in the 'syntax' sections of this book concern the various means that languages use for indicating that structure within the string. Linguists understand the structure in terms of multiple linked levels, including the surface form of words and their order, constituent structure, grammatical functions, and semantic predicate-argument structure. These various levels and their relevance to NLP will be discussed in Chapters 5–9.

In many languages, a lot of the information about sentence structure is reflected in the form of the words. The 'morphology' sections of this book concern the different kinds of information that can be expressed within a morphologically complex word (Chapters 2 and 4) and the relationship between the abstract morphological structure and its surface representation (Chapters 2–3).

In many NLP tasks, we want to extract from a sentence (as part of a text) precisely "who did what to whom" (and sometimes even "what kind of who did what kind of thing to what kind of whom, where, why, when and how"). Thus understanding how languages solve the inverse problem of encoding this information can help us more effectively design systems to extract it.

The subfield of linguistic typology is concerned with studying the range of variation across languages, both with an eye towards understanding the boundaries of that range (and thus universals of linguistic structure) as well as towards understanding the ways in which languages change over time and the various factors influencing those changes. Across all phenomena investigated by typologists, languages display interesting yet bounded variation. For example, to indicate 'who did what to whom', languages can and do use word order, case marking (differences in the form of the arguments), and agreement (differences in the form of the predicate), or a combination of those strategies, as described further in #78–80. The fact that languages vary in these ways, together with the fact that the range of variation is bounded, and in many cases, known, makes typology a very rich source of information for the design of NLP systems (see #6 and Bender 2011).

#4 Languages can be classified 'genetically', areally, or typologically.

Languages can be classified in several different ways. So called 'genetic' or 'genealogical' classifications group languages according to shared precursor languages. All languages change