THE WORLD OF TOKENS, TAGS AND TREES

Daniel Zeman

ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

STUDIES IN COMPUTATIONAL AND THEORETICAL LINGUISTICS

Daniel Zeman

THE WORLD OF TOKENS, TAGS AND TREES

Published by the Institute of Formal and Applied Linguistics
as the 19th publication in the series
Studies in Computational and Theoretical Linguistics. First edition, Prague 2018.

This book has been printed with the support of project 15-10472S of the Czech Science Foundation (GAČR).
Printed by MatfyzPress

ISBN 978-80-88132-09-7

to my family

Contents

Acknowledgement

This book is a result of a three-year research project conducted at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University in Prague, funded by the Czech Science Foundation (GAČR), project no. 15-10472S "Morphologically and Syntactically Annotated Corpora of Many Languages (MANYLA)".

I am indebted to all my wonderful colleagues at ÚFAL for their support, feedback and friendly atmosphere; in particular to Martin Popel, Zdeněk Žabokrtský, David Mareček, Rudolf Rosa, Loganathan Ramasamy, Jan Štěpánek and Jan Hajič – my team-mates from the HamleDT and MANYLA projects. I also want to thank the contributors and members of the ever growing Universal Dependencies community, including Joakim Nivre, Chris Manning, Filip Ginter, Marie de Marneffe, Fran Tyers, Sampo Pyysalo, Sebastian Schuster, Natalia Silveira, Teresa Lynn, Bill Croft and many others, for their hard work and fruitful discussions on extending the syntactic forest to new territories. Even deeper on the timeline, I am grateful to Philip Resnik and colleagues for inspiration and hospitality at the University of Maryland, where my work on delexicalized parsing and multilingual corpora began.

Finally, I would never be able to finish this book without the endless patience of my family: Klárka, Zuzka, Lucka and Martin. I love you and promise to spend more time with you from now on again.

Chapter 1

Introduction

This book is about corpora: large collections of sentences in natural language that serve as invaluable resources both for linguistic research and for computer applications that "learn" the human language by reading corpora and observing typical patterns. We are thus in the meeting point of two related and complementing fields: computational linguistics (CL) and natural language processing (NLP).

There are various types of corpora; even a simple collection of documents downloaded by a crawler program from the web can be regarded as a corpus. This book is about corpora that are manually annotated with additional information on the level of *morphology* (properties of individual words and their forms) and *syntax* (relations between words in the sentence). Syntactic relations are often represented as a hierarchical structure called *tree;* consequently, syntactically annotated corpora are called treebanks. We will be interested in one particular type of treebanks, which have become popular and common, and which are called **dependency treebanks.**

The oldest treebanks predate the bloom of natural language processing. Some of them can be traced back to the 1970s (Teleman, 1974; Einarsson, 1976; Těšitelová, 1983). In mid-1990s, the Penn Treebank of English (Marcus et al., 1993) became extremely popular and was used to train and test a large number of NLP models. Penn Treebank is based on immediate constituents, not on syntactic dependencies; algorithms that deal with dependency syntax had yet to wait for their heyday until about 2005. Dependency grammar has been traditionally more popular than constituency (phrase-based) grammar in certain parts of Europe and East Asia; dependencies are also easier to apply to languages with flexible word order. It is thus not surprising that the pioneering work in dependency treebanking was done in languages other than English. One of the largest and most influential dependency treebanks is the Prague Dependency Treebank of Czech (Hajič et al., 2000; Bejček et al., 2013).

In 2006, the CoNLL Shared Task in Multi-Lingual Dependency Parsing (Buchholz and Marsi, 2006) provided dependency treebanks of 13 languages and sparked the

Language	C2006	C2007	C2009	I2010	S2013	C2017
Arabic	yes	yes			yes	yes
Basque		yes			yes	yes
Bengali				yes		
Bulgarian	yes					yes
Catalan		yes	yes			yes
Chinese	yes	yes	yes			yes
Czech	yes	yes	yes			yes
Danish	yes					yes
Dutch	yes					yes
English		yes	yes			yes
French					yes	yes
German	yes		yes		yes	yes
Greek		yes				yes
Hebrew					yes	yes
Hindi				yes		yes
Hungarian		yes			yes	yes
Italian		yes				yes
Japanese	yes		yes			yes
Korean					yes	yes
Polish					yes	yes
Portuguese	yes					yes
Slovenian	yes					yes
Spanish	yes		yes			yes
Swedish	yes				yes	yes
Telugu				yes		
Turkish	yes	yes				yes
25 other						yes

Table 1.1: Languages in multi-lingual parsing shared tasks: CoNLL 2006 (Buchholz and Marsi, 2006), CoNLL 2007 (Nivre et al., 2007), CoNLL 2009 (Hajič et al., 2009), ICON 2010 (Husain et al., 2010), SPMRL 2013 (Seddah et al., 2013) and CoNLL 2017 (Zeman et al., 2017). The 25 extra languages in CoNLL 2017 were Ancient Greek, Buryat, Croatian, Estonian, Finnish, Galician, Gothic, Indonesian, Irish, Kazakh, Kurmanji, Latin, Latvian, North Sámi, Norwegian, Old Church Slavonic, Persian, Romanian, Russian, Slovak, Ukrainian, Upper Sorbian, Urdu, Uyghur and Vietnamese.

interest in both directions that are indicated in the title: building parsers that produce dependency trees, and evaluating them on multiple languages. Testing parsers on the CoNLL datasets (or at least on those treebanks that were freely available after the shared task) became a de-facto standard for several upcoming years. Other parsing shared tasks followed—see Table 1.1 for a brief overview of the languages involved. Various techniques were proposed for cross-linugual parser projection (Zeman and Resnik, 2008; McDonald et al., 2011; Tiedemann, 2014; Rosa and Žabokrtský, 2015).

Unfortunately, different treebanks use quite different annotation schemes, which makes any meaningful cross-linguistic comparison (including evaluation of parser projection techniques) difficult, if not impossible. Various efforts towards interoperability and harmonization of annotation schemes were launched, including Interset (Zeman, 2008),[1] the Universal POS Tagset (Petrov et al., 2012), HamleDT (Zeman et al., 2012, 2014; Rosa et al., 2014),[2] the "Google" Universal Dependency Treebank (McDonald et al., 2013), Universal Stanford Dependencies (de Marneffe et al., 2014) and Universal Dependencies (Nivre et al., 2016).[3] Especially the latter (UD) aims at uniting and superceding all the previous harmonization projects; with 129 treebanks and 76 languages in version 2.3 (Nivre et al., 2018), it is arguably the largest collection of freely available dependency treebanks in the world.

The aim of this book is to gather observations and experience accumulated during conversion of various annotation styles, first within the HamleDT project and later in Universal Dependencies. To provide an overview of design decisions taken in individual treebanks and addressing various phenomena in natural language; to compare the options, to show their advantages and downsides. It is not our primary goal to identify the ultimately "correct" annotation scheme. The current popularity and influence of UD may seem to suggest that whatever approach is taken in UD, is the "correct" way to go. This is not necessarily what we are trying to assert here. To be clear—the author does believe in UD and is an active member of the UD community. The contribution of the project to harmonized treebanking is enormous and undisputable. However, UD is and must be a compromise; not every aspect of the UD guidelines is necessarily the best possible solution for every purpose. This book is not (just) about UD. We will review non-UD and pre-UD treebanks and, by comparison of the diverse approaches their authors have taken, we hope to provide a more varied and multi-dimensional image. Our survey will help people who convert existing treebanks to UD but also those who want to use a UD treebank for a particular purpose and convert the UD-style annotation to a scheme that suits them better. Besides that, the survey should also help to better understand UD itself and to refine and particularize the future versions of the UD guidelines.

[1] http://ufal.mff.cuni.cz/interset/

[2] http://ufal.mff.cuni.cz/hamledt/

[3] http://universaldependencies.org/

While we will look at examples from many different languages and try to cover less known phenomena, we will still mostly deal with phenomena from the "big" languages and major language families. Such a bias is inevitable, given that we study annotation of machine-readable data. Less studied languages rarely have treebanks (or at least morphologically annotated corpora). We can (and sometimes will) theoretize about how particular constructions could be annotated in these languages, but the real complexity of a language can hardly be revealed before real data are annotated.

Chapter 3

Part of Speech Tags

3.1 Types of Tags

The part-of-speech category of each word is one of the most basic and most widespread piece of information found in annotated corpora. It is usually encoded as a short string, called part-of-speech (POS) tag. Many other elements of linguistic annotation could be considered various types of "tags"; however, if the words *tag* or *tagging* are used without further specification, it is usually the part of speech what is being discussed.

The part of speech itself is delimited quite vaguely and the exact list of categories depends on the intended use of the corpus. Even within one language, POS tagsets may vary from ten to several hundred tags. In morphologically rich languages, tags often encode various morphological features in addition to the POS category. It is then more appropriate to term them **morphological tags**[1] rather than POS tags, but the two terms are often used interchangeably. Such tags can be understood as a compact representation of a structure that consists of multiple feature-value pairs, each classifying the word along a different dimension. Some features, such as the part of speech proper, are *lexical:* they categorize the entire entry in the lexicon (lexeme), that is, all words belonging to the same lemma will have the same value in a lexical feature. Other features are *inflectional:* they categorize one word form in a paradigm. Ideally, the lemma plus the values of all inflectional features will uniquely identify the word form (but not all tagging schemes meet this desideratum).

Table 3.1 shows the English tagset of the Penn Treebank (Marcus et al., 1993). There are 45 tags, including 9 tags for various classes of punctuation symbols. The tags are rather atomic strings, although some of them actually encompass inflectional features: `NN` for singular nouns vs. `NNS` for plural, 6 tags for various verbal forms etc.

[1] Or *morphosyntactic descriptions*.

CC	coordinating conjunction	*and, or, but, &, nor*
CD	cardinal number	*million, billion, one, two*
DT	determiner	*the, a, an, this, some*
EX	existential *there*	*there*
FW	foreign word	*de, perestroika, glasnost, vs.*
IN	preposition or subord. conj.	*of, in, for, on, that*
JJ	adjective	*new, other, last, such, first*
JJR	adjective, comparative	*more, higher, lower, less, better*
JJS	adjective, superlative	*most, least, largest, latest, best*
LS	list item marker	*3, 2, 1, 4, First*
MD	modal auxiliary	*will, would, could, can, may*
NN	noun, singular/mass	*%, company, year, market*
NNS	noun, plural	*years, shares, sales, companies*
NNP	proper noun, singular	*Mr., U.S., Corp., New, Inc.*
NNPS	proper noun, plural	*Securities, Democrats*
PDT	predeterminer	*all, such, half, both, nary*
POS	possessive ending	*'s, '*
PRP	personal pronoun	*it, he, they, I, we*
PRP$	possessive pronoun	*its, his, their, our, her*
RB	adverb	*n't, not, also, only, as*
RBR	adverb, comparative	*more, earlier, less, higher*
RBS	adverb, superlative	*most, best, least, hardest, worst*
RP	particle	*up, out, off, down, in*
SYM	symbol	*a, c, *, **, b*
TO	*to*	*to*
UH	interjection	*yes, well, no, OK, oh*
VB	verb, base form	*be, have, make, buy, get*
VBD	verb, past tense	*said, was, were, had, did*
VBG	verb, gerund or present participle	*including, being, according*
VBN	verb, past participle	*been, expected, made, based*
VBP	verb, non-3rd person sing. pres.	*are, have, do, say, 're*
VBZ	verb, 3rd person singular present	*is, has, says, 's, does*
WDT	wh-determiner	*which, that, what, whatever*
WP	wh-pronoun	*who, what, whom, whoever*
WP$	possessive wh-pronoun	*whose*
WRB	wh-adverb	*when, how, where, why*
#	number sign	#
$	currency	*$, C$, US$, A$, HK$*
,	comma	*,*
.	period	*., ?, !*
``	opening quotation mark	*", '*
''	closing quotation mark	*", '*
-LRB-	opening bracket	*(, [, {*
-RRB-	closing bracket	*),], }*
:	other punctuation	*--, :, ;, ..., -*

Table 3.1: The English tagset of the Penn Treebank (Marcus et al., 1993) with examples.

Char	Meaning	Values
1	part of speech	NAPCVDRJTIZX
2	subpart of speech, mood	over 70
3	gender	MIFNXYTWHQZ
4	number	SDPWX
5	case	1234567X
6	possessor's gender	MF
7	possessor's number	SP
8	person	123
9	tense	MPF
10	degree of comparison	123
11	polarity	AN
12	voice	AP
13	*reserved*	
14	*reserved*	
15	style	12356789

Table 3.2: Character positions in the Czech tagset of the Prague Dependency Treebank (Hajič et al., 2000).

In contrast, a morphological tag in the Prague Dependency Treebank of Czech (Hajič et al., 2000) is always exactly 15 characters, each corresponding to a different feature.[2] The position of the character in the tag determines the feature; hence tagsets of this type are called *positional*. If the feature is not relevant in the context of the other features, its value is set to a hyphen, "-". Some features also allow the value "X", which is different from the hyphen. It means that the feature is relevant, but it is unknown or undeterminable for the particular word that bears the tag. Of course, the (un)determinability of a feature depends on how much we are willing to disambiguate from the context of the surrounding text. An example of a PDT tag is AGFS3-----A----. It says that the word is adjective (A), subtype verbal – present active participle (G), feminine (F), singular (S), dative (3), affirmative form (A). More than 4000 character combinations are licensed by the Czech morphological lexicon, although some of them are rare and not attested in the treebank.

Many other tagsets of morphologically rich languages adopt a similar positional approach, although they do not necessarily require that all tags have the same length. A common modification, used e.g. in the MULTEXT-EAST tagsets (Erjavec, 2012), is to allow variable set of features (that is, number of characters and their interpretation) for various parts of speech: for example, nouns will have 6 characters, the first

[2] In fact there are only 13 features because two positions have been reserved and never used.

character is N, and the other positions encode noun type, gender, number, case and animacy; adverbs will have 3 characters, the first character is R, and the other positions encode adverb type and degree of comparison. This way the number of hyphens for irrelevant features is reduced, though they still occur. Furthermore, trailing hyphens are omitted.

Some corpora encode features and their values more verbosely and list, for every token, a sequence of X=Y assignments, where X is the name and Y the value of the feature. There is still some variability about how verbose the scheme is, thus one corpus may say Pos=N | Gen=M | Num=S | Cas=3, while another will have pos=noun | gender=masculine | number=singular | case=dative. In Universal Dependencies, the main part-of-speech category is encoded separately as the universal, coarse-grained POS tag; more fine-grained lexical categories and all inflectional features are stored in a separate place. For instance, the universal POS tag may be NOUN and the accompanying features may be Gender=Masc | Number=Sing | Case=Dat. One of the most compact examples of an X=Y encoding is the Ajka tagset of Czech (Jakubíček et al., 2011), where every dimension consumes two characters, one identifying the feature and the other representing its value. Thus the tag k1gMnPc4 represents a noun (first category, k1), masculine (gM), plural (nP), accusative (fourth case, c4), while k5eAaImIp1nP is a verb (k5), affirmative (eA), imperfective (aI), indicative (mI), first person (p1) plural (nP).

All these variants are merely ways of encoding information. There is no principled difference in the amount or type of information that can be encoded. It is thus possible to design mutually equivalent and convertible encodings of the same set of tags in various shades of the X=Y feature mapping, or in a positional scheme. As long as two tagsets cover the same grammatical categories with the same degree of granularity, it does not really matter which encoding of the categories we choose. We can always convert them to the other representation if necessary.

However, tagsets typically are not equivalent. Even two different tagsets of one language are usually designed with varying level of granularity, as can be illustrated on two tagsets for Swedish: Mamba (Teleman, 1974; Nilsson et al., 2005) and SUC (Stockholm-Umeå Corpus) (Gustafson-Capková and Hartmann, 2006, p. 20–21). Mamba was used in the original version of Talbanken, the Swedish treebank from 1970s. The tagset defines 48 tags but 8 of them deal with phenomena specific to spoken dialogue and are not attested in the treebank.[3] Even the set of 40 attested tags (Table 3.3) is somewhat "unbalanced": there are 10 tags for different types of punctuation, and 10 tags for individual auxiliary verbs (besides the eleventh tag, VV, that covers all ordinary verbs). There are no morphological features. In contrast, the 25 POS tags of SUC (Table 3.4) include three types of punctuation, and a more mainstream selection of subclasses, such as interrogative/relative ("wh-" in English) adverbs, determiners and pronouns. These core tags are accompanied by values of 10 morphological features (Table 3.5), yielding over 150 possible tag strings attested in the treebank. It is

[3] We refer to the Talbanken data used in the CoNLL 2006 shared task.

obvious that mapping between the two tagsets is bound to lose information, unless the underlying text can be accessed and re-tagged.

3.2 Parallel and Serial Combination of Tags

3.2.1 Ambiguity

Tagsets come with different expectations about how much can and should be disambiguated by context. For example, the English word *can* is either a modal auxiliary (as in *I can give you a ride*), or a noun (as in *I have a can full of fruit*). We can also derive a verb from the noun (as in *How to can fruits*). The surface ambiguity between the first *can* and the other two is purely coincidental and we definitely want to disambiguate them in text. The second and third *can* are related, one is derived from the other, but we still want to distinguish them because the syntactic (distributional) rules applying to nouns and verbs are not compatible. On the other hand, words like *who* or *where* can be classified (and used) as either interrogative or relative, in English as well as in many other languages. It is usually not considered crucial to distinguish whether they are interrogative or relative in a given context, and thus tagsets often define one category that encompasses both functions (although this category may be defined multiple times, independently for pronouns, determiners and adverbs; cf. the "wh-" tags in the Penn Treebank tagset).

A more controversial example is the English tag `TO`, reserved for a single word, *to*. The word is either a preposition *(I give it to you)* or an infinitive marker *(I want you to come)*. The two functions and their distribution is different, and they would deserve to be disambiguated. After all, other prepositions are tagged `IN`. However, the word is very frequent and automatic taggers are likely to make a lot of errors; or at least it was likely in early 1990s when the tagset was designed. Indeed, (Marcus et al., 1993, p. 2) say: "the stochastic orientation of the Penn Treebank and the resulting concern with sparse data led us to modify the Brown Corpus tagset by paring it down considerably." They also argue that it is not fatal if they hide some distinctions in the tagset because the distinctions can be deduced from the syntactic structure.[4] Therefore, both functions of *to* are tagged with the same tag `TO`; if an application needs to disambiguate them, it has to do it on its own.

Similarly, the Czech tagset (Table 3.2) has ambiguous values for several features. Czech has four gender-animacy values (masculine animate, masculine inanimate, feminine and neuter) and two to three numbers (singular and plural, plus some surviving forms of the dual). However, the tagset has 11 values of gender and 5 values of number. The seven extra genders are various combinations of the four basic values. For instance, `Y` means either `M` or `I`, that is, masculine animate or inanimate. It is used with

[4] The creators of the Penn Treebank could not foresee the enormous popularity their tagset would gain over the years. It has been applied to many other datasets, regardless whether those datasets included syntactic structures and whether those structures, if present, were created manually or automatically.

++	coordinating conjunction	*och, eller, men, utan, samt*
AB	adverb	*inte, så, också, i, där*
AJ	adjective	*stor, olika, större, stora, nya*
AN	adjectival noun	*möjlighet, trygghet, möjligheter*
AV	the verb *vara* "to be"	*är, vara, var, varit, vore*
BV	the verb *bli* "to become"	*blir, bli, blivit, blev, bör*
EN	indef. article or numeral one	*en, ett, 1*
FV	the verb *få* "to get"	*får, få, fått, fick, finns*
GV	the verb *göra* "to do"	*göra, gör, gjort, gjorde, görs*
HV	the verb *ha* "to have"	*har, ha, hade, haft, hava*
I?	question mark	*?*
IC	quotation mark	*'*
ID	part of idiom	*att, Backberger, och, av, Hellsten*
IG	other punctuation	*..., /, =, ..., 1*
IK	comma	*,*
IM	infinitive marker	*att*
IP	period	*.*
IQ	colon	*:*
IR	parenthesis	*(,)*
IS	semicolon	*;*
IT	dash	*-, ---*
IU	exclamation mark	*!*
KV	*komma att* "to be going to"	*kommer, kommit, kom, komma, komer*
MV	the verb *måste* "must"	*måste, måsk*
NN	other noun	*äktenskapet, barn, äktenskap, familjen*
PN	proper name	*Barbro, Stig, Sverige, Gud, Hellsten*
PO	pronoun	*det, som, den, man, de*
PR	preposition	*i, av, på, för, med*
PU	pause	**, -*
QV	the verb *kunna* "can"	*kan, kunna, kunde, kunnat*
RO	numeral other than one	*två, tre, 20, 1968, 10*
SP	present participle	*kommande, bestående, gällande, växande*
SV	the verb *skola* "will, shall"	*skall, skulle, ska, skola*
TP	past participle	*ökade, ingångna, ökad, utlämnade*
UK	subordinating conjunction	*att, som, om, än, så*
VN	verbal noun	*uppfattning, betydelse, uppfostran*
VV	other verb	*finns, bör, tror, anser, säger*
WV	the verb *vilja* "to want"	*vill, vilja, ville, velat*
XX	unclassifiable	
YY	interjection	*ja, nej, jo, jodå, javisst*

Table 3.3: The Mamba tagset for Swedish (Teleman, 1974; Nilsson et al., 2005). The table shows 40 tags attested in the Talbanken corpus, example words are given in the third column. The tagset defines additional 8 tags, intended for other corpora and mostly dealing with spoken dialogue annotation.

AB	adverb	*inte, också, så, bara, nu*
DT	determiner	*en, ett, den, det, alla*
HA	interrog./relative adverb	*när, där, hur, som, då*
HD	interrog./relative determiner	*vilken, vilket, vilka*
HP	interrog./relative pronoun	*som, vilken, vem, vilket, vad*
HS	interrog./relative possessive	*vars*
IE	infinitive marker	*att*
IN	interjection	*jo, ja, nej*
JJ	adjective	*stor, annan, själv, sådan, viss*
KN	coordinating conjunction	*och, eller, som, än, men*
MAD	meaning separating punctuation	*., ?, :, !, ...*
MID	punctuation inside of sentence	*„ -, :, *, ;*
NN	noun	*år, arbete, barn, sätt, äktenskap*
PAD	paired punctuation	*', (,)*
PC	participle	*särskild, ökad, beredd, gift*
PL	particle	*ut, upp, in, till, med*
PM	proper name	*F, N, Liechtenstein, Danmark*
PN	pronoun	*han, den, vi, det, denne*
PP	preposition	*i, av, på, för, till*
PS	possessive pronoun	*min, din, sin, vår, er*
RG	cardinal numeral	*en, ett, två, tre, 1*
RO	ordinal numeral	*första, andra, tredje, fjärde, femte*
SN	subordinating conjunction	*att, om, innan, eftersom, medan*
UO	foreign word	*companionship, vice, versa, family*
VB	verb	*vara, få, ha, bli, kunna*

Table 3.4: The Stockholm-Umeå Corpus tagset for Swedish (Gustafson-Capková and Hartmann, 2006, p. 20–21) with example words.

Feature	Values
Gender	UTR, NEU, MAS
Number	SIN, PLU
Definiteness	IND, DEF
Case	NOM, GEN
Tense	PRS, PRT, SUP, INF
Voice	AKT, SFO
Mood	KON
Participle form	PRS, PRF
Degree	POS, KOM, SUV
Pronoun form	SUB, OBJ, SMS

Table 3.5: Features accompanying the tags in the Stockholm-Umeå Corpus of Swedish.

singular past tense forms of verbs, which do not distinguish animacy (e.g. *dělal* "he (Anim|Inan) did"). In contrast, plural past tense verbs have one form common for masculine inanimates and feminines (T=I|F, e.g. *dělaly* "they did"), while masculine animates (*dělali* "they did") and neuters (*dělala* "they did") are different. There are even values that are used only in certain combinations of gender and number: the gender Q=F|N is feminine or neuter, but it is only used together with the number W=S|P; together they denote forms that can be either feminine singular or neuter plural (but not feminine plural, nor neuter singular). All these ambiguities pertain to specific productive patterns of Czech morphology. They could be disambiguated by context but it was probably considered too risky given the accuracy of taggers at the time the tagset was designed. On the other hand, the feature of case is always disambiguated (except for indeclinable loanwords), although there are systematic ambiguities too: for example, the adjectives of so-called "soft declension" have just one form for all cases in the feminine singular. We can speculate that the reason for putting more stress on case disambiguation was the importance of case for syntax and valency.

3.2.2 Layered Features

In some languages, some features are marked more than once on the same word. For example, possessive pronouns (also called possessive determiners or adjectives in various terminological systems) may have two independent values of gender and two independent values of number. One of the values characterizes the possessor, the other characterizes the possessee. The possessor's gender and number is something that we observe also with normal personal pronouns: for instance, the English 3rd-person pronouns distinguish singular and plural, and they also distinguish three genders in the singular *(he, she, it)* but not in the plural *(they)*. Likewise, the corresponding possessive pronouns have three genders in singular *(his, her, its)* but only

		Sing	Sing	Plur
	Case	Masc/Neut	Fem	Masc/Fem/Neut
Prs	Nom	*on/ono*	*ona*	*oni/one/ona*
Prs	Gen	*njega*	*nje*	*njih*
	Number Gender Case			
Poss	Sing Masc Nom	*njegov*	*njezin*	*njihov*
Poss	Sing Fem Nom	*njegova*	*njezina*	*njihova*
Poss	Sing Neut Nom	*njegovo*	*njezino*	*njihovo*
Poss	Plur Masc Nom	*njegovi*	*njezini*	*njihovi*
Poss	Plur Fem Nom	*njegove*	*njezine*	*njihove*
Poss	Plur Neut Nom	*njegova*	*njezina*	*njihova*

Table 3.6: The nominative and genitive forms of Croatian 3rd person pronouns, and the nominative forms of the corresponding possessive pronouns. The rows represent various genders and numbers of the possessee, while the columns represent genders and numbers of the possessor.

one form in plural *(their)*. English does not mark the possessee's features morphologically, but other languages do.

Thus in Croatian, the 3rd person pronouns distinguish three genders and two numbers in the nominative case, but in the other cases and in the possessives, the singular masculine is often identical to the singular neuter, and the plural forms are mostly common for all three genders. In most cases, there are three distinct forms (Table 3.6). There are also possessive pronouns for three different categories of possessors: masculine/neuter singular *(njegov)*, feminine singular *(njezin)*,[5] and plural *(njihov)*. However, in Croatian the possessive pronouns behave like adjectives and agree in gender, number and case with the possessed (modified) noun. If the possessee is masculine singular, such as *pas* "dog", the possessive pronoun will acquire a masculine suffix: *njegov pas* "his dog", *njezin pas* "her dog", *njihov pas* "their dog". If the possessee is feminine singular, the form of the possessive changes and takes the feminine suffix: *njegova mačka* "his cat", *njezina mačka* "her cat", *njihova mačka* "their cat". Similarly for singular neuter (*njegovo polje* "his field"), plural masculine (*njegovi psi* "his dogs") etc.

We thus need tags that distinguish the ordinary agreement suffixes (i.e., the possessee's gender, number and case) from the possessor's gender and number, which is encoded in the stem. Universal Dependencies call this *layered features:* there are two layers of gender, and two layers of number. There is also a specific notation: if a word is annotated more than once with a feature, the layers must be identified by a predefined string given in square brackets. For instance, a masculine possessor would

[5] In fact, there are two feminine possessive variants: *njezin* and *njen*. We disregard the latter here.

be annotated as Gender[psor]=Masc. One layer can be treated as default and given without layer name; in our example, the agreement gender would be annotated simply as Gender=Masc. We will adopt the term *layered features* in this study, but not necessarily the notation, which always depends on the particular tagset.

3.2.3 Chained Features

In Sections 3.2.1 and 3.2.2, multiple tags or features were applied to a word in parallel. There are also situations where multiple tags or features apply to a word in sequence. We have seen examples in Section 2.4, where one orthographic word was segmented into multiple syntactic words, each with its own morphological tag. We have also seen examples of collapsed multi-word expressions in Section 2.3 and at least in the Alpino treebank, sequences of words that are collapsed into one token have also sequences of tags and features.[6] Hence, for example, Dutch *voor_het_geval* (lit. *for the case*) "in case of" is a multi-word unit and has the coarse-grained POS tag MWU, but its fine-grained tag is a sequence of three parts of speech: a preposition, an article and a noun – Prep_Art_N. Likewise, there are three sets of features, joined by underscore characters. The first feature, voor, says that this is a preposition *(voorzetsel)*, as opposed to postpositions, circumpositions and infinitive markers, which would also fall under the tag Prep. The second set of features, bep|onzijd|neut, says that the article is definite *(bepaald)*, neuter *(onzijd)* and neutral w.r.t. case *(neut)*. The third set of features, soort|ev|neut, says that the noun is common *(soort)*, singular *(enkelvoud)* and case-neutral.

In addition to the cases described above, some languages (especially agglutinating ones) allow repeated application of the same feature even in tokens that are not multi-word expressions or multi-word tokens. For example, Turkish has 5 basic voices: active is unmarked, the other four are marked by specific morphemes: passive, causative, reciprocal and reflexive. But there are words where multiple voice morphemes appear (e.g., causative + passive: "to be caused by someone to do something"), even the same voice can be applied multiple times (e.g., X caused Y to cause Z to do something). Similarly, there can be multiple tenses and multiple moods. If these operations are not analyzed as derivation rather than inflection, we have multiple values of one feature applied in sequence. They are not exactly layered features because the different voices (tenses, moods) do not refer to different entities and it is not clear what should be the labels and meanings of layers. The frequent approach in Turkish and similar languages is to segment the word into so-called *inflectional groups* and provide a sequence of tags that explain properties of each group. However, such tag sequences cannot be easily mapped to a Feature=Value model, where at most one assignment to each feature name is expected. So the current solution in UD, for instance, is

[6] We are referring to the version of the Alpino treebank that was released for the CoNLL 2006 shared task.

to define language-specific values that look like sequences of basic universal values, e.g., Voice=CauPass.

3.3 Harmonization Efforts

We showed in Section 3.1 how diverse the tagging approaches can be, depending on the use cases envisioned by their designers. From a more general point of view, such variability is disadvantageous, as significant effort is needed for users and tools to adapt to new corpora and tagsets. That is why there have been several attempts to standardize morphological tagsets, with varying level of success.

3.3.1 EAGLES, PAROLE and MULTEXT-EAST

The EAGLES project (EAGLES, 1996; Leech and Wilson, 1999) produced a set of recommendations for tagsets. The project report contained two complete tagsets for English and Italian but the recommendations were based on considering several other west European languages. The EAGLES guidelines were organized hierarchically, trying to standardize the most common concepts while leaving room for language-specific or project-specific extensions. The highest level corresponded to the major part-of-speech categories (Table 3.7).

On the next level, a set of recommended feature-value pairs was defined separately for each major part of speech (for instance, Table 3.8 shows the four recommended features of nouns, and Table 3.9 shows the eight recommended features of verbs). Not all features ("attributes" in the EAGLES terminology) are relevant in all languages, and some languages may need only a subset of the predefined values. However, it was expected that if a language makes a distinction captured by a recommended feature, the tagset would use the feature.

The third level corresponded to optional attributes (or new values of existing attributes) that could be added in concrete tagsets if needed. This way the guidelines could be extended to other languages beyond those considered in the original proposal.

EAGLES did not prescribe a single encoding of the categories and values it defined. It only defined encoding of so-called *intermediate tags* and required that an EAGLES-compliant tagset would operate on a compatible level of granularity, so that surface tags could be automatically mapped to intermediate tags. The intermediate tags are positional, starting with one or two letters denoting the major POS category, and followed by feature values expressed as Arabic digits. Thus, for example, the Italian verb *avere* "to have" has the intermediate tag `V00025101`. Following Table 3.9 it can be decoded as a non-finite verb (`2`), infinitive (`5`), present tense (`1`), used as a main (rather than auxiliary) verb (`1`). The features person, gender, number and voice are irrelevant for Italian infinitives, hence the value `0`. In the real tagset used by a tagger or in a corpus, *avere* would be tagged by a more compact and readable tag `VFY`; however, the

Tag	Category
N	Noun
V	Verb
AJ	Adjective
PD	Pronoun or determiner
AT	Article
AV	Adverb
AP	Adposition (preposition or postposition)
C	Conjunction
NU	Numeral
I	Interjection
U	Unique or unassigned
R	Residual
PU	Punctuation

Table 3.7: EAGLES obligatory major categories. The category U comprises categories with a unique or very small membership, such as "negative particle", which are unassigned to any of the standard part-of-speech categories. The residual category (R) contains tokens that stand outside the traditionally accepted range of grammatical classes, e.g., foreign words, mathematical formulae, symbols, acronyms or abbreviations.

	Feature	Values
(i)	Type	1. Common 2. Proper
(ii)	Gender	1. Masculine 2. Feminine 3. Neuter
(iii)	Number	1. Singular 2. Plural
(iv)	Case	1. Nominative 2. Genitive 3. Dative 4. Accusative 5. Vocative

Table 3.8: EAGLES recommended features for nouns.

	Feature	Values
(i)	Person	1. First 2. Second 3. Third
(ii)	Gender	1. Masculine 2. Feminine 3. Neuter
(iii)	Number	1. Singular 2. Plural
(iv)	Finiteness	1. Finite 2. Non-finite
(v)	Verb form / mood	1. Indicative 2. Subjunctive 3. Imperative 4. Conditional 5. Infinitive 6. Participle 7. Gerund 8. Supine
(vi)	Tense	1. Present 2. Imperfect 3. Future 4. Past
(vii)	Voice	1. Active 2. Passive
(viii)	Status	1. Main 2. Auxiliary

Table 3.9: EAGLES recommended features for verbs.

tagset definition table would map it to `V00025101` and thus define the tag in a unique and machine-readable way. The intermediate tags also have a mechanism for expressing alternatives. For example, in English it is useful to have one tag for the base form of a verb, but it corresponds to a number of possible morphological categories. Even if we leave out the non-finite use of the base form (as an infinitive), we still can interpret the word in many different ways (example taken from (EAGLES, 1996)): "[finite indicative present tense [plural or [first person or second person] singular] or imperative or subjunctive]". In the intermediate tag, this is represented with the help of the special symbols - (anything except the following subtag), | (disjunction of subtags) and [] (brackets for grouping): `V[[-301|002]111|000121|000130]01`.

EAGLES was followed by the EU-funded project LE-PAROLE (Volz and Lenz, 1996), whose main outcome was a multilingual corpus of 14 European languages, morphosyntactically annotated according to a common core PAROLE tagset, extended with a set of language-specific features in an EAGLES-compliant fashion. The fourteen languages covered were all EU languages of that time and one non-EU language: Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Irish, Italian, Norwegian, Portuguese, Spanish and Swedish. Applicability of the scheme to non-European languages remained an open question but at least the project could claim having made a step outside the Indo-European family (with Finnish belonging to Uralic languages).

A more recent example of a practical application of EAGLES is the FreeLing tool (Padró and Stanilovsky, 2012), which contains a tagger producing EAGLES-compliant morphological tags.[7] In version 4.0, FreeLing supports 14 languages: Asturian, Catalan, Croatian, English, French, Galician, German, Italian, Norwegian, Portuguese, Russian, Slovenian, Spanish and Welsh.

[7] `https://talp-upc.gitbooks.io/freeling-4-0-user-manual/content/tagsets.html`

Another multilingual corpus with common tagset is MULTEXT (Ide and Véronis, 1994) for six European languages (English, Dutch, German, French, Spanish, Italian), and later its more vital spin-off MULTEXT-EAST (Erjavec, 2012). It offers a parallel, morphologically annotated corpus (the 1984 novel by George Orwell), lexicons and harmonized tagsets ("morphosyntactic descriptions"). There were several releases since late 1990s; in version 4 (Erjavec, 2010),[8] MULTEXT-EAST covers 17 languages from two families: Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Macedonian, Persian, Polish, Resian, Romanian, Russian, Serbian, Slovak, Slovenian and Ukrainian. For some languages (e.g., Bulgarian, Slovenian, Serbian), the MULTEXT-EAST-derived tagset became the most-widely used tagset of the language. For others (e.g., Czech), it did not win out the competition with already established tagsets, and its usage is more or less limited to the MULTEXT-EAST project, as a means of cross-linguistic comparison.

The MULTEXT-EAST tagsets are positional, starting with an uppercase letter identifying the part-of-speech category (Table 3.10) and following with lowercase letters and digits that encode feature values. The tags are EAGLES compliant and can be mapped on the intermediate tagset of EAGLES. There is a large number of optional attributes and values, partially because of the more detailed approach of MULTEXT, and partially due to the morphological richness of the languages covered (for example, nouns have up to 14 features including the 4 basic features recommended in EAGLES; the case feature has 31 possible values (cf. the 5 cases in EAGLES), though no single language uses all of them). The sets of categories are mostly based on concepts used in the grammatical tradition of the individual languages. So for instance, the category of determiners is used in English, Romanian and Persian but not in the Slavic languages, where the corresponding words are traditionally subsumed into pronouns.

The morphological complexity of the Central and East European languages makes the harmonization endeavor in MULTEXT-EAST inherently more difficult than PAROLE. However, the MULTEXT-EAST tagsets are not perfectly harmonized, i.e., there are still phenomena that are tagged differently in different languages. For example, in Slavic languages there is a verbal form that behaves syntactically as an adverb and is variously termed adverbial participle, transgressive, gerund or converb. The MULTEXT-EAST tagsets of Polish, Russian, Ukrainian and Bulgarian tag this form as a verb with the feature VForm=gerund (g). In Czech and Slovak, the form is also verb but with VForm=transgressive (t), following the local terminology. In Serbian and Macedonian, the form is classified as adverb with the feature Type=verbal (v). And finally in Slovenian, the form is tagged as an adverb with Type=participle (r).

[8] http://nl.ijs.si/ME/V4/

Tag	Category
N	Noun
V	Verb
A	Adjective
P	Pronoun
D	Determiner
T	Article
R	Adverb
S	Adposition
C	Conjunction
M	Numeral
Q	Particle
I	Interjection
Y	Abbreviation
R	Residual

Table 3.10: MULTEXT-EAST major word categories (POS). Compare it with the EAGLES categories in Table 3.7. The two sets align quite well. MULTEXT does not have a tag for punctuation, which is distinguished already at the level of XML markup. The "unique-unassigned" category from EAGLES roughly corresponds to particles in MULTEXT. The PD category is split to pronouns and determiners, and abbreviations are separated from other residual tokens.

3.3.2 Indian Languages

India is after Europe another part of the world where NLP technology has to tackle many different languages. There are four main language families found in India: Indo-European, Dravidian, Sino-Tibetan and Austro-Asiatic. Most Indian languages belong to the first two. The families are very different typologically, yet there are similarities too, thanks to centuries of language contact on the Indian subcontinent.

Several tagsets have been designed to cover multiple Indian languages. One of the early solutions was the IIIT tagset (Bharati et al., 2006b), which bears some resemblances to the Penn Treebank English tagset. A hierarchical, EAGLES-inspired common POS-tagset framework was later proposed by (Baskaran et al., 2008). It is supposed to cover the morphosyntactic details of Indian languages and to offer advantages such as flexibility, cross-linguistic compatibility and reusability. Subsequently, the proposal was refined following input from IIIT and other researchers, and it was eventually submitted to the Bureau of Indian Standards (BIS) (Lata et al., 2010).[9] See Table 3.11 for the list of the tags in this tagset. A full tag is constructed by joining the coarse and the fine-grained tag, e.g., `V_VM_VINF` denotes an infinitive of a main verb. While the tagset is supposed to accommodate languages from all four Indian families, the proposal demonstrates its application to 12 languages (8 Indo-European and 4 Dravidian): Bangla, Gujarati, Hindi, Kannada, Konkani, Maithili, Malayalam, Marathi, Punjabi, Tamil, Telugu and Urdu.

3.3.3 Interset, UPOS and Universal Dependencies

The projects mentioned so far aimed at standardization of primary tagsets used in corpus annotation. Another wave of harmonization efforts was sparked by the need for interoperability between NLP tools.

(Zeman, 2008) proposed Interset,[10] a set of morphosyntactic features applicable to a large number of languages. Its original purpose was to aid conversion of tagsets in the context of cross-linguistic transfer of machine-learned models (Zeman and Resnik, 2008). The role of the universal set of features in tag conversion was similar to the role of Interlingua in Interlingua-based machine translation (Richens, 1958) or the role of Unicode among character sets. Features from tagset A were first mapped to the universal set of features, then to features of tagset B; the mapping between each physical tagset and Interset could be reused in conversion of any other tagset to and from tagsets A and B. As a side-effect, Interset itself became a useful means for description of morphosyntax. Its feature-value inventory is meant to be universal and cover anything that one may want to encode in a morphological tag. It is built bottom-up and new features or values are added as the need arises, hence there was at least initially a bias towards big and well-resourced languages, as with most other harmonization

[9] `http://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf`

[10] `http://ufal.mff.cuni.cz/interset`

Tag	Category	Fine Tag	Category
N	Noun	NN	Common noun
		NNP	Proper noun
		NNV	Verbal noun
		NST	Spatiotemporal noun
PR	Pronoun	PRP	Personal pronoun
		PRF	Reflexive pronoun
		PRC	Reciprocal pronoun
		PRL	Relative pronoun
		PRQ	Wh-pronoun
		PRI	Indefinite pronoun
DM	Demonstrative	DMD	Deictic demonstrative
		DMR	Relative demonstrative
		DMQ	Wh-demonstrative
		DMI	Indefinite demonstrative
V	Verb	VM_VF	Finite main verb
		VM_VNF	Non-finite main verb
		VM_VINF	Infinitive of main verb
		VM_VNG	Gerund of main verb
		VAUX_VF	Finite auxiliary verb
		VAUX_VNF	Non-finite auxiliary verb
		VAUX_VINF	Infinitive of auxiliary verb
		VAUX_VNG	Gerund of auxiliary verb
		VAUX_VNP	Participle noun
JJ	Adjective	JJ	Adjective
RB	Adverb	RB	Manner adverb
PSP	Postposition	PSP	Postposition
CC	Conjunction	CCD	Coordinating conjunction
		CCS	Subordinating conjunction
		CCS_UT	Quotative subordinator
RP	Particle	RPD	Default particle
		CL	Classifier
		INJ	Interjection
		INTF	Intensifier
		NEG	Negation
QT	Quantifier	QTF	General quantifier
		QTC	Cardinal numeral
		QTO	Ordinal numeral
RD	Residual	RDF	Foreign word
		SYM	Symbol
		PUNC	Punctuation
		UNK	Unknown
		ECH	Echo word

Table 3.11: Categories defined in the Bureau of Indian Standards (BIS) tagset.

efforts. In version 3.012,[11] Interset covers 68 tagsets of 41 languages and defines 63 different features (including the main part-of-speech category) and 386 feature values.

(Petrov et al., 2012) focused just on the parts of speech, assuming that harmonizing these categories will be sufficient for many downstream NLP tasks. They proposed a set of 12 universal POS classes (sometimes dubbed Google UPOS, referring to the affiliation of the authors). Conversion tables from a number of other tagsets, especially those of the treebanks from the CoNLL 2006 and 2007 shared tasks, were also provided. However, the conversion was often based on the names of the categories and did not reflect their internal definition. For instance, some tagsets classify ordinal numerals as a special type of numerals, others as a special type of adjectives. The conversion tables half-blindedly copy the top-level category and do not attempt to put ordinal numerals in one target category, even though the source tagset is fine enough to distinguish them from other words. Instead, they will be tagged `ADJ` or `NUM`, depending on the preferences of the source tagset.

The morphological layer of Universal Dependencies (Nivre et al., 2016) combines an extended set of the universal POS tags and selected feature-value pairs from Interset. There are 17 UPOS categories that should be sufficient for any natural language; if an additional distinction is needed, it should be encoded as a feature. Features, not UPOS, are extensible. A set of core features and values is defined in the UD guidelines but additional language-specific and task-specific features or values may be added when necessary. Unlike the conversion tables supplied with Google UPOS, the UD guidelines try to provide a cross-linguistic definition of each category, and it is assumed that a conversion procedure will respect the definition. Ordinal numerals should be tagged `ADJ` and use the feature `NumType=Ord`, even if other tagsets group them with cardinal numerals. Table 3.12 compares the Google and Universal Dependencies versions of UPOS and Table 3.13 provides an overview of the 21 features defined in UD v2. There are no restrictions at the universal level on which feature can appear with which UPOS, although individual languages may have such restrictions.

When language-specific features and values are included (and when every layer of layered features is counted separately), the release 2.2 of UD data contains 85 distinct features and 411 distinct feature-value pairs.

3.3.4 UniMorph

Another recent attempt to cover morphology of all languages is UniMorph (Kirov et al., 2018; Sylak-Glassman, 2016).[12] It defines over 300 atomic tags called "features" and organized along 25 "dimensions of meaning" (see Table 3.14). In the terms of Interset and Universal Dependencies, UniMorph dimensions of meaning correspond to

[11] `https://metacpan.org/pod/Lingua::Interset`

[12] `https://unimorph.github.io/`

Google	Category	UD	Category
NOUN	Noun	NOUN	Common noun
		PROPN	Proper noun
VERB	Verb	VERB	Main verb
		AUX	Auxiliary verb or particle
ADJ	Adjective	ADJ	Adjective
PRON	Pronoun	PRON	Pronoun
DET	Determiner	DET	Determiner
ADV	Adverb	ADV	Adverb
ADP	Adposition	ADP	Adposition
CONJ	Conjunction	CCONJ	Coordinating conjunction
		SCONJ	Subordinating conjunction
NUM	Numeral	NUM	Numeral
PRT	Particle	PART	Particle
		INTJ	Interjection
X	Other	X	Other
		SYM	Non-punctuation symbol
.	Punctuation	PUNCT	Punctuation

Table 3.12: UPOS: The universal part-of-speech tags, Google and UD version. Examples of the X category are foreign words. The original Google proposal also included typos and abbreviations in X, while in UD these should use the category of the unabbreviated correct word.

Feature	Values
PronType	Art Dem Emp Exc Ind Int Neg Prs Rcp Rel Tot
NumType	Card Dist Frac Mult Ord Range Sets
Poss	Yes
Reflex	Yes
Foreign	Yes
Abbr	Yes
Gender	Com Fem Masc Neut
Animacy	Anim Hum Inan Nhum
Number	Coll Count Dual Grpa Grpl Inv Pauc Plur Ptan Sing Tri
Case	Abs Acc Erg Nom Abe Ben Cau Cmp Cns Com Dat Dis Equ Gen Ins Par Tem Tra Voc Abl Add Ade All Del Ela Ess Ill Ine Lat Loc Per Sub Sup Ter
Definite	Com Cons Def Ind Spec
Degree	Abs Cmp Equ Pos Sup
VerbForm	Conv Fin Gdv Ger Inf Part Sup Vnoun
Mood	Adm Cnd Des Imp Ind Jus Nec Opt Pot Prp Qot Sub
Tense	Fut Imp Past Pqp Pres
Aspect	Hab Imp Iter Perf Prog Prosp
Voice	Act Antip Cau Dir Inv Mid Pass Rcp
Evident	Fh Nfh
Polarity	Neg Pos
Person	0 1 2 3 4
Polite	Elev Form Humb Infm

Table 3.13: Universal features defined in the Universal Dependencies v2 guidelines. For details on individual features, see the guidelines at `http://universaldependencies.org/u/feat/index.html`.

feature names and UniMorph features correspond to feature values. However, UniMorph feature values are unique and do not have to be qualified by the dimensions.[13] For example, the UD value Inv is ambiguous and must be qualified by the feature name to distinguish the inverse number (Number=Inv) from the inverse voice (Voice=Inv). In contrast, UniMorph uses the value INVN in the number dimension, and INV for voice. Every word form can be decomposed to a lemma and a "bundle of UniMorph features", e.g., Spanish *hablé* "I spoke" is represented by the lemma *hablar* "to speak" and the bundle [V;FIN;IND;PFV;PST;1;SG] (verb, final, indicative, perfect, past, first person, singular).

Features of one dimension can be combined if necessary—one could view it as splitting the dimension to several subdimensions. For instance, Uralic languages have a complex system of morphological cases expressing location and movement. Hungarian *ház* "house" can be inflected to *házban* "in the house", *házba* "into the house" and *házból* "out of the house". In UD, these three cases have distinct names and feature values: inessive (Case=Ine), illative (Case=Ill) and elative (Case=Ela), respectively. UniMorph decomposes the system to two subdimensions, location and movement. It uses the case feature IN to identify the location in the house (as opposed to "next to the house", "on top of the house", "under the house" etc.), and another case feature to specify movement: inessive will be [IN;ESS] (no movement), illative [IN;ALL] (movement to the location), and elative [IN;ABL] (movement from the location).

Some UniMorph features are templatic and correspond to a combination of other features. They occur in places where UD would introduce layered features: argument and possessor marking. A noun may have a morpheme signalling that the noun is possessed "by me" (first person singular possessor); UniMorph will tag it with the feature PSS1S. If the noun is instead possessed "by them", it will be flagged PSS3P (third person plural possessor). Examples can be taken again from Hungarian: *házam* "my house", *házuk* "their house".

3.4 How to Define a Part-of-Speech Category

Various grammatical traditions distinguish different sets of part-of-speech categories and may also provide different definitions of a same-named category. Traditional definitions are often based on a mixture of morphological, syntactic and semantic criteria. It is sometimes useful to oversimplify and utter approximations like "nouns are words that denote persons, animals or things," and "verbs denote actions, events or states". But statements of this kind never provide sufficient means to classify all words in a language. Thus it is obvious that the words *child*, *dog* or *rock* are nouns in English, and it is very likely that their translations in other languages will be nouns as well. But it is impossible to extend a purely semantic classification to abstract concepts like *love* or *colonization*. These words denote states and events, a semantic category prototypically associated with verbs. Yet we want to classify them as nouns in English

[13] The only exception is the PROX feature. It appears both in the case and in the deixis dimension; but it seems to be an oversight in the first draft of the specification rather than intention.

Dimension	Values (“features”)
Aktionsart	STAT DYN TEL ATEL PCT DUR ACH ACCMP SEMEL ACTY
Animacy	ANIM INAN HUM NHUM
Argument Marking	ARG[NO|AC|AB|ER|DA|BE][1|2|3][S|P]
Aspect	IPFV PFV PRF PROG PROSP ITER HAB
Case	NOM ACC ERG ABS NOMS DAT BEN PRP GEN REL PRT INS COM VOC COMPV EQTV PRIV PROPR AVR FRML TRANS BYWAY INTER AT POST IN CIRC ANTE APUD ON ONHR ONVR SUB REM PROX ESS ALL ABL APPRX TERM VERS
Comparison	CMPR SPRL AB RL EQT
Definiteness	DEF INDF SPEC NSPEC
Deixis	PROX MED REMT REF1 REF2 NOREF PHOR VIS NVIS ABV EVEN BEL
Evidentiality	FH DRCT SEN VISU NVSEN AUD NFH QUOT RPRT HRSY INFER ASSUM
Finiteness	FIN NFIN
Gender / Noun Class	MASC FEM NEUT BANTU1-23 NAKH1-8
Information Structure	TOP FOC
Interrogativity	DECL INT
Language-Specific	LGSPEC1 LGSPEC2…
Mood	IND SBJV REAL IRR AUPRP AUNPRP IMP COND PURP INTEN POT LKLY ADM OBLIG DEB PERM DED SIM OPT
Number	SG PL GRPL DU TRI PAUC GPAUC INVN
Part of Speech	N PROPN ADJ PRO CLF ART DET V ADV AUX V.PTCP V.MSDR V.CVB ADP COMP CONJ NUM PART INTJ
Person	0 1 2 3 4 INCL EXCL PRX OBV
Polarity	POS NEG
Politeness	INFM FORM ELEV HUMB POL MPOL AVOID LOW HIGH STELV STSUPR LIT FOREG COL
Possession	ALN NALN PSSD PSS1S PSS2S PSS2SM PSS2SF PSS2SINFM PSS2SFORM PSS3S PSS3SM PSS3SF PSS1D PSS1DI PSS1DE PSS2D PSS2DM PSS2DF PSS3D PSS3DM PSS3DF PSS1P PSS1PI PSS1PE PSS2P PSS2PM PSS2PF PSS3P PSS3PM PSS3PF
Switch-Reference	SS SSADV DS DSADV OR CN_R_MN SIMMA SEQMA LOG
Tense	PRS PST FUT IMMED HOD 1DAY RCT RMT
Valency	IMPRS INTR TR DITR REFL RECP CAUS APPL
Voice	ACT MID PASS ANTIP DIR INV AGFOC PFOC LFOC BFOC ACFOC IFOC CFOC

Table 3.14: Dimensions of meaning and features defined in the v2 draft of the UniMorph guidelines. For details on individual features, see (Sylak-Glassman, 2016).

because they adhere to the same grammatical rules as *child*, *dog* and *rock*: for instance, they can be accompanied by an article or a preposition. In other languages, we may be able to delimit word classes on morphological grounds: in Czech, nouns inflect for seven cases and two numbers, which holds for translations of all five English examples mentioned above: *dítě* "child", *pes* "dog", *skála* "rock", *láska* "love", *kolonizace* "colonization". Czech verbs can inflect for number but not for case, therefore the examples are not verbs.

A concrete example of translation equivalents that do not preserve the part-of-speech category is the phrase *father's house*. In English, *father* is a noun (we could say that it is in the genitive case but the *'s* is a clitic that attaches to the last word of the noun phrase, thus it is probably better analyzed as a separate word). When the phrase is translated to a Slavic language such as Croatian, there are two options. Either we may preserve *father* as a noun in the genitive; this is necessary if the noun is further modified, as in *kuća mog oca* "house of my father". In other contexts however, it is preferable to translate it with a possessive adjective: *očeva kuća* "father's house". Here, *očeva* is derived from the noun *otac* "father" but it acquires morphology that is doubtlessly adjectival. It inflects for gender in order to agree with the modified (possessed) noun: *očeva kuća* is feminine, while e.g. *očev ranč* "father's ranch" is masculine and *očevo polje* "father's field" is neuter.

While the possessive constructions require special care to identify language-specific borderline between adjectives and nouns, in other languages adjectives may blend with verbs and it may be quite difficult to delimit adjectives as a separate category at all. Examples from (Schachter and Shopen, 2007, p. 18) illustrate that Chinese words corresponding to English adjectives and words corresponding to English stative verbs behave the same way, regardless whether they are used predicatively or attributively:

(7) 那個女孩子漂亮。 *(Nàgè nǚháizi piàoliang.)* "That girl (is) beautiful."

(8) 那個女孩子瞭解。 *(Nàgè nǚháizi liǎojiě.)* "That girl understands."

(9) 漂亮的女孩子 *(piàoliang de nǚháizi)* "beautiful girl"

(10) 瞭解的女孩子 *(liǎojiě de nǚháizi)* "understanding girl"

We can require that (quoting (Schachter and Shopen, 2007)) "assignment of words to part-of-speech classes is based on properties that are grammatical rather than semantic, and often language-particular rather than universal," but we can still assume "that the *name* that is chosen for a particular part-of-speech class in a language may appropriately reflect universal semantic considerations." In other words, if a language has a category that contains the local equivalents of *child*, *dog* and *rock*, we will call the words in the category *nouns*, even if the category does not extend to less prototypical concepts such as *colonization*, which may be expressed using other grammatical means. However, if we want to preserve parallelism to other languages and define a category that has no special status in the grammar of the language (such as the adjec-

tival subcategory of the verbo-adjectival category in Chinese), we may have to resort to semantic criteria.

Borderlines between individual part-of-speech categories are often blurry even within one language, especially if morphology does not help with disambiguation. For example, the English word *that* and its Spanish counterpart *que* both function as relative pronouns, as in the following parallel example where they replace the subject noun phrase:

(11) *Existe toda una gama de virus **que** provoca este tipo de enfermedades.*

(12) *There is a whole range of viruses **that** cause this type of disease.*

Both words can also act as complementizers (i.e., subordinating conjunctions), as in (13) and (14). Their forms are identical to pronouns and it is possible that diachronically they are derived from the pronouns, but their syntactic function has changed significantly: they do not represent an argument or adjunct in the subordinate phrase. It is desirable that we distinguish the two functions and tag the two instances of *que/that* as two different parts of speech. We cannot rely on the lexicon here; instead, we must look at the context, just like the more arbitrary instances of ambiguity discussed in Section 3.2.1.

(13) *Es cierto **que** las viviendas son malas.*

(14) *It is true **that** the houses are bad.*

Finally, the English word *that* can also function as a demonstrative pronoun (replacing a noun phrase) or determiner (modifying a noun phrase), as in (15) and (16). The Spanish word *que* cannot be a determiner; there is a homophonous interrogative-relative determiner but it is spelled *qué*, as in (17).

(15) *I appreciate **that**.*

(16) *I put **that** application in the round file.*

(17) *… para ver de **qué** color era el traje…*
"… to see what color was the costume…"

Here again we may want to distinguish the pronominal *that* from the determiner *that*. Intuitively it seems less urgent because the distribution of these two functions is less divergent than with the pronoun-complementizer distinction: one could also claim that the pronominal usage is an instance of ellipsis, where the word is actually a determiner and modifies an elided noun phrase *(I appreciate that **thing**.)* Nevertheless, there is no measurable and universally valid criterion that would tell us if two functions of a word are "divergent enough". And indeed, some tagsets and corpora put more weight on the lexicon while others resort earlier to the sentential context.

In (McDonald et al., 2013), words in multi-word named entities are tagged as proper nouns even if they are adjectives, articles or prepositions, as in Spanish *La Rioja*,

Aduana Vieja or *Raúl de Zárate*. If the Spanish names were used in English text, it would be natural to tag all the words as proper nouns because the determiner *la* or the preposition *de* do not have their original function in English; even the adjective *vieja* "old" does not work as English adjectives. However, these examples are taken from a Spanish treebank where it is more disputable whether the words lost their part-of-speech category by being used in names.

An extreme case of functional shift is when a word is cited rather than used. In (18), the prepositions *on* and *in* function as labels of words and fill positions normally filled by nouns. Similarly, *yes* in (19) functions as a noun rather than interjection. Again, approaches to tagging of such words differ. Corpora that are more context-oriented will tag these words as nouns. In other corpora (including Universal Dependencies), the annotation guidelines state that cited words shall keep their category from the lexicon.

(18) English: *You should use 'on' with days of week, but 'in' with names of months.*

(19) English: *I am waiting for her 'yes' on the matter.*

When we want to harmonize annotation across multiple languages, it is sometimes impossible to stay true to traditional terminology. There are many traditions and conflicts are not uncommon—a category may have different names, and one name may be used for two different things in two different terminologies. In Section 3.3.1, we saw how the different traditions resulted in different MULTEXT-EAST tags for the Slavic converbs (gerunds, transgressives or adverbial participles). At the same time, the term *gerund* is used in English for words that are closer to verbal nouns, and Spanish *gerundio* could be compared to English present participles (which have the same form as gerunds, but different syntactic distribution).

A good example of varying approaches is the class of pronouns, determiners and related words. The determiner category is not exactly "traditional" or "classical"—it is said to have been first used by Leonard Bloomfield in 1933 (OED, 1989). In classical grammars, determiners would be classed along with either adjectives or pronouns. Nevertheless, it is now a standard category in formal descriptions of English and in English tagsets. It comprises definite and indefinite articles together with other words (such as demonstratives, possessives and quantifiers) that may introduce a noun phrase instead of an article. The category has been applied to Romance languages, which also have articles, although there are some differences: for instance, the Italian possessives are used with a definite article instead of replacing it (*il mio paese*, lit. *the my country*). We can infer that the limit of one determiner per noun phrase is English-specific rather than universal;[14] or alternatively, one could claim that unlike English, the Italian possessives are not determiners. However, these words should

[14] Even English has exceptions from the rule. A few words function like determiners but they also occur together with an article: ***all*** *the people,* ***both*** *the rules,* ***such*** *a thing*. In tagsets of English, such words may be put to a separate class of *predeterminers*.

	Entities	Modifiers of entities	Circumstances
Content	Nouns	Adjectives	Adverbs
Reference	Pronouns	Determiners	Pronominal adverbs

Table 3.15: The system of common pronominal classes and the corresponding content words.

not be mixed with ordinary adjectives because they are referential (deictic), just like pronouns. And they should stay separate from pronouns, because their morphology is closer to adjectives and they modify nouns, just like adjectives.

This reasoning can be extended to other languages, notably the Slavic languages and German, where the category of pronouns traditionally comprises referential words of substantive (nominal) and attributive (adjectival) nature. Here the tradition is still reflected in influential corpora and tagsets, such as the Stuttgart-Tübingen Tagset of German, the Prague tagset of Czech or the MULTEXT-EAST tagsets of Slavic languages. Moreover, the tagset of the Bulgarian BulTreeBank (Simov and Osenova, 2005) extends the category of pronouns to pronominal adverbs such as *къде (kăde)* "where", *кога (koga)* "when", *как (kak)* "how". If annotation is to be harmonized across languages, these locally acknowledged notions of "pronoun" must be adjusted so that their meaning and extent in one language overlaps with pronouns in other languages as much as possible. A somewhat simplified model of pronominal categories and the corresponding non-referential descriptive word classes is given in Table 3.15. Note that in some languages the picture will be more complex; for instance, some Slavic quantifiers tend to form a category of their own, rather than to align with either pronouns or determiners.

3.5 Part-of-Speech Categories

3.5.1 Nouns

Nouns seem to be the most frequent category in all languages. As mentioned above, the prototypical semantics of nouns is to identify persons, animals, things and uncountable substances, but it also extends to abstract concepts, ideas, as well as to properties, actions and events if they can be transformed to a nominalized form.

Nouns describe or name entities that participate in an action as arguments of a verb *(The **dog** chased the **cat**.)* Nouns can also act as adjuncts that describe place, time and other circumstances of an event *(He slept the whole **day** on the **beach**.)* Nouns can be predicates, either with a copula verb, as *to be* in English, or without it as in Russian *(The dog is a **rottweiler**.)* And nouns can modify other nouns *(the house of my **father**)*.

Tagsets often divide the noun category into common and proper nouns. A proper noun is used to identify one particular person, place, institution or product (e.g., *Lon-*

don) instead of using a more general term that describes the type of the entity (e.g., *city*). That does not mean that a proper noun must be globally unique: there is one London in England, one in Ontario, and quite a few others around the world. Proper nouns in English are grammatically distinct because they are inherently definite, thus appearing without the definite article *the*. In other languages (e.g., Czech), the grammatical rules specific to proper nouns affect only the written language, as the initial letter is required to be uppercase. In yet other languages, the writing system does not distinguish lowercase and uppercase letters, and separation of common and proper nouns is a purely semantic distinction. Note that being part of a multi-word named entity does not automatically mean that the word is a proper noun—although there are corpora that pretend the opposite! In the name *Red River*, *red* is an adjective and *river* is a common noun. At least as long as we apply part-of-speech tags to individual words and not to the entire name. The syntactic annotation may tell us that this is a multi-word name, and so may do a special layer of named entity annotation. But at the word level, there is nothing "proper-nounish" on either *red* or *river*.

Orthogonally to the common-proper distinction, nouns in many languages can be classified into classes, partially on semantic grounds and partially arbitrarily. The classes may enforce distinct morphology on the noun, as well as on other words that cross-reference the entity denoted by the noun, or words that modify the noun and must morphologically agree (be congruent) with it: verbs, adjectives, pronouns etc. The classes are reflected in the morphological features of **gender, animacy** or **noun class.**

Somewhat related to noun classes are *classifiers* (also called *measure words*), function words that identify the semantic class of a noun and that must accompany the noun in certain contexts defined by the grammar. For example, in Chinese quantified noun phrases there must be a classifier between the cardinal numeral and the counted noun: in 三項工程 *(sān xiàng gōngchéng)* "three projects", the word 項 *(xiàng)* is a classifier for the class of principles, clauses, tasks etc., but it can also independently mean "a thing, item, sum of money, back of neck". Classifiers may be viewed as grammaticalized nouns and tagged as a special type of nouns, as in the Sinica Treebank (Huang et al., 2000); or they may be tagged as a separate part-of-speech category, as in the Penn Chinese Treebank (Xia, 2000).

Depending on language, nouns can inflect for **number, case** and/or **definiteness.** Nouns can also show the feature of **polarity,** although it is a marginal phenomenon. Since a noun typically denotes a set of entities with certain properties, negation of the noun will denote the complement of the set, as in Czech *nepolitik* "non-politician". In Uralic, Turkic and other languages, nouns can take **possessive suffixes** that cross-reference person and number of the possessor of the entity denoted by the noun. Nouns derived from verbs may show other features inherited from the verb; this will be discussed in later sections.

Nouns have a lot in common with pronouns, and there are tagsets that merge the two categories. In the Sinica Treebank of Chinese (Huang et al., 2000), pronouns are

a subclass of nouns, tagged Nh; a similar approach is taken also in Interset. In the Russian treebank SynTagRus (Boguslavsky et al., 2000), pronouns and nouns are both tagged S and they are not distinguished by subsequent features: the full tag S ЕД МУЖ ВИН ОД is used for both *его (ego)* "him" and *мальчика (mal'čika)* "boy".

In some languages, nouns are not easily distinguished from adjectives and may share a part-of-speech tag. As (Busa, 1980) says about Index Thomisticus, a monumental corpus of medieval Latin: "… the distinction between adjectives and nouns, for example, appeared without doubt to be syntactical … in Latin … no morpheme in the structure of a word ever differentiates an adjective from a noun."

Certain non-finite verb forms (gerunds, masdars,[15] verbal nouns or infinitives) may show morphological and syntactic behavior similar to nouns. Some tagsets will tag them as verbs, some as deverbal nouns (with a nominal lemma), and in some corpora the context may be used to decide whether individual occurrences are nouns or verbs. See also Section 3.5.2.

Nouns denoting locations in space or time are sometimes difficult to tell apart from adverbs. In the English Penn Treebank, the word *tomorrow* is tagged as a noun (NN). It occurs in some noun-like positions, such as possession and prepositional phrases *(tomorrow's, people of tomorrow, scheduled for tomorrow)*. It can even function as an object or subject: *Tomorrow never dies.* However, it does not occur with an article, which is otherwise typical of English nouns; when used as a temporal adjunct, it is quite similar to adverbs, such as *now*: *It begins tomorrow/now.* With respect to spatial locations, the noun *home* can be used as an adverb in *I go home*; other nouns would require a preposition in this context. However, here the Penn Treebank uses the adverb tag (RB) for *home*. In contrast to English, the Czech word *zítra* "tomorrow" is clearly an adverb. It cannot inflect for case; if we need the "tomorrow" meaning in a nominal function, we must employ morphological derivation and create the noun *zítřek*: *Zítřek nikdy neumírá* "Tomorrow never dies." In the common POS tagset for Indian languages (Lata et al., 2010), spatio-temporal nouns have a separate tag N_NST. There is a good reason to acknowledge a special status of these nouns: they are often used in compound postpositions. For instance, Hindi ऊपर *(ūpara)* could be translated as "upper side", and it is used in the compound postposition meaning "on top of": बस के ऊपर *(basa ke ūpara)* (lit. *bus of upper-side*) "on top of the bus". Another example of a noun grammaticalized into a secondary adposition is the Czech *prostřednictvím* "by means of", originally the instrumental form of the noun *prostřednictví* "mediation, instrumentality".

[15] *Masdar* is an Arabic term for verbal nouns. It has been proposed as a general term for such words in linguistic typology.

3.5.2 Verbs

Verbs are the core part-of-speech category in all languages. They denote actions, events or states, although the semantics is not a sufficient criterion—actions, events and states can also appear in the form of nouns.

Verbs, and especially finite verbs are the prototypical predicates (although many languages have also non-verbal clauses where the core of the predicate is another part of speech): *The dog **chased** the cat.* A phrase headed by a predicate is called *clause* and it can fill subordinate functions in other, larger clauses. Clauses can be subjects *(**Eating** here makes you feel at home)*, object-like complements *(I think she **has** no interest in me)*, adverbial adjuncts *(I drove there to **place** my order)* or modifiers of nouns *(the pleasure of **learning** the language; the need to **write** a review)*. Verbs usually license one or more *arguments* in a particular surface form (such as morphological case or preposition) and assign it a semantic role. (Some other parts of speech, such as nouns and adjectives, can sometimes license arguments as well.)

Some verbal forms in some languages are *periphrastic*, i.e., they are combinations of the main verb and one or more auxiliary words; the auxiliaries may be other verbs, or they may be uninflectable particles. If the language has verbal auxiliaries, its tagset may distinguish main and auxiliary verbs. This distinction is typically context-dependent, as it is not uncommon that a verb can be used as an auxiliary *(Mary **has** left)* or as a main verb *(Mary **has** a baby.)* (Huddleston and Pullum, 2002, p. 92) list 4 non-modal *(be, have, do, use)* and 8 modal *(can, may, will, shall, must, ought, need, dare)* auxiliary verbs in English. They say that "auxiliaries differ very strikingly from lexical verbs in their syntactic behaviour" and offer four diagnostic tests for determining whether a verb is auxiliary. Similar tests are not necessarily available in all languages, and corpora sometimes disagree in what verbs are classified as auxiliary. For example, the modal verbs may be treated differently in different languages. Likewise, copula verbs (such as *to be* in *He **is** a teacher.*) may receive a dedicated tag, although their function is very similar to that of auxiliaries: they provide a (nonverbal) predicate with verbal features.

A central distinction in verbal morphology is the borderline between finite and non-finite forms. Finiteness is not well-defined crosslinguistically (Koptjevskaja Tamm, 1993, p. 29) and (Haspelmath, 1995, p. 4–7); it rather seems to be a "language-specific constellation of syntactic properties" (Sylak-Glassman, 2016, sec. 5.10 p. 26); the usual characteristics of finite verbs include taking nominal subjects in nominative, ergative or absolutive case, morphological cross-referencing of gender, number and person of an argument of the verb, and ability to govern independent clauses (while nonfinite clauses are mostly subordinate). Infinitives and participles are prototypical nonfinite verb forms; other forms are variously termed supines, converbs, transgressives, gerundives, gerunds, masdars or verbal nouns. The terminology differs depending on their function and behavior in each language, as well as on the selection of the

other forms that the language distinguishes. Non-finite forms often have a mixture of verbal properties and properties of another part of speech: participles can be viewed as verbal adjectives, masdars (and sometimes also infinitives) as verbal nouns, and converbs as verbal adverbs. It thus depends on the local tradition and on the tagset design whether their main tag is still that of verb, or they are classified as a special subclass of the other part of speech, or they are even granted a part-of-speech category of their own. If they are not tagged as verbs, they are treated as words morphologically derived from verbs. That also means that their lemma may change. For example, the common citation form of verbs in German is the infinitive, e.g., *entsprechen* "to correspond". If the present participle *entsprechend* "corresponding" is considered just an inflection of the verb, its lemma will be the infinitive. However, if it is tagged as a derived adjective, its lemma will be *entsprechend* and it will cover only the other forms of the adjective, such as *entsprechende, entsprechendes, entsprechendem* etc.

Morphological features typical for verbs are **mood, tense, aspect, voice** and **evidentiality**. In some languages, verbs can contain **negative** or **interrogative** morphemes. Finite verbs may cross-reference their subject and other arguments by mirroring their nominal and pronominal features: **person, number, gender, animacy, politeness** and **clusivity**. In addition, non-finite verb forms may acquire nominal features such as case, definiteness or possession.

3.5.3 Adjectives

Adjectives prototypically denote properties or states of entities (nouns). There are two main ways of connecting adjectives with nouns: *attribution (The **big** dog chased the cat.)* and *predication (The dog that chased the cat was **big**.)*

In some languages, adjectives can morphologically express the **degree** of the property they denote, either absolute, or, more commonly, relative in comparison to the same property of other entities *(big, bigger, biggest)*. Similarly, they can be negated, thus expressing **polarity** *(necessary, unnecessary)*. In addition, adjectives may inflect for various nominal features in order to show agreement with the noun they modify. This type of agreement is not attested in English but it occurs in other languages. For instance, Polish adjectives agree with nouns in **gender, number** and **case:** *wysoki dąb* "a tall oak"; *wysokie dęby* "tall oaks"; *pod wysokimi dębami* "under the tall oaks"; *wysoka jodła* "a tall fir" etc. In Arabic, adjectives will also reflect the nouns' **definiteness.**

Language-particular morphology may reveal differences between adjectives and nouns that are connected to the same lexical meaning. Adjectives can be derived from nouns, including proper nouns, as in Czech *Jizera* (name of a river) → *jizerský*: *Jizerské hory* "Jizera Mountains" (while in English the first word in the name of the mountain range is still the same proper noun, in Czech it is an adjective). Conversely, some nouns are derived from adjectives: *clever* → *cleverness*.

A special kind of adjectives derived from nouns are possessive adjectives, like the Croatian *očev* "father's", discussed in Section 3.4. These should not be confused with

deictic possessives, which may be called in various grammatical traditions possessive pronouns, possessive determiners but also possessive adjectives (Italian *mio* "my"). We discuss deictic possessives in Section 3.5.5.

Another special kind of adjectives are ordinal numerals. Many tagsets actually do not classify them as adjectives; they are traditionally clustered with cardinal numbers, from which they are often (but not always) derived. The traditional concept of numerals is defined semantically as anything pertaining to a definite quantity. Syntactically however, ordinals behave like adjectives. They denote the rank of an entity, and rank is just another type of property. If a language has distinctive adjectival morphology, ordinal numerals are likely to use it as well. Thus in Polish we can observe agreement between the ordinals and the ranked nouns: *szósty dąb* "the sixth oak", *pod szóstym dębem* "under the sixth oak", *szósta jodła* "the sixth fir".

Participles, or verbal adjectives, may be viewed either as part of the verbal paradigm, or as adjective-like words derived from verbs. They can be used predicatively or attributively, although some participial forms in some languages can only be used as predicates. If the language has distinctive adjectival morphology, such as gender agreement with nouns, participles are likely to inflect similarly. In addition to adjectival features, participles typically show some verbal features, too: for instance, the Russian participles in (20) and (21) are active, (22) and (23) are passive; (20) and (22) are in the present tense, while (21) and (23) are in the past tense; and (23) has the perfective aspect, while the other examples are imperfective.

(20) *студент, **читающий** журнал (student, **čitajuščij** žurnal)* "student that is reading a journal"

(21) *студент, **читавший** журнал (student, **čitavšij** žurnal)* "student that was reading a journal"

(22) *журнал, **читаемый** студентом (žurnal, **čitaemyj** studentom)* "journal that the student is reading"

(23) *журнал, **прочитанный** студентом (žurnal, **pročitannyj** studentom)* "journal that the student has read"

On the syntactic level, participles commonly take arguments, which is not so common (but possible) with ordinary adjectives.

As mentioned in Sections 3.4, 3.5.1 and 3.5.2, in some languages all adjectives (and not just participles) are hard to tell apart from verbs, and in other languages, adjectives may be very similar to nouns.

3.5.4 Adverbs

Adverbs prototypically denote circumstances of events and states, such as location, time, manner, cause etc. They also denote degree or extent. Most commonly, adverbs modify clausal predicates *(Do it **here** and **now**, and do it **well**!)*; but note that the same

function can be also fulfilled by noun phrases *(Do it at this place, in this moment and without any mistakes.)* Adverbs can also modify adjectives *(**very** good)* or other adverbs *(**very** well)*. Certain adverbs may also be used with noun phrases (including adpositional phrases) to emphasize them in one or another way *(**especially** the king)*.

Some adverbs in some languages can inflect for the **degree of comparison** just like adjectives: Czech *chytře, chytřeji, nejchytřeji* "cleverly, more cleverly, most cleverly". Similarly, adverbs may show **polarity** (English *necessarily, unnecessarily*).

In Section 3.5.1 we showed how spatial and temporal adverbs overlap with nouns. In good many languages, adverbs are also easily confused with adjectives. In German, for example, adjectives take agreement suffixes when they are used attributively (cf. the feminine form *drastische* "drastic" in (24)) and they omit the suffix in predicate position (*drastisch* in (25)). Nevertheless, the same form can also be used as an adverb (26).

(24) *eine **drastische** Änderung* "a drastic change"

(25) *Die Änderung ist **drastisch**.* "The change is drastic."

(26) *Es hat sich **drastisch** geändert.* "It changed drastically."

In colloquial English, one can observe adjectives used in place of adverbs, as in (28); the standard English version is in (27).

(27) *He gets along **well** with his co-workers.*

(28) *He gets along **good** with his co-workers.*

Converbs (also called transgressives, gerunds, verbal adverbs or adverbial participles), may be viewed either as part of the verbal paradigm, or as adverb-like words derived from verbs. They modify finite verbs or other predicates, providing another event as a circumstance of the main event. Unlike ordinary adverbs, they may show some verbal features such as tense or aspect. For instance, Russian (29) is in the imperfective aspect, (30) is perfective. It is also customary to classify the former as the present and the latter as the past tense, although here the reference point is the event of the main clause rather than the moment of the utterance.

(29) ***Читая** книгу, Маша думала о своих друзьях. (**Čitaja** knigu, Maša dumala o svoih druz'jah.)* "While reading a book, Masha thought about her friends."

(30) ***Прочитав** газету, я лёг спать. (**Pročitav** gazetu, ja lëg spat'.)* "Having read a newspaper, I went to bed."

On the syntactic level, converbs commonly take arguments, which cannot be said about ordinary adverbs.

Adverbs also overlap with numerals and pronouns. There are quantitative or ordinal modifiers whose syntactic distribution is adverbial, and depending on linguistic

tradition, they may be tagged as numerals or adverbs. Examples include Czech *poprvé* "for the first time" and *třikrát* "three times". There is also a class of pronominal or deictic adverbs such as *where, when, how, there, then, so*. In some tagsets, these adverbs may be classified as pronouns (see also Section 3.5.5).

Negative particles, such as English *not* or German *nicht*, may be considered an extreme case of degree adverbs, as they modify the clause predicate and set the degree of the state or action to zero. Treating them as adverbs seems to be traditional especially in the Romance languages; others may prefer a tag for particles.

Speaking of particles, there is a larger and vaguely defined group of words that may be considered adverbs (because they are modifier words at the clause level) but some authors argue that their communicative function is sufficiently different from adverbs and that they should be classified as particles. Examples include words like *unfortunately* or *only*: they express the attitude of the speaker towards the event, rather than a circumstance of the event.

Finally, some adverbs have grammaticalized as conjunctions. For instance, English *so* is an adverb in *She is so beautiful!* but it is more like a conjunction in *She came late, so I could not show her the sunset over the lake.*

3.5.5 Pronouns, Determiners and Quantifiers

Words covered in this section are almost always distributed into multiple categories but the boundaries are fuzzy, and differences between tagsets may be dramatic. Pronouns, as their name suggests, are words that can be substituted for nouns. Instead of describing or naming the entity, they *refer* to an entity supposedly known or imaginable by the speaker and the addressee. This situation-dependent referring is called *deixis*, hence pronouns are sometimes characterized as deictic words. Personal pronouns are the core kind of pronouns: they refer to the speaker *(I)*, the addressee *(you)* or somebody / something else *(he, she, it)*. There may be also an impersonal pronoun for general statements, such as German *man* and French *on* (their usual English translation is "one", as in *One does not normally go this way.*) In object position, languages have reflexive pronouns (English *himself*) and reciprocal pronouns (German *einander* "each other").

Interrogative pronouns *(who, what)* are used to refer to an unknown entity whose identity we demand in questions. Various other pronoun types function as referents for unknown entities in declarative sentences (indefinite pronouns: *somebody, something, anybody, anything*), referents for all possible entities *(everybody, everything)* or even for excluding all entities in negative sentences *(nobody, nothing)*.

Languages often have another set of words whose deictic functions are parallel to those just listed, but their morphosyntactic behavior resembles adjectives rather than nouns, and they can be used to modify a noun phrase rather than to replace it (though many of them can still occur without a modified noun, which can be understood as an elided generic entity). Depending on language and tagset, these words may be still

labeled as pronouns, or put in a separate category of *determiners.* English examples would include the words *which, some, any, every, no*. In languages where adjectives show morphological agreement with nouns they modify, determiners are likely to do the same; especially if the language distinguishes genders and a deictic word can inflect for gender, it is likely a determiner. Let's take once again some Polish trees as examples: *każdy dąb* "every oak", *każda jodła* "every fir".

A very common and somewhat special class of deictic words are demonstratives, i.e. words "pointing" at a particular entity, either in the previous discourse, or in the scene visible to both the speaker and the addressee, often also indicating whether the entity is close or distant. Demonstratives in many European languages are determiners because they modify noun phrases *(this dog, that cat)* but they can also stand alone like pronouns *(I have this and you have that)*. In some languages, demonstratives also perform the function of third-person pronouns, e.g. in Hindi यह *(yaha)* "this, he, she, it", वह *(vaha)* "that, he, she, it".

Another special class are relative pronouns (or determiners). They occur in subordinate clauses modifying a noun phrase, and they represent the modified noun phrase within the structure of the modifying clause *(the dog **that** chases the cat)*. Like in the other classes, some relative words may have distinctive adjectival morphology and thus show affinity to determiners; yet in the syntactic structure, they stand alone without the modified noun (because that noun is in a superordinate clause), as in Polish (31). In contrast, Hindi uses a different pattern where the modified noun may occur within the subordinate clause (32). Also note that while English and Polish relatives overlap with demonstratives and interrogatives, respectively, Hindi is an example of a language where relatives form a distinct set, separate from other classes.

(31) *dąb, pod **którym** siedzimy* "the oak under which we sit"

(32) जो आदमी बाहर खड़ा है वह विदेशी है। *(**jo** ādamī bāhara khaṛā hai vaha videśī hai .)* (lit. ***which** man outside standing is that foreigner is .*) "The man who is standing outside is a foreigner."

Possessive pronouns (or determiners) relate to personal pronouns. They too have different forms for different persons, but here it is the person of the possessor. In some languages, possessive pronouns are simply the personal pronouns in the genitive case (Japanese 私 *(watashi)* "I", 私の *(watashino)* "my"). In other languages, possessives have adjective-like morphology and agree with the modified (possessed) noun—cf. the Croatian possessives in Table 3.6. They would thus fall into what we call determiners here. English possessives possess no morphology that would clearly mark them as determiners, but they do modify (rather than replace) noun phrases and they also replace articles, just like other determiners in English (but not necessarily in other languages). Despite the evidence, even the English tradition calls these words "possessive pronouns" and they may be classified as such in tagsets.

Articles are sometimes awarded a category of their own, sometimes they are subsumed under determiners. As mentioned earlier, in English they have a similar distribution with other determiners (that is, a noun occurs either with an article, or another determiner, but not with both); in other languages however, this does not hold (at least not if we use the term for all deictic, noun-modifying words). Articles can be described as determiners that contribute the single feature of definiteness. It is not uncommon that definite articles resemble demonstratives (cf. English *the – this – that*) and indefinite articles are ambiguous with the number one (cf. German *ein* "one / a").

In a broader sense, cardinal numbers *(one, two, three)* can be viewed as determiners as well; but in tagsets they are practically always defined as a separate category. However, there are other quantifiers that are deictic and refer to indefinite quantities. These may be tagged as a special type of numerals, as in the Prague tagset of Czech, or, more commonly, as a type of determiners or pronominal adverbs. English examples are *many* and *few*. A wider range can be observed in Czech where we have an interrogative quantifier *kolik* "how many / how much", demonstrative *tolik* "so many / so much", indefinite *několik* "several", as well as *mnoho* "many / much" and *málo* "few / little".

Besides cardinals, there are other types of number-based expressions that are—again depending on grammatical tradition—either treated as subclasses of numerals, or as subclasses of other categories whose morphosyntactic behavior they resemble. We have thus seen adjectival ordinal numerals in Section 3.5.3 (Czech *první, druhý, třetí* "first, second, third") and adverbial ordinal numerals in Section 3.5.4 (Czech *poprvé, podruhé, potřetí* "for the first time, for the second time, for the third time"). Similarly, there are multiplicative numerals that work like adjectives (*dvojí, trojí* "twofold, threefold") and those that work like adverbs (*jednou, dvakrát, třikrát* "once, twice, three times"). Special type of cardinal numerals may be used for counting sets, such as pairs of shoes (Czech *jedny, dvoje, troje* "one set of, two sets of, three sets of"; cf. the standard cardinals *jeden, dva, tři*). Some cardinals, especially those with high values, may be undistinguishable from nouns: they form plurals and occasionally appear without the counted noun, as in English *Thousands protest peacefully in London.* Most of the more peculiar numeral types have corresponding interrogatives, demonstratives and indefinites, e.g. *kolikátý* (question word asking for an ordinal numeral, i.e., "what rank"), *pokolikáté, kolikerý, kolikrát, kolikery* etc. These words may be tagged either as a subtype of numerals (the Prague Czech tagset) or as determiners and pronominal adverbs (Universal Dependencies).

Finally, adverbs themselves have a pronominal (deictic) subclass, consisting of interrogatives / relatives *(where, when, how, why)*, demonstratives *(here, there, now, then)*, indefinites *(somewhere, sometime, somehow)*, universals *(everywhere)* and negatives *(nowhere, never)*.

Before leaving this very diverse section, let us now turn to a language family that is very different from English or Czech, namely to the Philippine branch of the Aus-

tronesian languages. Noun phrases in these languages are often introduced by function words that serve multiple purposes. For instance, the Tagalog sentence (33) contains phrase markers *ng*, *ang* and *sa*.

(33) *Aalisan* ***ng*** *babae* ***ng*** *bigas* ***ang*** *sako para* ***sa*** *bata.* "A/the woman will take some rice out of the sack for a/the child."

One of their functions is pragmatic, the function word *ang* marks the topic of the sentence. However, as the topic-focus articulation is one of the main principles around which the Philippine-type grammar is organized, these words can also be said to mark the various arguments of the verb; as such, they are similar to prepositions in European languages. The phrase markers may also bear a trace of definiteness, as the topic noun phrase is always definite (but the other phrases may be definite or indefinite). This bit makes the words similar to determiners in European languages. Both determiners and prepositions can be broadly defined as function words that introduce noun phrases and supply them with additional features or specify their role in the clause. Both categories could be extended to the Philippine phrase markers and can be found in the literature, e.g., (Schachter and Shopen, 2007, p. 35) and (Dryer, 2007, p. 121).

3.5.6 Adpositions, Conjunctions, Linkers and Particles

Adposition is an umbrella term for prepositions and postpositions. Prepositions occur at the beginning of a noun phrase, postpositions at the end; languages usually have strong preferences towards one of these types but there may be exceptions. For instance, English is a prevailingly prepositional language, but in *two days ago*, the word *ago* is a postposition. Most adpositions are *case markers*, i.e., their function is similar to that of morphological case: they help specify the role of the noun phrase as an argument of a predicate, or its relation to another constituent, its spatiotemporal location, movement etc.

Japanese postpositions are traditionally called particles but many of them are case markers just like European cases or prepositions: for example, を *((w)o)* marks the direct object and corresponds to the accusative case; に *(ni)* can be translated as the dative or indirect object; and の *(no)* corresponds to the genitive case or the English preposition "of".

Philippine-type languages have phrase markers that can be classified as either prepositions or determiners; see Section 3.5.5 for details.

Verbal particles or separable verb prefixes in Germanic languages are often prepositions or adverbs by origin. It depends on how much context-oriented the tagset is whether they retain the original tag or get a new, language-specific category. Syntactically, they form compounds with verbs, despite the fact that they are not necessarily adjacent (English *pick it* ***up***, German *zieh dich* ***an*** "get dressed").

Primary adpositions tend to be short and very frequent; they usually rank among the most frequent words in the language. Some languages also have secondary adpositions (such as the Czech *prostřednictvím* "by means of" discussed in Section 3.5.1) and compound adpositions; the latter are multi-word expressions, typically composed of prepositions and nouns (e.g. English *in contrast to*) and as long as they are written as multiple words, each word gets its part-of-speech tag individually. The same holds for circumpositions, i.e., fixed combinations of a preposition and a postposition, e.g., ***from** that moment **on***.

Conjunctions are words that connect words, phrases or clauses into larger constituents. They are divided into coordinators (which connect same-level constituents) and subordinators (which mark one constituent as dependent on the other). Subordinating conjunctions resemble adpositions but unlike adpositions, they are typically used with clauses rather than noun phrases. Prototypical subordinators are *that, if, because*; sometimes an adposition can be used with a clause as a subordinator too, as in English *I have to vacuum the room **before** she returns home.* Infinitive markers such as English *to* and German *zu* are similar to subordinators, although a tagset may choose to tag them as particles (or prepositions, because that is what these two examples originally were).

Prototypical coordinating conjunctions are *and, or, but* and their equivalents. They may connect clauses, noun phrases, adjectives, adverbs and even function words such as prepositions *(There are regular connections **from and to** Berlin.)* There is a strong tendency though that the connected constituents are of same or at least compatible type. Languages have other conjunctions that are classified as coordinators, but sometimes the borderline between coordination and subordination appears blurry or arbitrary. For example, the Czech conjunctions *protože* and *neboť* are both equivalents of English *because*, but the former is subordinating and the latter coordinating. Similarly, a preposition can sometimes be used to convey the same meaning as coordination, cf. *Petr a Pavel* "Peter and Paul" vs. *Petr s Pavlem* "Peter with Paul". Some coordinating conjunctions are multi-word expressions *(either-or, both-and, neither-nor)*.

Linkers are function words or morphemes that mark relation between words and that are required by certain grammatical constructions in some languages. This broad definition makes them a generalization of adpositions and conjunctions; however, the term *linker* usually denotes words that are not considered adpositions or conjunctions by the traditional grammar. For example, in Ilokano (34) (Rubino, 1998), the linker *nga* links an adjective to the noun it modifies.

(34) *ti nalaíng **nga** ubíng* (lit. *the smart LINK child*) "the smart child"

Languages have various other function words that are either categorized as particles, or receive dedicated, language-specific tags. Some of them connect or modify phrases, others operate at clause or sentence levels. To name just a few, Chinese has a multi-purpose particle that is pronounced *de* but written variously 的, 得 or 地, depending on its function. For instance, 的 is often used in situations where English

would use the preposition *of*. Many languages have auxiliary particles that function similarly to auxiliary verbs, but they are invariant and not verbal. Examples include the Slovak *by* for conditional, or Greek *ζα (za)* for future tense. There are negative particles such as English *not* and German *nicht* (we noted in Section 3.5.4 that some tagsets will treat them as adverbs). And many languages have question particles that mark yes-no questions: Polish *czy*, Hindi क्या *(kyā)*, Japanese か *(ka)*.

3.5.7 Interjections and Onomatopoeia

Interjections are words that express a spontaneous feeling or reaction. They include exclamations *(ouch!, wow!)*, curses *(damn!)*, greetings *(hi, bye)*, response particles (although some tagsets will tag these as particles: *yes, no, okay*) and hesitation markers *(uh, um)*. In some grammatical traditions, interjections include the related category of onomatopoeic words, while in others this category will be separate. Onomatopoeia refers to phonetic imitation of a sound in the word that describes or represents the sound. Languages have words that are commonly used to represent the noises made by certain animals *(woof, miaow, oink)* and things *(tick tock, beep-beep, vroom)*.

Some exclamations originate as words of other categories and depending on tagset, they may be tagged as interjections or as the original part of speech *(Jesus!, help!, fuck!, thanks!)* Other exclamations are multi-word expressions and like in other MWEs, the individual words may not be interjections *(Excuse me! Oh my God!)*

Interjections and onomatopoeic words tend to be morphologically invariant (except for inherited—and usually frozen—morphology of secondary interjections). Since they form an independent utterance, they are not tightly integrated as clause constituents. One exception is when such a word is used instead of a verb as the predicate, as in Czech (35):

(35) *Utrhni jablko a* ***šup*** *s ním do košíku!* (lit. *Pluck apple and whoosh with it in basket!*)
"Pluck the apple and put it in the basket!"

3.5.8 Other

This final section lists some other word-like elements that can be found in language, especially in written language. We do not try to categorize them; some tagsets will provide ad-hoc tags, others may put these words into a residual (or "garbage can") class. The purpose here is merely to point out that they exist.

Occasionally a subword unit appears in text independently. In German, the first part of a compound may represent the compound in coordination even if it cannot be used as an independent word otherwise: ***Landes-*** *und Kreisstraßen* "state and county roads". In other languages a similar situation may arise just because of tokenization; for instance, if the Czech adjective *francouzsko-německý* "French-German" is split to three tokens, the first token *francouzsko* has a suffix that would never occur with a full adjective.

Separate morphemes may occur also when we want to account for two or more morphological variants. For instance, if we want to suggest that both singular and plural readings are possible, we may say *Bring your friend**(s)**!* In languages with gender, we may want to account for multiple gender variants: Czech *nezletilí**/é** studenti**/ky** musí mít svolení rodičů* "underage.MASC/.FEM students.MASC/.FEM must have a consent from their parents".

Some languages work with reduplication, i.e., two copies of a word appear in sequence. Both may get the same tag, or the second copy may get a special tag for reduplicatives. For instance, Indonesian uses reduplication to signal the plural number. In Hindi, this would add the meaning of distribution ("one rupee each"), separation ("sit separately"), variety, diversity or just emphasis: कभी — कभी *(kabhī – kabhī)* "sometimes", whereas single कभी *(kabhī)* also means "sometimes"; एक एक *(eka eka)* "one each", whereas एक *(eka)* means "one".

Hindi also has so-called echo words: The word rhymes with a previous word but it is not identical to it and typically it does not have any meaning of its own. In Hindi it generalizes the meaning of the previous word and eventually translates as "or something", "etc." etc. चाय वाय *(cāya vāya)* "tea or something" (as in "Have some tea or something.") A similar phenomenon can be, even if much less frequently, observed in other languages. The Czech expression *projít křížem krážem* means to criss-cross an area; *projít* is the verb "to go", *křížem* is the instrumental form of *kříž* "cross", but *krážem*, despite looking like a noun in instrumental, never occurs in the language outside this expression.

Finally, there are various symbols, alphanumeric product identifiers, e-mail addresses, time and date specifications. Their classification is very diverse across corpora. They may be clustered with nouns, numbers, punctuation, put in a residual category or, as always, have dedicated ad-hoc tags.

Chapter 6

Some Concluding Tokens

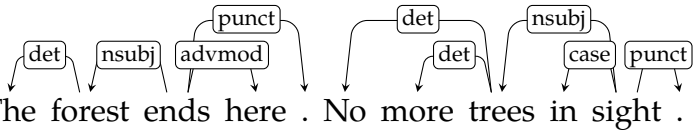
We have attempted to present a cross-linguistic survey of morphological and dependency-based syntactic annotation. We looked at

1. what one may want to annotate in the various languages of the world;
2. how has it been annotated in existing corpora; and
3. what are the advantages and disadvantages of possible approaches.

In the time of the preparation of this study, Universal Dependencies has become the dominant annotation style. While we do discuss other styles a lot, the outcome is inevitably a comparison of UD vs. the rest. Sometimes we argue that the advantages of the UD approach prevail (e.g. subordinate clauses), elsewhere it is not as clear (e.g. copula constructions) or there is no optimal solution (coordination). However, no annotation style is optimal for all purposes. This leads us to one ultimate objective: while picking one annotation approach, we should lose as little information as possible, so that the data can be automatically converted to another style if desirable.

As for the item 1 above, we tried to be rather thorough in the domain of part-of-speech tags and morphological features, and we discussed many features of endangered languages that have little or no machine-readable resources. Still, we admit significant bias towards big languages, and there are morphological peculiarities that would deserve much more attention if the less familiar languages were given equal space. On the level of syntax, we were only able to discuss core phenomena such as noun phrases, verbal and nonverbal clauses, subordination and coordination. Arguably, a monograph larger than this one could be written on syntax alone. This book is not an annotation manual; if it was, then we would have to discuss numerous other constructions, such as ellipsis (only briefly mentioned by us in other sections), comparative constructions, transitivity, reflexivity, valency-changing operations such as passivization, and a lot more. It would be certainly interesting to expand this study into all the missing corners of syntax; however, a good annotation scheme is coherent and follows the same objectives and principles in all decisions it makes. Therefore,

the annotation of the undiscussed constructions more or less depends on the decisions taken in the core parts that we have discussed here.



Summary

This monograph presents a comparative study of annotation approaches to morphology and syntax of natural languages, with emphasis on applicability in a multilingual environment. Annotation is understood as adding linguistic categories and relations to digitally encoded natural language text, resulting in annotated corpus; as syntactic relations are often represented in the form of dependency trees, the annotated corpora covered by the monograph are dependency treebanks. Many treebanks exist and their annotation styles vary significantly, which hampers their usefulness for linguists and language engineers. We survey several harmonization efforts that tried to come up with cross-linguistically applicable annotation guidelines, including the most recent and broadest effort to date, Universal Dependencies. We examine language description on three levels: 1. tokenization and word segmentation, 2. morphology, and 3. surface dependency syntax. For each language phenomenon we provide a comparison of its analysis and annotation in various existing treebanks (or other corpora, for tokenization and morphology), pointing out advantages and disadvantages of the competing approaches. On the morphological layer, we go even beyond the currently available corpora and provide a typological survey of features that will be needed when less-resourced languages are covered by an annotation project. We conclude that no single approach is suitable for all purposes, but a good approach must not lose information, so that annotation can be converted to another style when necessary.

There are grammar descriptions, there are linguistic typological works, and there are annotation manuals for corpora in individual languages. However, there are not many studies that take the corpus-annotation perspective and compare a large number of languages. There is a gap on the market, and this book can fill it.

List of Figures

List of Tables

Bibliography

Alexander Adelaar and Nikolaus P. Himmelmann. The Austronesian Languages of Asia and Madagascar. Routledge Language Family Series. Routledge, Oxon/New York, 2005.

Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. "Floresta sintá(c)tica": a treebank for Portuguese. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC), pages 1968–1703, 2002.

Alexandra Y. Aikhenvald. Evidentiality. Oxford University Press, Oxford, UK, 2004.

Avery D. Andrews. The major functions of the noun phrase. In Timothy Shopen, editor, Language Typology and Syntactic Description. Volume I: Clause Structure, pages 132–223. Cambridge University Press, Cambridge, UK, second edition, 2007. ISBN 978-0-521-58156-1.

Nart B. Atalay, Kemal Oflazer, and Bilge Say. The annotation process in the Turkish treebank. In In Proceedings of the 4th International Workshop on Linguistically Interpreteted Corpora (LINC), 2003.

David Bamman and Gregory Crane. The Ancient Greek and Latin dependency treebanks. In Language Technology for Cultural Heritage, Theory and Applications of Natural Language Processing, pages 79–98. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-20227-8.

Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Monojit Choudhury, Girish Nath Jha, Rajendran S., Saravanan K., Sobha L., and KVS Subbarao. A common parts-of-speech tagset framework for Indian languages. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. URL `http://www.lrec-conf.org/proceedings/lrec2008/summaries/337.html`.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague dependency treebank 3.0, 2013. URL `http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3`. LINDAT/CLARIN digital library at the Institute of

Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. Natural Language Processing – A Paninian Perspective. Prentice-Hall of India, New Delhi, India, 2006a. ISBN 978-81-203-0921-9.

Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. AnnCorra : Annotating corpora guidelines for pos and chunk annotation for indian languages, December 2006b. URL https://www.researchgate.net/publication/268414162_AnnCorra_Annotating_Corpora_Guidelines_For_POS_And_Chunk_Annotation_For_Indian_Languages. [online; accessed 2018-08-29].

D. N. Shankara Bhat. Pronouns. Oxford University Press, Oxford, UK, 2004.

Agnė Bielinskienė, Loïc Boizou, Jolanta Kovalevskaitė, and Erika Rimkutė. Lithuanian dependency treebank ALKSNIS. In I. Skadiņa and R. Rozis, editors, Human Language Technologies – The Baltic Perspective, pages 107–114, 2016. doi: 10.3233/978-1-61499-701-6-107.

Igor Boguslavsky, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Kreidlin, and Nadezhda Frid. Dependency treebank for Russian: Concept, tools, types of information. In Proceedings of the 18th conference on Computational linguistics-Volume 2, pages 987–991. Association for Computational Linguistics, Morristown, NJ, USA, 2000.

Igor Boguslavsky, Leonid Iomdin, Vadim Petrochenkov, Victor Sizov, and Leonid Tsinman. A case of hybrid parsing: Rules refined by empirical and corpus statistics. In Kim Gerdes, Eva Hajičová, and Leo Wanner, editors, Computational Dependency Theory, volume 258, pages 226–240. IOS Press, Amsterdam, Netherlands, 2013. ISBN 978-1-61499-351-3. doi: 10.3233/978-1-61499-352-0-226.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER treebank. In Proceedings of the Workshop on Treebanks and Linguistic Theories, Sozopol, 2002.

Penelope Brown and Stephen C. Levinson. Politeness: Some Universals in Language Usage. Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge, UK, 1987.

Sabine Buchholz and Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing. In Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X), pages 149–164. Association for Computational Linguistics, 2006. URL http://anthology.aclweb.org/W/W06/W06-29.pdf\#page=165.

Roberto Busa. The annals of humanities computing: The index thomisticus. Computers and the Humanities, 14:83–90, 1980. URL http://www.alice.id.tue.nl/references/busa-1980.pdf.

Key-Sun Choi, Hitoshi Isahara, Kyoko Kanzaki, Hansaem Kim, Seok Mun Pak, and Maosong Sun. Word segmentation standard in chinese, japanese and korean. In Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 179–186, Suntec, Singapore, August 2009. ACL and AFNLP. URL http://www.

aclweb.org/anthology/W09-3426.

Mihaela Călăcean. Data-driven dependency parsing for Romanian. Master's thesis, Uppsala University, August 2008. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.6068&rep=rep1&type=pdf.

Bernard Comrie. Linguistic politeness axes: Speaker-addressee, speaker-referent, speaker-bystander. Pragmatics Microfiche, 1.7(A3), 1976.

Bernard Comrie. Tense. Cambridge University Press, Cambridge, UK, 1985.

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. The Szeged treebank. In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, Text, Speech and Dialogue, 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005, Proceedings, volume 3658 of Lecture Notes in Computer Science, pages 123–131. Springer, 2005. ISBN 3-540-28789-2. URL http://dx.doi.org/10.1007/11551874_16.

Irvine Davis. The language of Santa Ana Pueblo (anthropological papers, no. 69). Smithsonian Institution Bureau of American Ethnology, Bulletin, 191(68–74):53–190, 1964.

Marie-Catherine de Marneffe, Natalia Silveira, Timothy Dozat, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal Stanford dependencies: A cross-linguistic typology. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavík, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

R. M. W. Dixon. A Grammar of Yidiny. Cambridge University Press, Cambridge, UK, 1977. ISBN 978-0-521-21462-9.

Matthew S. Dryer. Word order. In Timothy Shopen, editor, Language Typology and Syntactic Description. Volume I: Clause Structure, pages 61–131. Cambridge University Press, Cambridge, UK, second edition, 2007. ISBN 978-0-521-58156-1.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdeněk Žabokrtský, and Andreja Žele. Towards a Slovene dependency treebank. In Proceedings of the Fifth International Language Resources and Evaluation Conference, LREC 2006, pages 1388–1391, Genova, Italy, 2006. European Language Resources Association (ELRA). URL http://hnk.ffzg.hr/bibl/lrec2006/summaries/133.html.

EAGLES. EAGLES. recommendations for the morphosyntactic annotation of corpora, 1996. URL http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html.

Jan Einarsson. Talbankens skriftspråkskonkordans; Talbankens talspråkskonkordans, 1976.

Tomaž Erjavec. MULTEXT-East version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, 2010.

Tomaž Erjavec. MULTEXT-East: Morphosyntactic resources for central and eastern european languages. Language Resources and Evaluation, 46(1):131–142, 2012.

Wolfdietrich Fischer. Classical Arabic. In Robert Hetzron, editor, The Semitic Languages, Routledge Language Family Series. Routledge, Oxon/New York, 1997.

ISBN 978-0-415-41266-7.

William A. Foley. A typology of information packaging in the clause. In Timothy Shopen, editor, Language Typology and Syntactic Description. Volume I: Clause Structure, pages 362–446. Cambridge University Press, Cambridge, UK, 2007. ISBN 978-0-521-58156-1.

GB/T. Xinxi chuli yong xiandai Hanyu fenci guifan (Contemporary Chinese Language Word Segmentation Specification for Information Processing) GB/T 13715-92. Biaozhun chubanshe, Beijing, China, 1993.

Sofia Gustafson-Capková and Britt Hartmann. Manual of the stockholm umeå corpus version 2.0, December 2006. URL https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf. [online; accessed 2018-07-26].

Jan Hajič, Alena Böhmová, Eva Hajičová, and Barbora Vidová Hladká. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Anne Abeillé, editor, Treebanks: Building and Using Parsed Corpora, pages 103–127. Amsterdam:Kluwer, 2000.

Jan Hajič, Otakar Smrž, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. Prague Arabic dependency treebank: Development in data and tools. In Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools, pages 110–117, Cairo, Egypt, May 2004. URL http://ufal.mff.cuni.cz/padt/PADT_1.0/docs/papers/2004-nemlar-padt.pdf.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Prague Czech-English dependency treebank 2.0, 2011. LDC2012T08.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5, Boulder, Colorado, USA, 2009.

Martin Haspelmath. The converb as a cross-linguistically valid category. In Martin Haspelmath and Ekkehard König, editors, Converbs in Cross-Linguistic Perspective: Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerunds, Empirical Approaches to Language Typology, pages 1–56. Mouton de Gruyter, Berlin, Germany, 1995. URL https://www.researchgate.net/publication/238336969_The_converb_as_a_cross-linguistically_valid_category.

Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao, and Kuang-Yu Chen. Sinica treebank: Design criteria, annotation guidelines, and on-line interface. In Second Chinese Language Processing Workshop, pages 29–37, Hong Kong, China, October 2000. Association for Computational Linguistics. doi: 10.3115/1117769.1117775. URL http://www.aclweb.org/anthology/W00-1205.

Rodney Huddleston and Geoffrey K. Pullum. The Cambridge Grammar of the English Language. Cambridge University Press, Cambridge, UK, 2002.

Richard Hudson. Are determiners heads? Functions of Language, 11(1):7–42, 2004. ISSN 0929-998X. URL http://dickhudson.com/wp-content/uploads/2013/07/dets.pdf.

Samar Husain, Prashanth Mannem, Bharat Ambati, and Phani Gadde. The ICON-2010 tools contest on Indian language dependency parsing. In Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing, Kharagpur, India, 2010.

Nancy Ide and Jean Véronis. Multext (multilingual tools and corpora). 1994. URL http://www.aclweb.org/anthology/C/C94/C94-1097.pdf.

Miloš Jakubíček, Vojtěch Kovář, and Pavel Šmerk. Czech morphological tagset revisited. In Aleš Horák and Pavel Rychlý, editors, Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011, pages 29–42. Tribun EU, 2011.

Sylvain Kahane. Bubble trees and syntactic representations. In In Proceedings of the 5th Meeting of the Mathematics of the Language, DFKI, Saarbrucken, 1997.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. UniMorph 2.0: Universal Morphology. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9. URL http://www.lrec-conf.org/proceedings/lrec2018/summaries/789.html.

Maria Koptjevskaja Tamm. Nominalizations. Routledge, London, UK, 1993.

Matthias T. Kromann, Line Mikkelsen, and Stine Kern Lynge. Danish dependency treebank, 2004. URL http://code.google.com/p/copenhagen-dependency-treebank/.

Swaran Lata, Girish Nath Jha, Somnath Chandra, Dipti Misra Sharma, Somi Ram, Uma Maheswara Rao G, Sobha L, Menak S, Kalika Bali, Pushpak Bhattacharyya, Malhar Kulkarni, Lata Popale, Kirtida Shah, Mona Parakh, Jyoti Pawar, Madhavi Sardesai, Ramnath, Aadil Kak, Nazima, Richa, Mazhar Mehdi Hussain, Prashant Verma, and Swati Arora. Unified parts of speech (pos) standard in indian languages – draft standard – version 1.0, 2010. URL http://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf. [online; accessed 2018-08-29].

Geoffrey Leech and Andrew Wilson. Standards for tagsets. In Syntactic Wordclass Tagging. Text, Speech and Language Technology, pages 55–80. Kluwer Academic

Publishers, Dordrecht, The Netherlands, 1999. ISBN 0-7923-5896-1.

Markéta Lopatková, Martin Plátek, and Vladislav Kuboň. Modeling syntax of free word-order languages: Dependency analysis by reduction. In Proceedings of the International Conference on Text, Speech and Dialogue (TSD), pages 140–147, Berlin / Heidelberg, 2005. Springer. URL http://ufal.mff.cuni.cz/~lopatkova/literatura/05-TSD-RA.pdf.

Witold Mańczak. Ile jest rodzajów w języku polskim? Język Polski, 36:116–121, 1956.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn treebank. Computational Linguistics, 19(2): 313–330, 1993.

Jiří Maršík and Ondřej Bojar. Trtok: A fast and trainable tokenizer for natural languages. The Prague Bulletin of Mathematical Linguistics, 98:75–85, 2012. ISSN 0032-6585.

Nicolar Mazziotta. Coordination of verbal dependents in Old French: Coordination as a specified juxtaposition or apposition. In Proceedings of International Conference on Dependency Linguistics (DepLing 2011, 2011.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP), pages 62–72, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D11-1006.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. Universal dependency annotation for multilingual parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 2013. Bălgarska akademija na naukite, Association for Computational Linguistics. ISBN 978-1-937284-50-3.

Igor A. Mel'čuk. Dependency Syntax: Theory and Practice. State University of New York Press, 1988.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. Building the Italian syntactic-semantic treebank. In Anne Abeillé, editor, Building and using Parsed Corpora, Language and Speech series, pages 189–210, Dordrecht, 2003. Kluwer.

Jens Nilsson, Johan Hall, and Joakim Nivre. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. In Proceedings of the NODALIDA Special Session on Treebanks, 2005. URL http://www.msi.vxu.se/users/nivre/research/Talbanken05.html.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The CoNLL 2007 shared task on dependency parsing. In

Proceedings of the CoNLL 2007 Shared Task. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 915–932. Association for Computational Linguistics, 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1.pdf\#page=949.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal dependencies v1: A multilingual treebank collection. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pages 1659–1666, Paris, France, 2016. European Language Resources Association. ISBN 978-2-9517408-9-1.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Kamil Kopacewicz, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek,

Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jing Xian Wa. Universal dependencies 2.3, 2018. URL `http://hdl.handle.net/11234/1-2895`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Derek Nurse and Gérard Philippson. The Bantu Languages. Routledge Language Family Series. Routledge, Oxon/New York, 2003. ISBN 978-0-415-412-650.

OED. The Oxford English Dictionary. Oxford University Press, second edition, 1989. ISBN 978-0-19-861186-8.

Lluís Padró and Evgeny Stanilovsky. FreeLing 3.0: Towards wider multilinguality. May 2012.

Marco Passarotti and Felice Dell'Orletta. Improvements in parsing the index thomisticus treebank. revision, combination and a feature model for medieval latin. Training, 2:61–024, 2010.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, Proceed-

ings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 2089–2096, İstanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zdeněk Žabokrtský. Coordination structures in dependency treebanks. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 517–527, Sofija, Bulgaria, 2013. Association for Computational Linguistics. ISBN 978-1-937284-50-3.

Gregory Pringle. Thoughts on the universal dependencies proposal for japanese, 2016. URL `http://www.cjvlang.com/Spicks/udjapanese.html`. [online; posted 2016-09-18].

Prokopis Prokopidis, Elina Desipri, Maria Koutsombogera, Harris Papageorgiou, and Stelios Piperidis. Theoretical and practical issues in the construction of a Greek dependency treebank. In In Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT), pages 149–160, 2005.

Loganathan Ramasamy and Zdeněk Žabokrtský. Prague dependency style treebank for Tamil. In Proceedings of LREC 2012, pages 23–25, İstanbul, Turkey, 2012. European Language Resources Association. ISBN 978-2-9517408-7-7.

Mohammad Sadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, and Behrouz Minaei-Bidgoli. A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank. In 5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics, pages 227–231, Poznań, Poland, 2011.

Richard Hook Richens. Interlingual machine translation. The Computer Journal, 1 (3):144–147, 1958.

Rudolf Rosa. Multi-source cross-lingual delexicalized parser transfer: Prague or stanford? In Eva Hajičová and Joakim Nivre, editors, Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 281–290, Uppsala, Sweden, 2015. Uppsala University, Uppsala University. ISBN 978-91-637-8965-6.

Rudolf Rosa and Zdeněk Žabokrtský. $\mathrm{KL}_{\mathrm{cpos}^3}$ – a language similarity measure for delexicalized parser transfer. In ACL (2), pages 243–249, 2015.

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, and Zdeněk Žabokrtský. Hamledt 2.0: Thirty dependency treebanks stanfordized. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pages 2334–2341, Reykjavík, Iceland, 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Carl R. Galvez Rubino. Ilocano: Ilocano–English/English–Ilocano Dictionary and Phrasebook. Hippocrene Books, New York, USA, 1998.

Paul Schachter and Timothy Shopen. Part-of-speech systems. In Timothy Shopen, editor, Language Typology and Syntactic Description. Volume I: Clause Structure, pages 1–60. Cambridge University Press, Cambridge, UK, second edition, 2007.

ISBN 978-0-521-58156-1.

Roy Schwartz, Omri Abend, and Ari Rappoport. Learnability-based syntactic annotation design. In Proceedings of COLING, Mumbai, India, 2012.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clérgerie. Overview of the SPMRL 2013 shared task: Cross-framework evaluation of parsing morphologically rich languages. In Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages, pages 146–182, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-4917.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. The meaning of the sentence in its semantic and pragmatic aspects. Riedel / Academia, Dordrecht / Praha, 1986.

Natalia Silveira and Christopher Manning. Does UD need a parsing representation? an investigation of English. In Eva Hajičová and Joakim Nivre, editors, Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 281–290, Uppsala, Sweden, 2015. Uppsala University, Uppsala University. ISBN 978-91-637-8965-6.

Kiril Simov and Petya Osenova. Extending the annotation of BulTreeBank: Phase 2. In The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005), pages 173–184, Barcelona, Spain, December 2005.

Otakar Smrž, Viktor Bielický, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, and Petr Zemánek. Prague Arabic dependency treebank: A word on the million words. In Proceedings of the Workshop on Arabic and Local Languages (LREC 2008), pages 16–23, Marrakech, Morocco, 2008. European Language Resources Association. ISBN 2-9517408-4-0.

Jan Štěpánek. Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konzistence dat) [Capturing a Sentence Structure by a Dependency Relation in an Annotated Syntactical Corpus (Tools Guaranteeing Data Consistence)]. PhD thesis, Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czech Rep., 2006.

Milan Straka, Jan Hajič, and Jana Straková. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 2016. European Language Resources Association. ISBN 978-2-9517408-9-1.

STTS. Das Stuttgart-Tübingen Wortarten-Tagset – Stand und Perspektiven. Journal for Language Technology and Computational Linguistics, 28(1), 2013. ISSN 2190-6858. URL http://www.jlcl.org/2013_Heft1/H2013-1.pdf.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In Proceedings of CoNLL, 2008.

Roland Sussex and Paul Cubberley. The Slavic Languages. Cambridge Language Surveys. Cambridge University Press, 2011. ISBN 978-0-521-29448-5.

Logan Sutton. Noun class and number in kiowa-tanoan: Comparative-historical research and respecting speakers' rights in fieldwork. Fieldwork and Linguistic Analysis in Indigenous Languages of the Americas, pages 57–89, May 2010. URL http://hdl.handle.net/10125/4451.

John Sylak-Glassman. The composition and use of the universal morphological feature schema (UniMorph schema), working draft, v. 2, June 2016. URL https://unimorph.github.io/doc/unimorph-schema.pdf. [online; accessed 2018-08-23].

Mária Šimková and Radovan Garabík. Sintaksičeskaja razmetka v slovackom nacional'nom korpuse (Синтаксическая разметка в Словацком национальном корпусе). In Trudy meždunarodnoj konferencii Korpusnaja lingvistika (Труды международной конференции Корпусная лингвистика) – 2006, pages 389–394, Sankt-Peterburg, Russia, 2006. St. Petersburg University Press. ISBN 5-288-04181-4.

Marko Tadić. Building the Croatian dependency treebank: the initial stages. Suvremena Lingvistika, 63(1):85–92, May 2007. URL https://www.researchgate.net/publication/228614382_Building_the_Croatian_Dependency_Treebank_the_initial_stages.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. Universal dependencies for japanese. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1651–1658, Paris, France, May 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1. URL http://www.lrec-conf.org/proceedings/lrec2016/summaries/122.html.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. AnCora: Multilevel annotated corpora for Catalan and Spanish. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May – 1 June 2008, Marrakech, Morocco. European Language Resources Association, 2008. URL http://www.lrec-conf.org/proceedings/lrec2008/summaries/35.html.

Ulf Teleman. Manual för grammatisk beskrivning av talad och skriven svenska (Mamba), 1974.

Lucien Tesnière. Eléments de syntaxe structurale. Klincksieck, Paris, France, 1959.

Jörg Tiedemann. Rediscovering annotation projection for cross-lingual parser induction. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1854–1864, August 2014.

Marie Těšitelová. K jazyku věcného stylu z hlediska kvantitativního (on the language of non-fiction style from the quantitative point of view). Slovo a slovesnost, 44(4): 275–283, 1983. URL http://sas.ujc.cas.cz/archiv.php?art=2911.

Stephen Tratz and Eduard Hovy. A fast, accurate, non-projective, semantically-enriched parser. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1257–1268, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D11-1116.

UNICODE. Unicode normalization forms. Unicode® standard annex #15, 2018. URL http://unicode.org/reports/tr15/.

Leonoor van der Beek, Gosse Bouma, Jan Daciuk, Tanja Gaustad, Robert Malouf, Gertjan van Noord, Robbert Prins, and Begoña Villada. Chapter 5. the Alpino dependency treebank. In Algorithms for Linguistic Processing NWO PIONIER Progress Report, Groningen, The Netherlands, 2002. URL http://odur.let.rug.nl/~vannoord/trees/Papers/report_ch5.pdf.

Norbert Volz and Suzanne Lenz. Multilingual corpus tagset specifications. MLAP PAROLE 63–386 WP 4.1.4, 1996. URL http://www.elda.org/catalogue/en/text/doc/parole.html.

Laurel J. Watkins. A grammar of Kiowa. University of Nebraska Press, Lincoln, USA, 1984.

James R. Wenger. Some Universals of Honorific Language with Special Reference to Japanese (Ph.D. thesis). University of Arizona, Tucson, AZ, USA, 1982.

Arok Elessar Wolvengrey. Semantic and pragmatic functions in Plains Cree syntax (PhD thesis). LOT, Utrecht, Netherlands, 2011. ISBN 978-94-6093-051-5.

Alina Wróblewska and Marcin Woliński. Preliminary experiments in Polish dependency parsing. In Pascal Bouvry, MieczysławA. Kłopotek, Franck Leprévost, Małgorzata Marciniak, Agnieszka Mykowiecka, and Henryk Rybiński, editors, Security and Intelligent Information Systems, volume 7053 of Lecture Notes in Computer Science, pages 279–292. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-25260-0. doi: 10.1007/978-3-642-25261-7_22. URL http://dx.doi.org/10.1007/978-3-642-25261-7_22.

Fei Xia. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0), October 2000.

Daniel Zeman. Reusable tagset conversion using tagset drivers. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, Proceedings of the Sixth International Language Resources and Evaluation Conference, LREC 2008, pages 28–30, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. URL http://www.lrec-conf.org/proceedings/lrec2008/summaries/66.html.

Daniel Zeman. Slavic languages in Universal Dependencies. In Natural Language Processing, Corpus Linguistics, E-learning, pages 151–163, Lüdenscheid, Germany, 2015. RAM-Verlag. ISBN 978-3-942303-32-3. URL http://ufal.mff.cuni.cz/

biblio/servlet/File?field=File&id=4326707699154676324.

Daniel Zeman. Core arguments in Universal Dependencies. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), Linköping Electronic Conference Proceedings, pages 287–296, Pisa, Italy, 2017. Linköping University Electronic Press. ISBN 978-91-7685-467-9.

Daniel Zeman and Philip Resnik. Cross-language parser adaptation between related languages. In Workshop on NLP for Less-Privileged Languages, IJCNLP, Hyderabad, India, 2008.

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. HamleDT: To parse or not to parse? In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 2735–2741, İstanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. Hamledt: Harmonized multilanguage dependency treebank. Language Resources and Evaluation, 48(4):601–637, 2014. ISSN 1574-020X.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadova, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics, 2017.

Fernando Zúñiga and Beatriz Fernández. Grammatical relations in Basque (draft). 2014. URL http://basdisyn.net/pdf/Zuniga\%20&\%20Fernandez\%202014\%20Basque\%20GRs\%20270614.pdf.

Language Index