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.

Editor-in-chief: Jan Hajič

Editorial board: Nicoletta Calzolari, Mirjam Fried, Eva Hajičová,
Petr Karlík, Joakim Nivre, Jarmila Panevová,
Patrice Pognan, Pavel Straňák, and Hans Uszkoreit

Reviewers: Ing. Alexandr Rosen, Ph.D.
Mgr. Barbora Vidová Hladká, Ph.D.

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 2

Tokenization and Segmentation

The basic unit in a treebank is the **token.** For the most part, tokens correspond to words and we use the two terms interchangeably unless explicitly said otherwise. However, a token also covers elements that do not fall under the purely linguistic notion of word: punctuation symbols, numbers, e-mail addresses, URLs etc. A majority of words (tokens) in most writing systems[1] are delimited by spaces on each side. There are numerous exceptions to this rule of thumb. Typographic guidelines often dictate that a punctuation mark be attached to the neighboring word without intervening space; but it is practical in NLP to separate the two as independent tokens. On the other hand, we probably do not want to split decimal numbers *(3.14)* or e-mail addresses *(me@universe.org)* into multiple tokens. And occasionally we may want to join two or more adjacent token candidates into one token. For instance, thousands in large numbers are separated by spaces in some languages *(1 000 000)* but we want to keep the whole number as one token. The process of identifying and marking tokens in running text is called **tokenization.** Formally, it is a function that takes a sequence of characters as input and returns a sequence of tokens.

2.1 Methods of Tokenization

Tokenization is almost always an automatic process (as opposed to higher-level corpus annotation, which must be done manually by human annotators if we want to avoid errors). One of the simplest approaches to tokenization is based on the Unicode[2] tables of character properties. Every character belongs to a class such as letters, digits or punctuation symbols. Any whitespace character (space, tabulator, line break) sep-

[1] In this work we focus on written corpora. To some extent, our observations can be extended to transcribed spoken data too; however, we largely ignore the issues of transcription, such as missing graphical representation of word and sentence boundaries.

[2] `http://www.unicode.org/`

arates tokens; in addition, there is a token boundary between any two non-whitespace characters that belong to different classes.

Such a low-level approach has downsides, too. Decimal point is not a digit, therefore it splits the number into three tokens: *3.14* is tokenized as "3 . 14". Long numbers in languages like French or Czech will crumble because of the spaces they contain: *1 000 000*. E-mail addresses and other atomic strings will not be kept together. Tokenizers thus often include rules that specify situations where token candidates should be merged. And then there are also tokenizers that can automatically learn the rules from a set of seed examples (Maršík and Bojar, 2012) or simply from tokenized text (Straka et al., 2016).

Tokenization rules may be regular expressions describing typographical conventions applicable to many languages, but they may also refer to concrete lexical material in a particular language. For example, many corpora choose to treat the abbreviation-marking period as a part of the abbreviated token: *dr.* ("doctor") is one token that includes two letters and the period. In other corpora however (Hajič et al., 2000), the period is not part of the abbreviated token. There is evidence that the period actually should be kept separate. Sometimes it performs a double function—besides marking the abbreviation, it also terminates a sentence:

(1) English: *We have seen rising prices of dairy products, bread, soft drinks etc.*

Here, *etc.* is an abbreviation of Latin *et cetera* and the period marks the abbreviation. But the period also happens to occur at the end of the sentence, which should also be marked by a period. Instead of typing two periods in sequence, the convention is to merge them into one, double-duty symbol.

On a similar note, many languages use the hyphen (optionally or compulsorily) in certain kinds of compounds, e.g. *Anglo-American* in English or *česko-německý* "Czech-German" in Czech. A tokenizer based on character categories will split such a compound into three tokens: "Anglo - American". Indeed, this is what some corpora do (Hajič et al., 2000). The advantages of such an approach are simplicity of tokenization and reduced data sparsity for machine learning (storing individual parts of compounds in a dictionary is easier and we are more likely to encounter a large proportion of possible parts than of all their combinations). On the other hand, the initial part of a two-part compound sometimes differs in form from independent words (*Anglo* is not a self-contained English adjective). That is why other corpora (Brants et al., 2002) keep such compounds as one token, similarly to what is usually done with compounds that are written without a hyphen (German *angloamerikanisch* "Anglo-American").

There is no single and universally valid set of tokenization rules. The same can be said about the inverse procedure, detokenization, although it may not be immediately obvious. Rules for space insertion vary (for instance, quotation marks should be adjacent to the first/last word of the quoted text in Czech and English, but not in French). It is not always possible to distinguish opening from closing symbols, or hyphens from dashes. And it is not guaranteed that the original text strictly follows

the rules. For all these reasons it is desirable that the tokenized corpus preserves the information about spaces before tokenization. Being able to restore the underlying text means that we can train a tokenizer as a statistical model, rather than a set of human-designed rules. Unfortunately, only a fraction of treebanks preserves this sort of information.

2.2 Normalization of Forms

It is customary to normalize certain types of tokens in corpora, for various reasons. Sometimes there is more than one way to write (or digitally encode) a token, without any *linguistic* differences. Even if national typographic rules require one particular way of presentation, the influence of English and of the internet on one side, and technical limitations on the other side cause people to deviate from the standard. Thus we can encounter the typographically correct and directed quotation marks (like “here” – using `U+201C`[3] for the opening mark and `U+201D` for the closing mark), and in other texts there will be the unidirectional ASCII mark `"`, corresponding to `U+0022`. Some people will use the “English” quotes in languages where they are not appropriate (for example, the „Czech“ rule requires a low-9 mark, `U+201E`, as the opening symbol, and a high-6 mark, identical to the English opening mark, as the closing symbol). There are several other options of encoding the quotation marks in digital text, including the sequences used in the LaTeX typesetting system: two grave accents (``) instead of the opening mark, and two ASCII apostrophes ('') instead of the closing mark. Authors of corpora often choose to normalize the quotation marks to one particular type, so that the marks can be easily identified throughout the corpus. It is desirable that information is not lost in the process of normalization. Normalization of all quotes to the unidirectional ASCII symbol would be a poor decision because it would obliterate the difference between opening and closing marks. Ideally, in any kind of normalization, the original encoding should be preserved in the corpus, perhaps as a special attribute of the token, so that it can be accessed and restored if necessary.

Some other normalization examples include the following choices:

- Various kinds of hyphens, short and long dashes. The ellipsis character (…, `U+2026`) vs. three consecutive full stop characters (`U+002E`).
- Decimal numbers may use the decimal point, as in English *(3.14159)* or the decimal comma, as in German *(3,14159)*.
- Order groups in large numbers may be separated by commas as in English *(1,000,000)*, by spaces as in French *(1 000 000)*, or not separated at all *(1000000)*.[4]

[3] `U+xxxx` refers to a Unicode character using its hexadecimal address. `U+201C` is the character at the code point $201C_{hex}$, or 8220_{dec}.

[4] Even the grouping by thousands cannot be taken for granted: in Indian English, one does not use millions and billions, but lakh (hundred thousand) and crore (ten million), with different position for the comma separator: western 7,000,000,000 would be written 7,00,00,00,000 in India. However, one could argue that this is a difference between Indian and non-Indian English, which should be preserved in the corpus.

- Languages with non-Latin writing systems, such as Arabic or Hindi, have their own sets of digits, but the western "Arabic" digits are often used as well. Thus in a Hindi text written in the Devanagari script, we can encounter both the western digits *0, 1, 2, 3, 4, 5, 6, 7, 8, 9* (U+0030 to U+0039) and the Devanagari digits ०, १, २, ३, ४, ५, ६, ७, ८, ९ (U+0966 to U+096F).
- Besides the standard space character (U+0020), the text may contain a "no-break space" (U+00A0), signalling that a word processor should not break the line on this space. Such a space may or may not indicate a token boundary. Furthermore, Unicode provides several other space characters of varying sizes and behavior (U+2002 to U+200B).
- The character stream may contain various "invisible" characters such as the zero width non-joiner (ZWNJ, U+200C). It is used to prevent ligature of neighboring letters, which would be otherwise automatically created by a typesetting system. For example, the Persian plural suffix ـها *(-hā)* is always attached to the word stem via a ZWNJ. The Unicode category of ZWNJ is not "letter" but "other". Therefore, a naive category-based tokenizer would insert a token boundary before a Persian plural suffix.
- For similar reasons, Catalan uses a "middle dot" (U+00B7) between two L's to prevent pronounciation normally associated with *ll*: *col·lecció* "collection".
- Unicode often provides more than one way of encoding the same string. A set of "combining diacritics" is available so that any letter can be combined with any diacritical mark; however, more common accented letters are available as independent characters, too. Hence the Latin small letter *a* with grave accent, *à*, occupies the codepoint U+00E0 but it can be also encoded as two characters, the letter *a* (U+0061) and combining grave accent (U+0300). In contrast, the grave accent can also be combined with a preceding schwa (U+0259 U+0300: *ə̀*) but there is no single codepoint for schwa with grave accent. The possibilities are even more complex when two diacritical marks attach to one base letter. For example, the Latin small letter *u* with diaeresis and caron *(ǚ)* can be encoded as a single character (U+01DA) but also as various combinations of the letters *u* (U+0075), *ü* (U+00FC) and even *ǔ* (U+01D4)[5], and the combining marks diaeresis (U+0308) and caron (U+030C). The Unicode consortium provides recommendations how to normalize text with respect to this kind of variation (UNICODE, 2018).

2.3 Multi-Word Expressions

Various types of fixed expressions, compounds and named entities may be understood as single syntactic units (tree nodes), although orthographically they are com-

[5] But note that changing the order of the two diacritics may result in different appearance of the final glyph.

posed of multiple words. Indeed, some treebanks collapse multi-word expressions (MWEs) into single tokens (often using the underscore, "_", as the glue character). The exact demarcation of MWEs varies among the treebanks that take this approach.

For example, in the Alpino treebank of Dutch (van der Beek et al., 2002), personal names are treated as single tokens *(A._G._van_der_Spek)*. Similar treatment is given to names of organizations, which sometimes leads to very long tokens (*Departement_voor_sabotage_van_macht_en_geweld_van_de_Oranje_Vrijstaat* "Department for Sabotage of Power and Violence of the Orange Free State"), and to various other named entities (*zondag_1_april* "Sunday April 1"). Besides named entities, many fixed expressions are collapsed as well. Examples include multi-word prepositions (*op_basis_van* "on the basis of") and adverb-like expressions (*bij_voorbeeld* "for example").

In the SynTagRus treebank of Russian (Boguslavsky et al., 2000, 2013), MWEs are also collapsed but no underscore character is used to glue the tokens together; instead, tokens are allowed to contain spaces, and XML markup is used to mark token boundaries. Examples: *несмотря на (nesmotrja na)* "despite", *до сих пор (do sih por)* "so far".

In the Persian Dependency Treebank (Rasooli et al., 2011), analytical verb forms are treated as multi-word expressions and collapsed into one token. For instance, خواهند کرد *(xwāhand kard)* "they will do" is one such token.

In contrast, Universal Dependencies only allows tokens with spaces in a narrow set of cases, such as the long numbers where spaces help to make the number more easily readable. Words in fixed multi-word expressions are technically independent nodes, but they are connected using special "non-syntactic" relations that signal the MWE (Figure 2.1). Note that this approach allows to assign to each member word its lemma, part-of-speech tag and morphological features. If words are merged to a single node, one can either annotate the node with a sequence of morphological attributes (as in the Alpino treebank; but this will complicate processing of the data), or try to come up with a "morphological" annotation of the whole MWE (it does not make much sense with lemma and inflectional features, but it may be actually useful with part-of-speech tags, which are determined by both morphology and syntax).

Finally, there are treebanks where nodes in trees correspond to chunks, i.e., sequences of words rather than individual words. Examples include the treebanks of Hindi, Bengali and Telugu used in the ICON 2010 NLP Tools Contest (Husain et al., 2010). One chunk node may comprise a whole verbal group with all auxiliaries, or a noun with its postposition(s). Chunks are not considered equivalent to tokens, but at least on the syntactic level they are treated as if they actually were equivalent.

2.4 Word Segmentation

While the term *multi-word expression* refers to a sequence of surface (orthographic) words that act as one syntactic unit, there is also the opposite phenomenon: one sur-

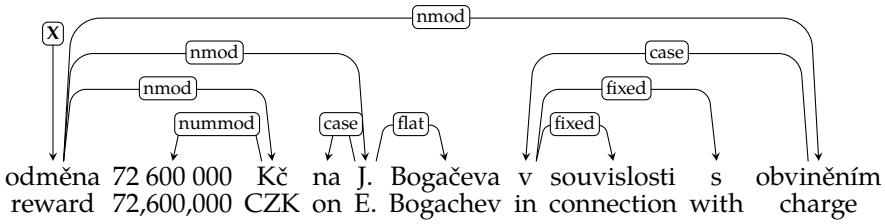


Figure 2.1: A Czech sentence fragment that illustrates the UD approach to multi-word expressions: the relations `flat` and `fixed` are special and connect tokens that make up one syntactic unit, without asserting that one node dominates the others. There is also a token with spaces, the number *72 600 000*.

face (orthographic) word is composed of multiple syntactic units. UD terms such words *multi-word tokens (MWT);* elsewhere, they may be called *contractions, fused words* etc. The terminology is slightly confusing here. One has to recognize that "multi-word" in "multi-word expression" refers to multiple orthographic words, whereas "multi-word" in "multi-word token" refers to multiple *syntactic* words.

An Arabic example is the contraction وبالفالوجة *(wabiālfālūjah)* = *wa*/`CONJ` + *bi*/`PREP` + *alfālūjah*/ `NOUN_PROP` = "and in al-Falujah". Several European languages have fused prepositions with articles or pronouns. For instance, German *zum* is a contraction of the preposition *zu* "to" and the dative definite article *dem* "the". Another example is the English *don't*: naive tokenization would split it on the apostrophe, but it is actually a contraction of *do not*, hence the three characters *n't* should stay in the same token.

There are several possibilities how to deal with contractions in annotation. In Arabic, word segmentation (also called tokenization) is traditionally done together with morphological analysis, and the resulting morphological hypotheses may also differ in word boundaries. Tokenization is no longer simple and language-independent; but in return, we do not have to struggle with lemma and part-of-speech tag for a syntactically heterogenous string. In German, the class of fused words is much more restricted and closed. The Stuttgart-Tübingen Tagset (STTS, 2013), which is one of the most popular POS tagsets for German, has a special tag for preposition-article contractions: `APPRART`. However, such a solution cannot be used in a language-agnostic annotation scheme such as Universal Dependencies. Therefore, UD has two levels of segmentation: the lower level corresponds to simple tokenization, i.e., separation of punctuation symbols and real words. The higher level, word segmentation, splits multi-word tokens into syntactic words, which are then treated as independent tree nodes, with their own lemma, POS tag, morphological features and dependency relations. The UD guidelines make sure that the original surface word form and its

mapping to syntactic words is preserved, so that NLP tools can be trained to segment new text the same way.

Finally, there are writing systems where spaces do not delimit words. Either spaces are not used, as in Chinese, or they are used to delimit sub-word units, as in Vietnamese. The visible unit in both languages is a monosyllabic morpheme. It may or may not correspond to a word.[6] Even though words are not visible in the text, it is advantageous to define them for the purpose of natural language processing and linguistics. Thus the surface text in (2) should be segmented as in (3).

(2) 現在我們在瓦倫西亞。
Xiàn zài wǒ men zài wǎ lún xī yǎ.

(3) 現在 我們 在 瓦倫西亞 。
Xiànzài wǒmen zài wǎlúnxīyǎ .
Now we in Valencia .

"We are now in Valencia."

A sophisticated, language-specific algorithm is needed to segment words in Chinese text. One also needs a standard that defines what is a word in Chinese. Should 北海 *(Běihǎi)* "North Sea" be one word as in German *(Nordsee)* and Dutch *(Noordzee)*, or two words as in English *(North Sea)* and Czech *(Severní moře)*? Should 火车站 *(huǒchēzhàn)* (lit. *fire vehicle stand*) "railroad station" be one word, two, or three? In mainland China, there is a national standard that sets the word segmentation rules (GB/T, 1993); other standards are have been proposed for other languages (Choi et al., 2009). Despite it, controversies sometimes arise, as in the case of the Universal Dependencies for Japanese (Tanaka et al., 2016; Pringle, 2016). So there are multiple possible segmentations of the example (4), varying in whether and which morphemes are considered inflectional morphology, or separate auxiliary verbs, postpositions and other function words (Figure 2.2).

(4) 経堂の美容室に行ってきました。
Kyō dō no mi yō shitsu ni i t te ki ma shi ta.

"I went to the beauty salon of Kyōdō."

Vietnamese uses a writing system based on the Latin alphabet. A space is normally inserted between every two syllables, whether they belong to the same word or not. Vietnamese is thus not unlike Chinese, where every character corresponds to a syllable. Polysyllabic loanwords from foreign languages are sometimes exempt from this rule and written together, as *bêtông* "concrete" (from French *béton*) in (5); but even

[6] With the exception of polysyllabic loanwords in Vietnamese.

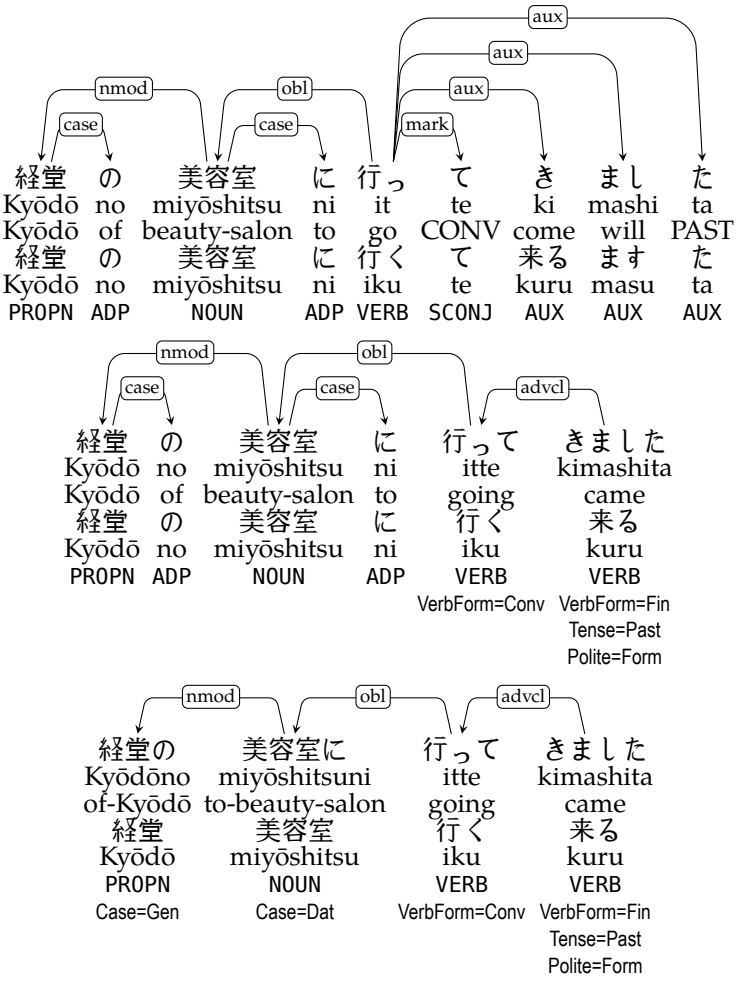
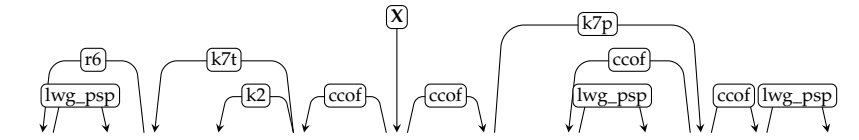


Figure 2.2: Three segmentation options for the Japanese sentence in (4), with English glosses, lemmas, part-of-speech tags and morphological features. The first option is the one actually used in Universal Dependencies 2.0.



"they gamble on the Diwali festival but [they do so] at home or hotel"

Figure 2.3: A sentence from the Hindi Dependency Treebank (Husain et al., 2010) that demonstrates the usage of a NULL (empty) node for a deleted conjoined predicate.

here, it is possible to write the syllables separately *(bê tông)*. Besides *bêtông*, there are three bisyllabic words in the example: *Tất cả, nội đồng* and *thành quả*.

(5) Tất cả đường bêtông nội đồng là thành quả …
All road concrete country is achievement …

"All the concrete country roads are the result of…"

2.5 Empty Nodes

Occasionally the syntactic annotation of a sentence contains empty nodes (variously called also NULL nodes, Phantom nodes etc.) that do not correspond to any surface word. This is done only in some treebanks, usually those that strive to get closer to meaning. Empty nodes signal ellipsis and represent participants that have been deleted from the utterance because they are known from the context. Empty nodes are especially useful if the missing word has orphaned dependents in the sentence; however, in some treebanks they are also employed for dropped subjects (pro-drop languages). A prototypical type of ellipsis where empty nodes are useful is *gapping*: coordinate clauses with a shared predicate but different set of arguments (Figure 2.3).

The AnCora treebanks of Catalan and Spanish (Taulé et al., 2008) are examples of pro-drop languages where empty nodes are used to represent subjects. Another example would be the tectogrammatical layer of the Prague family of treebanks (Hajič et al., 2000). Here is a Spanish example:

(6) _ Afirmó que _ sigue el criterio europeo y que _ trata de
He said that he follows the standard European and that he works on
incentivar el mercado donde no lo hay.
encouraging the market where not it exists.

> "He said he follows the European standard and encourages the market where there is none."

The three underscores all mark subjects of the following verbs and could be translated as "he".

2.6 Sentence Segmentation

Sentence is the next level of segmentation, the higher unit immediately above word. In treebanks, one tree corresponds to one sentence. In most types of text and in most writing systems it is not difficult to define sentence boundaries, at least for a human annotator. Some sort of punctuation symbols is usually needed to mark sentence end in writing; corpora that lack punctuation thus have to use other criteria for sentencehood. Transcribed spoken language often resists attempts to cut it in sentences, especially if it contains dialogues. Old texts in classical languages often lack punctuation, too.

One possible approach is to use clauses as segments, instead of sentences: clauses can be determined syntactically (Călăcean, 2008). The drawback is that syntactic relations between clauses will not be captured. The other extreme is to take entire paragraphs as segments; the texts of the Prague Arabic Dependency Treebank (PADT) (Smrž et al., 2008) are close to this extreme, although some paragraphs are segmented into several sentence-like units. Here the obvious drawback is processing complexity, both by humans and by statistical models.

Finally, some special types of text have their own units that work as sentence-level segments. For instance, some corpora contain poetry or excerpts from the Bible, and they are segmented by verses.

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 non-finite 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 4

Morphological Features

We showed in Section 3.1 that many tagsets are composed of the main part-of-speech category, plus a number of additional feature-value pairs (or "dimensions of meaning" in the terminology of (Sylak-Glassman, 2016)). We also discussed several features that often appear in morphological tags but can be understood as a more fine-grained partition of the part-of-speech space. These features are lexical, that is, they have the same value for all forms of one lexical unit (lemma). Here we only briefly recapitulate some more common representants of this type of features:

Pronominal (deictic) type. As discussed in the previous section, this feature applies to multiple top-level parts of speech, such as (substantive) pronouns, (attributive) determiners, pronominal quantifiers and adverbs. **Personal** *(I, you, they)* and **possessive** *(my, your, their)* pronouns (determiners) are distinguished in many languages, although possessives are in a sense personal as well. **Reciprocal** pronouns (German *einander* "each other") are used as objects of reciprocal transitive predicates. **Reflexive** pronouns (English *myself, yourself, themselves*; German *sich*) are used to co-index an object or adjunct with the subject of the clause; as a matter of fact, the reflexive category is orthogonal to the other pronominal types, as there are also **reflexive possessives,** expressing that something is possessed by the subject (Czech *svůj*). In some languages, the reflexives have multiple functions, they can also act as reciprocal pronouns or mark the passive/middle voice. In English, the reflexive pronouns can be used as **emphatic** words *(John **himself** did this)* while other languages have a separate type of pronoun for emphasis (Romanian *însuși*, Czech *sám*). There are **interrogative** and **relative** words; these two classes overlap in English but they are disjoint in other languages *(who, what, which, where, when)*. Strictly speaking, these are again orthogonal because there are also **interrogative/relative possessives** *(whose)*. Other classes include **demonstratives** *(this, that, here, there, now, then)*, **indefinites** *(one, something,*

anything, somewhere, sometimes), **universals** *(everybody, everything, every, everywhere, always)* and **negatives** *(nobody, nothing, no, nowhere, never)*.

Numeral type. This feature may apply to multiple top-level parts of speech if the various numeric words are first classified by their morphosyntactic behavior: nouns, adjectives, adverbs, and, of course, numerals. If the language has deictic quantifiers (such as Czech *kolik* "how many"), then the word will have both the numeral type and a pronominal type. Numeral types include **cardinals** *(one, two, three)*, **ordinals** *(first, second, third)* or **multiplicatives** *(once, twice)*.

Adverbial type. A mostly semantic division to adverbs of **location, time, manner, cause** etc. Applies to both descriptive and pronominal adverbs: for example, *how, so* and *hastily* are all manner adverbs. The BulTreeBank tagset of Bulgarian uses a more general feature called **referent type** and applies it to all deictic words. For deictic adverbs, the referent types are the adverbial types mentioned above; for a deictic quantifier, the referent type is 'quantity'; for a pronoun, it is 'entity'.

Verb form. We list this feature here because it has been already discussed in more detail among the parts of speech; however, it is a lexical feature only if the various nonfinite forms are treated as derived words, i.e., either separate parts of speech or special types of a non-verb part of speech. The verb form usually has a big impact on the applicability of the other morphological features. This can be also said about mood (see Section 4.18), and mood only applies to finite verbs, hence some tagsets conflate verb form and mood into one feature. Verb forms include **finite verbs** (Spanish *conozco, conoces, conoce* "I know, you know, he knows"), **infinitives** (*conocer* "to know"), **participles** (*conocido* "known"), **converbs** or adverbial participles (Czech *vědouce* "knowing"), **verbal nouns** (*vědění* "knowing"), **gerundives** and **supines** (both sometimes considered special types of other verb forms).

We will now turn to features that relate to morphological inflection. Note that some of them are actually sometimes lexical, depending on the part of speech of the word. For instance, the feature of gender (Section 4.1) is a lexical feature of nouns, but an inflectional feature of adjectives or verbs.

4.1 Gender

Gender is one of the features that categorize nouns (i.e., for nouns this feature is lexical) but other parts of speech inflect for gender and show agreement with nouns, in order to signal some sort of syntactic relation. For example, Czech adjectives agree in gender with the nouns they modify. The same holds for subject-verb agreement, but only if the verb has a participial form, e.g. the past tense. Finite verbs do not inflect for gender in Czech, but they do so e.g. in Arabic.

Gender inflection has vanished from English, yet the grammatical category still exists in English and is demonstrated in the third-person pronouns. All English nouns can thus be classified by one of three gender values, depending on whether they would be referred to using *he, she* or *it*. Nevertheless, the gender feature is usually not encoded in tags of English nouns (it is sometimes encoded in tags of pronouns).

There are various gender-like systems with various numbers of genders. Some language families lack the feature completely. The Indo-European system is originally based on three values: **masculine, feminine** and **neuter.** In some languages this system has been simplified to two values: either masculine-feminine, or **common**-neuter (also called utrum-neutrum). Some other values will be discussed later as separate features of animacy (Section 4.2) and noun class (Section 4.3).

When there is correlation between gender and biological sex, it often works one way only. Nouns denoting male humans are likely to belong to the masculine gender, and female humans usually belong to the feminine gender. However, the English approach of putting everything else in the neuter class is not typical: many languages categorize nouns that denote things as either masculine or feminine, and the choice is rather arbitrary. Hence the noun *castle* is neuter in English, but its German counterpart *Burg* is feminine and the Czech for "castle", *hrad*, is masculine.

In the masculine-feminine systems (as seen in most Romance languages), all thing-denoting nouns have been distributed into one of the originally sex-based classes. In the common-neuter systems (seen in Scandinavian languages and Dutch, for example), the borderline is also arbitrary but nouns denoting human beings will likely have the common gender, as it is a merger of ancient masculine and feminine.

Czech is a language where determiners, adjectives, certain numerals and verb forms agree with the modified (or cross-referenced) noun in gender—note that the form of the first, second and fourth word changes while the corresponding words in the English translation stay the same:

(36) *Tento veselý muž vyhrál závod.* "This cheerful man won the race." (masculine)

(37) *Tato veselá žena vyhrála závod.* "This cheerful woman won the race." (feminine)

(38) *Toto veselé dítě vyhrálo závod.* "This cheerful child won the race." (neuter)

Determiners and adjectives agree with nouns also in Spanish (adjectives follow the noun in this case):

(39) *un hombre viejo* "an old man" (masculine)

(40) *una mujer vieja* "an old woman" (feminine)

However, Spanish also has a considerable group of adjectives that do not inflect and can be used both with masculine and feminine nouns. These adjectives are not in the common gender, at least not in the same sense as this gender is defined in the Scandinavian languages (albeit some tagsets will use the common label for them). It

would be more appropriate to say that the adjective is either masculine or feminine (depending on context) but its form is ambiguous.

(41) *un hombre feliz* "a happy man" (masculine)

(42) *una mujer feliz* "a happy woman" (feminine)

The examples (43) and (44) demonstrate how Arabic finite verbs are sensitive to the gender of the subject:

(43) رجل يقرأ كتابا *(raǧulun yaqraʾu kitāban)* "A man reads a book."

(44) امرأة تقرأ كتابا *(imraʾatun taqraʾu kitāban)* "A woman reads a book."

4.2 Animacy

Similarly to gender (Section 4.1) and to African noun classes (Section 4.3), animacy is usually a lexical feature of nouns and an inflectional feature of other parts of speech that mark agreement with nouns. Some languages (e.g., the Romance group) distinguish gender but not animacy. Other languages (e.g., Basque) distinguish animacy but not gender. Yet others distinguish both. The main distinction in Slavic languages is gender, but a few forms are further differentiated by animacy. In some languages, the distinction is **animate-inanimate,** in others, **human-nonhuman** (e.g. Yuwan, a Ryukyuan language). There are also three-way systems like human-animate nonhuman-inanimate.

Czech is an example of an animate-inanimate system. This distinction is much more semantic than the rather arbitrary genders: animate nouns denote people and animals, inanimate nouns denote things. By extension, mythical beings, puppets and other human- or animal-like objects are also treated as animate. However, the distinction matters only if the gender is masculine, and only in a few forms. In singular, the accusative form of animate masculines is identical to the genitive (45), while the accusative form of inanimate masculines is identical to the nominative (46). In plural, the nominative and vocative form of masculine inanimates (48) is identical to feminine forms (49), while masculine animates (47) have a different form.

(45) *Vidím nějakého starého muže.* "I see an old man."

(46) *Vidím nějaký starý stůl.* "I see an old table."

(47) *Tito staří muži vyhráli závod.* "These old men won the race." (masculine animate)

(48) *Tyto staré vozy vyhrály závod.* "These old cars won the race." (masculine inanimate)

(49) *Tyto staré ženy vyhrály závod.* "These old women won the race." (feminine)

The interconnectedness of gender and number in Czech led some authors to merge the two features into one: a set of four genders (masculine animate, masculine inanimate, feminine and neuter). Indeed, this is how the PDT tagset models the situation. In contrast, the MULTEXT-EAST tagsets or the features in Universal Dependencies encode the two features independently.

In other Slavic languages, e.g., Polish, the animacy borderline is different in singular, where it leads between animates and inanimates, and in plural, where it leads between humans and nonhumans. Therefore, the interrogative determiner *który* "which" in singular accusative has the form *którego* if referring to a masculine animate noun, and *który* when referring to a masculine inanimate. In plural nominative, the form is *którzy* when referring to masculine humans, but *które* if referring to nonhumans. Finally, plural accusative is *których* for masculine humans and *które* for nonhumans. The other number-case combinations do not change. Like in Czech, the animacy distinctions are limited to masculine forms, therefore some tagsets have labels for three masculine genders (M1, M2, M3) plus two other genders (feminine and neuter).[1]

4.3 Noun Class

In a sense, noun classes are a generalization of gender (Section 4.1) and animacy (Section 4.2), although they may overlap with other features as well. To a large part, noun class is a lexical feature of nouns, while other parts of speech inflect for it to show agreement.

The distinction between gender and noun class is not sharp and is partially conditioned by the traditional terminology of a given language family. In general, the feature is called gender if the number of possible values is relatively low (typically 2–4) and the partition at least partially correlates with sex of people and animals. In language families where the number of categories is high (10–20), the feature is usually called noun class. Noun classes occur in various parts of the world but perhaps the most widely known are the noun classes of the Bantu languages.

In Bantu languages, the noun class also encodes grammatical number (Section 4.4), and some classes encode locative meanings that are similar to cases in other languages (Section 4.5); therefore it is a lexical-inflectional feature of nouns.

The set of values of this feature is specific for a language family or group. Within the group, it is possible to identify classes that have similar meaning across languages (although some classes may have merged or disappeared in some languages in the group). For instance, there is a standardized class numbering system accepted by

[1] As a matter of fact, describing the Polish masculine genders in terms of animacy is an approximation that is not entirely adequate. The "animate" (M2) category includes *trup* "corpse" or names of dances such as *walc* "waltz" and *fokstrot* "foxtrot". The word *babsztyl* "hag" is grouped together with masculine nonhuman animates, although semantically it is human and feminine. The borderline is arbitrary at places, just like the borders between genders. The three masculine genders have been introduced by (Mańczak, 1956) and are widely accepted in modern Polish linguistics.

No	Prefix	Typical meaning	Pair
1	*m-, mw-, mu-*	singular: persons	
2	*wa-, w-*	plural: persons	1
3	*m-, mw-, mu-*	singular: plants	
4	*mi-, my-*	plural: plants	3
5	*ji-, j-*	singular: fruits	
6	*ma-, m-*	plural: fruits	5, 9, 11, 1
7	*ki-, ch-*	singular: things	
8	*vi-, vy-*	plural: things	7
9	*n-, ny-, m-, Ø-*	singular: animals, things	
10	*n-, ny-, m-, Ø-*	plural: animals, things	9, 11
11	*u-, w-, uw-*	singular: no clear semantics	
15	*ku-, kw-*	verbal nouns	
16	*pa-*	locative meanings: close to something	
17	*ku-*	indefinite locative or directive meaning	
18	*mu-, m-*	locative meanings: inside something	

Table 4.1: Noun classes in Swahili. Class numbers correspond to all-Bantu classification; some Bantu classes are not present in Swahili. The **Pair** column refers from plural classes to corresponding singular classes.

scholars of the various Bantu languages; similar numbering systems should be created for the other families that have noun classes.

Table 4.1 shows the 15 classes used in Swahili (out of 23–24 total classes identified in Bantu languages). The class numbers correspond to the numbers of similar classes used in other Bantu languages, therefore some numbers are missing. As an example, *mtoto* "child" is class 1; *watoto* "children" is class 2.

4.4 Number

Number is usually an inflectional feature of nouns and, depending on language, other parts of speech (pronouns, adjectives, determiners, numerals, verbs) that mark agreement with nouns. The main number distinction is **singular** (one) vs. **plural** (many), although there are languages with other distinctions, as described below. Some languages may use the base ('singular') form even for multiple entities, if it is not necessary to emphasize that they are many. Other languages (e.g., Japanese) lack the morphological feature completely; if they need to express multitude, they can use a quantifier word.

Examples (50) and (51) demonstrate that English demonstratives and verbs agree with nouns in number, while the form of adjectives does not change.

(50) *This new car is expensive.* (singular)

(51) *These new cars are expensive.* (plural)

In German, attributive adjectives inflect for number but predicatively used adjectives do not. Determiners and verbs agree with the noun, too.

(52) *Kein neues Auto ist billig.* "No new car is cheap." (singular)

(53) *Keine neuen Autos sind billig.* "No new cars are cheap." (plural)

(54) *Kein billiges Auto ist neu.* "No cheap car is new." (singular)

(55) *Keine billigen Autos sind neu.* "No cheap cars are new." (plural)

Some languages have **dual,** a special form for two entities. The Proto-Slavic language had it and it survived to-date in Slovenian and Sorbian. It also survives in Arabic. Note that while in singular-plural systems plural means "more than one", in singular-dual-plural systems the meaning of plural has shifted to "more than two". The following examples are from Slovenian.

(56) *Ena dama je prišla pozno.* "One lady has come late." (singular)

(57) *Dve dami sta prišli pozno.* "Two ladies have come late." (dual)

(58) *Tri dame so prišle pozno.* "Three ladies have come late." (plural)

Even fewer languages have **trial,** referring to three entities. For example, Biak, an Austronesian language from islands off the New Guinea coast, has a 3rd person trial pronoun *sko* (while in the first two persons, only singular, dual and plural is distinguished). There are also bound pronominal morphemes that cross-reference the subject on the verb. Hence we have: *ibiser* "he/she/it is hungry", *subiser* "they two are hungry", *skobiser* "they three are hungry" and *sibiser* "they are hungry" (Adelaar and Himmelmann, 2005, p. 801).

Paucal means "a few" and is attested e.g. in Warlpiri, a Pama-Nyungan language from Australia: *karli* (singular) means "boomerang", *karlipatu* (paucal) means "a few boomerangs". **Greater paucal** means "more than several but not many". It occurs in Sursurunga, an Austronesian language. And **greater plural** may mean "many, all possible" entities. Precise semantics varies across languages.

In Kiowa (a North American language spoken in Oklahoma), nouns fall into one of four classes and each class has a basic, default number interpretation. If we want to express a different number than the default, we need to transform the noun into what is called the **inverse number** (Sutton, 2010). Thus *tógúl* "young man" is class I and its basic number meaning is singular or dual. If we transform it to the inverse form, we get the plural: *tógú:dɔ́* "young men". In contrast, class II nouns are dual or plural by default: *ɔ̀nsó:* "feet". The inverse form then gives the singular: *ɔ̀nsôy* "foot". Class III nouns are dual by default and singular or plural in the inverse form: *álɔ̀* "two apples"

vs. *álɔ̀:bɔ̀* "apple, apples". And finally, nouns from class IV have no inverse form and can be interpreted as singular, dual or plural: *hóldà* "dress, dresses". The number is cross-referenced also by the verb (Watkins, 1984, p. 97). In (59) and (60), the subject *ɸ́:dè sân* is in the basic form, which corresponds to singular or dual. The verb further disambiguates the number by using the dual prefix in (60), whereas the absence of the prefix in (59) signals the singular. In (61), all three words are in their inverse forms. Since the subject is a class I noun, it is interpreted as plural.

(59) *ɸ́:dè sân khópdɔ́:* "This child is sick." (basic, singular)

(60) *ɸ́:dè sân ɸ̀khópdɔ́:* "These two children are sick." (basic, dual)

(61) *ɸ́:gɔ̀ sɸ̂:dɔ̀ ɸ̀khópdɔ́:* "These children are sick." (inverse, plural)

There are two options how to encode this in a tagged corpus of Kiowa. One may try to resolve the context to one of the three target meanings, singular, dual, and plural. Or one may introduce new features for basic vs. inverse number, thus modeling more adequately the actual morphology of the language, and leaving for downstream applications to decide the actual meaning.

Assigning a default number value to a lexical unit is not restricted to Kiowa and related languages. In a much subtler way, something remotely similar happens also in European languages. There is a class of nouns called **plurale tantum;** these words occur only in plural, even though they denote a single entity (semantic singular).[2] English examples include *scissors* and *pants*; their Czech counterparts, *nůžky* and *kalhoty*, are plurale tantum as well. Some tagsets will tag them with a different tag than normal plurals, although the plural tag is obviously not wrong, as they behave grammatically like plurals. If the language also marks gender, the non-existence of singular form sometimes means that the gender is unknown. In Czech, special type of numerals is used when counting nouns that are plurale tantum: compare *tři košile* "three shirts" vs. *troje kalhoty* "three pairs of pants".

Similarly, there are nouns that normally occur only in the singular form, although some of them may describe a set of entities (semantic plural). They are called variously **singulare tantum,** collective or mass nouns. Although in theory they might be able to form plural, in practice it would be rarely semantically plausible. Sometimes, the plural form exists and means "several sorts of" or "several packages of". Example: Czech *lidstvo* "mankind". Again, a few tagsets have a special tag for this lexical singular, while the obvious default is to tag them just as singular.

Since morphosyntactic descriptions should provide a different tag for each form of a lemma, one occasionally has to add new tags (features) that account for special forms. In the domain of grammatical number, one such form is attested in Bulgarian and Macedonian: the so-called **count plural** (also known as "counting form" or

[2] Historically, this single entity may be viewed as a pair of parts, which explains why the grammar treats it as plural.

"quantitative plural" (Sussex and Cubberley, 2011, p. 324)). It is a special plural form of nouns if they occur after numerals. The form originates in the Proto-Slavic dual but it should not be marked as dual because 1. the dual has vanished from Bulgarian and 2. the form is no longer semantically tied to the number two. For instance, consider the noun *стол (stol)* "chair". Its normal plural form (without a number) is *столове (stolove)*. However, the count plural is different: *три стола (tri stola)* "three chairs".

4.5 Case

Case is usually an inflectional feature of nouns, pronouns, and, depending on language, other parts of speech that mark agreement with nouns (adjectives, determiners, numerals and even verbs).

Case helps specify the role of the noun phrase in the sentence, especially in free-word-order languages. For example, the nominative and accusative cases often distinguish subject and object of the verb, while in fixed-word-order languages these functions would be distinguished merely by the positions of the nouns in the sentence. Other cases may express relations between two noun phrases (most notably the genitive case), and yet other cases specify the semantics of various adjunct constituents, such as location or instrument. The functions of morphological case forms are very similar to the functions of adpositions. English almost completely lost cases (except for the nominative-accusative forms of some pronouns), but it can use the preposition *of* where other languages would use the genitive case. Many Indo-European languages have mixed systems with both adpositions and morphological cases. At the other end of the scale, Hungarian has relatively few adpositions, but more than 20 cases that include various temporal, local and directional meanings.

Sometimes it is merely a question of the writing conventions whether a morpheme is considered an affix, or a separate word – adposition. If several words in the noun phrase change their forms according to the case, then we have genuine case inflection. But if the 'affix' appears always only on the first or the last word of the noun phrase, then it is more like an adposition; nevertheless, if it is written together with the host word and tokenization rules do not cut it off, the word can be tagged as a case form.

In languages that use both adpositions and morphological cases, the adposition may require a particular case form of its noun phrase. Sometimes more than one case is permitted, typically with different meaning. For instance the German preposition *auf* "on" occurs either with dative noun phrases, meaning location (*auf der Insel* "on the island"), or with accusative noun phrases, meaning direction (*auf die Insel* "to the island"). Some tagsets will encode the case feature of prepositions, expressing their valency (saying that the adposition requires its argument to be in that case).

English has two case forms for some personal pronouns: *he, she, who* (nominative) vs. *him, her, whom* (accusative). German has four cases that are rarely marked morphologically directly on the noun, but they can be recognized by the forms of the article or adjective (combined with the knowledge of the noun's gender): *der Mann* "the man"

(nominative), *des Mannes* (genitive), *dem Mann* (dative), *den Mann* (accusative). Czech has seven cases: *matka* "mother" (nominative), *matky* (genitive), *matce* (dative), *matku* (accusative), *matko* (vocative), *matce* (locative), *matkou* (instrumental). While the dative and locative forms in this example are identical, the cases are different in plural: *matkám* (dative) vs. *matkách* (locative).

Note that Indian corpora based on the so-called Paninian model use a related feature called **vibhakti.** It is a merger of the case feature described here and of various postpositions. Values of the feature are language-dependent because they are copies of the relevant morphemes (either bound morphemes or postpositions). Vibhakti can be mapped on the case values described here if we know 1. which source values are bound morphemes (postpositions are separate nodes for us) and 2. what is their meaning. For instance, the genitive case in Bengali is marked using the suffix -র *(-ra)*, i.e., vibhakti=era. In Hindi, the suffix has been split off the noun and it is now written as a separate word – the postposition का/की/के *(kā/kī/ke)*. Even if the postpositional phrase can be understood as a genitive noun phrase, the noun is not in genitive. Instead, the postposition requires that it takes one of three case forms that are marked directly on the noun: the oblique (accusative) case.

There are many different cases in the languages of the world and it is unlikely that we have managed to describe all of them here. Also note that a case may have more than one name. As with other linguistic terms, the exact meaning is never the same in two different languages, but the core semantics of two differently named cases may still be similar enough to justify covering them by one label.

The descriptions of the individual case values below include semantic hints about the prototypical meaning of the case. Bear in mind that quite often a case will be used for a meaning that is totally unrelated to the meaning mentioned here. Valency of verbs, adpositions and other words will determine that the noun phrase must be in a particular grammatical case to fill a particular valency slot (semantic role). It is much the same as trying to explain the meaning of prepositions: most people would agree that the central meaning of English *in* is location in space or time but there are phrases where the meaning is less locational: *In God we trust. Say it in English.*

In the rest of this section we describe individual case values. We organize them into three groups. Core cases is a relatively small but important group of cases that mark the core grammatical relations such as subject and object. The second group is a diverse mix of cases that are not core but do not have a local or directional meaning, which is the common theme in the third group.

4.5.1 Core Cases

In languages where it exists, **nominative** is the base form of the noun, typically used as the citation form (lemma). This is the word form typically used for subjects of clauses. If the language has only two cases, which are called 'direct' and 'oblique', the direct case corresponds to nominative. Nominative is sometimes claimed to be unmarked,

but in fact there often is a nominative-specific affix that must be attached to the stem in order to create the word form.

Perhaps the second most widely spread morphological case is the **accusative.** Its main function is to mark the direct object of a transitive verb. Accusative normally occurs in languages that also have nominative: nominative is used for the single argument of intransitive verbs and for the more actor-like argument of transitive verbs, while accusative is used for the patient-like argument of transitive verbs. Latin, Greek, some Germanic and Slavic languages can serve as examples of nominative-accusative systems. If a language has just two cases (like the English pronouns), the object case corresponds to the accusative, although it is sometimes called oblique and appears also with prepositions.

Some languages (e.g., Basque in Europe and many languages in Caucasus and Australia) do not use nominative-accusative to distinguish subjects and objects. Instead, they use the contrast of absolutive-ergative. The **absolutive** case marks the single argument of an intransitive verb and the more patient-like argument (direct object) of a transitive verb.[3] Absolutive forms are often (but not always) morphologically unmarked and serve as citation forms in their languages.

The **ergative** case marks the more actor-like argument of a transitive verb. It is usually marked with an affix. So the main difference from the nominative-accusative systems is that the undergoer in transitive clauses aligns with (has the same form as) the single argument of intransitive clauses (which can be understood as either doer or undergoer, depending on the verb). In nominative-accusative systems, the doer of the transitive clause aligns with the single argument of intransitives.

The following examples (62) to (64) show the Czech equivalents of "storm" and "ship" in nominative *(bouře, loďka)* and accusative *(bouři, loďku)*. Examples (65) and (66) show the Basque equivalents in absolutive *(ekaitza, itsasontzia)* and ergative *(ekaitzak, itsasontziak)*; *-k* is the typical ergative suffix in Basque.

(62) Czech: *Bouře byla silná.* (lit. *storm.NOM was strong*) "The storm was strong."

(63) Czech: *Loďka se potopila.* (lit. *ship.NOM REFL sunk*) "The ship has sunk."

(64) Czech: *Bouře potopila loďku.* (lit. *storm.NOM sunk ship.ACC*) "The storm has sunk the ship."

(65) Basque: *Itsasontzia hondoratu zen.* (lit. *ship.ABS sunk has*) "The ship has sunk."

(66) Basque: *Ekaitzak itsasontzia hondoratu du.* (lit. *storm.ERG ship.ABS sunk has*) "The storm has sunk the ship."

[3] The claim that the single argument of intransitives is in the absolutive is a simplification. Some languages show "split intransitivity", where arguments of some verbs align with doers of transitive verbs, others align with undergoers.

4.5.2 Non-core Non-local Cases

Many languages have a **dative** case that marks indirect objects of verbs. Typically they are the addressees or beneficiaries in clauses describing giving or other transfer of ownership. The dative may also mark the goal in clauses describing movement from one place to another. An example:

(67) German: *Ich gebe meinem Bruder ein Geschenk.* (lit. *I.NOM give my.DAT brother.DAT a.ACC present.ACC*) "I give my brother a present."

Another widespread case is called **genitive.** Its prototypical meaning is that the noun phrase somehow belongs to its governor; it would often be translated by the English preposition *of*. English has the 'Saxon genitive' formed by the suffix *'s*; but English tokenization will typically separate the suffix from the noun, hence the noun does not need to be marked as genitive.

Note that despite considerable semantic overlap, the genitive case is not necessarily the same as the feature of possessivity. Semantics of possessivity is much more clearly defined while the genitive (as many other cases) may be required in situations that have nothing to do with possessing. For example, Czech *bez prezidentovy dcery* "without the president's daughter" is a prepositional phrase containing the preposition *bez* "without", the possessive adjective *prezidentovy* "president's" and the noun *dcery* "daughter". The possessive adjective is derived from the noun *prezident* "president" but it is really an adjective (with separate lemma and paradigm), not just a form of the noun. In addition, both the adjective and the noun are in their genitive forms (the nominative would be *prezidentova dcera*). There is nothing possessive about this particular occurrence of the genitive. It is there because the preposition *bez* always requires its argument to be in genitive. Another example:

(68) Czech: *Praha je hlavní město České republiky.* (lit. *Prague.NOM is capital.NOM city.NOM Czech.GEN Republic.GEN*) "Prague is the capital of the Czech Republic."

Basque has two genitives: **possessive genitive** and **locative genitive.** The former is semantically closer to what we described above: *diktadorearen erregimena* (lit. *dictator.PGEN regime.ABS*) "dictator's regime"; the absolutive of "dictator" is *diktadore*.

The **vocative** case is a special form of noun used to address someone. Thus it predominantly appears with animate nouns (see animacy in Section 4.2). Nevertheless this is not a grammatical restriction and inanimate things can be addressed as well.

(69) Czech: *Co myslíš, Filipe?* (lit. *what.ACC you-think, Filip.VOC?*) "What do you think, Filip?"

The name of the **instrumental** case suggests that it is often used for nouns that denote the instrument used to do something (as in Czech *psát perem* "to write using a

pen"). Many other meanings are possible, e.g. in Czech the instrumental is required by the preposition *s* "with" and thus it includes the meaning expressed in other languages by the comitative case.

In Czech the instrumental is also used for the agent-object in passive constructions (cf. the English preposition *by*):

(70) Czech: *Tento zákon byl schválen vládou.* (lit. *this.NOM bill.NOM was approved government.INS*) "This bill has been approved by the government."

A semantically similar case called **instructive** is used rarely in Finnish to express "with (the aid of)". It can be applied to infinitives that behave much like nouns in Finnish.

(71) Finnish: *lähteä* "to leave" ⇒ *2003 lähtien* "since 2003" (second infinitive in the instructive case)

(72) Finnish: *yllättää* "to surprise" ⇒ *sekaantui yllättäen valtataisteluun* (lit. *was-involved-in by-surprise.INS power-struggle.ILL*) "was involved in unexpected power struggle"

The **partitive** is attested e.g. in Finnish. It expresses indefinite identity and unfinished actions without result. It is also used for counted nouns (other languages may use the genitive in similar contexts).

(73) Finnish: *kolme taloa* "three houses" (the base form of "house" is *talo*)

(74) Finnish: *rakastan tätä taloa* "I love this house"

(75) Finnish: *saanko lainata kirjaa?* "can I borrow the book?" (the base form of "book" is *kirja*)

(76) Finnish: *lasissa on maitoa* "there is (some) milk in the glass"

(77) Finnish: *ammuin karhun* (lit. *I-shot bear.ACC*) "I shot a bear (and I know that it is dead)"

(78) Finnish: *ammuin karhua* (lit. *I-shot bear.PAR*) "I shot at a bear (but I may have missed)"

Using accusative instead of partitive may also substitute the missing future tense:

(79) Finnish: *luen kirjan* (lit. *I-read book.ACC*) "I will read the book"

(80) Finnish: *luen kirjaa* (lit. *I-read book.PAR*) "I am reading the book"

The **distributive** case conveys that something happened to every member of a set, one in a time. Or it may express frequency. It is attested in Hungarian.

(81) Hungarian: *fejenként* "per capita"

(82) Hungarian: *esetenként* "in some cases"

(83) Hungarian: *hetenként* "once per week, weekly"

(84) Hungarian: *tízpercenként* "every ten minutes"

The **essive** case expresses a temporary state, often it corresponds to English "as a …". A similar case in Basque is called **prolative.** (Sylak-Glassman, 2016) uses the term 'essive' as a synonym for locative, that is, a union of inessive, adessive and superessive. Prolative/translative are mentioned there as synonyms denoting movement along or across a referent point (see also perlative below).

(85) Finnish: *lapsi* "child" ⇒ *lapsena* "as a child / when he/she was child"

(86) Estonian: *laps* "child" ⇒ *lapsena* "as a child"

(87) Basque: *erreformista* "reformer" ⇒ *erreformistatzat* "as a reformer"

The **translative** case in Uralic languages expresses a change of state ("it becomes X, it changes to X"). Also used for the phrase "in language X". In the Szeged Treebank (Csendes et al., 2005), this case is called **factive.** (Sylak-Glassman, 2016) mentions 'prolative/translative' as two synonyms roughly corresponding to what other authors call perlative (see below).

(88) Finnish: *pitkä* "long" ⇒ *kasvoi pitkäksi* "grew long"

(89) Finnish: *englanti* "English language" ⇒ *englanniksi* "in/into English"

(90) Finnish: *kello kuusi* "six o'clock" ⇒ *kello kuudeksi* "by six o'clock"

(91) Estonian: *kell kuus* "six o'clock" ⇒ *kella kuueks* "by six o'clock"

(92) Hungarian: *Oroszlány halott várossá válhat.* (lit. *Oroszlány dead city.TRA could-become.*) "Oroszlány could become a dead city."

The **comitative** (also called **associative**) case corresponds to English "together with …"

(93) Estonian: *koer* "dog" ⇒ *koeraga* "with dog"

The **abessive** case corresponds to the English preposition "without".

(94) Finnish: *raha* "money" ⇒ *rahatta* "without money"

Noun in the **causative** case is the cause of something. In Hungarian it also seems to be used frequently with currency ("to buy something *for* the money") and it also can mean the goal of something. Other grammatical descriptions may call this case **motivative** or **purposive.**

(95) Hungarian: *Egy világcég benzinkútjánál 7183 forintért tankoltam.* (lit. *a world-wide.company petrol.station.ADE 7183 forint.CAU refueled*) "I refueled my car at the petrol station of a world-wide company for 7183 forints."

(96) Hungarian: *Elmentem a boltba tejért.* (lit. *went the shop.ILL milk.CAU*) "I went to the shop to buy milk."

(97) Basque: *jokaera* "behavior" ⇒ *jokaeragatik* "because of behavior"

The **benefactive** case corresponds to the English preposition "for". Some grammatical descriptions use the term **destinative** for a case with similar meaning.

(98) Basque: *mutil* "boy" ⇒ *mutilarentzat* "for boys"

The **considerative** case denotes something that is given in exchange for something else. It is used in Warlpiri (Andrews, 2007, p. 164).

(99) Warlpiri: *miyi* "food" ⇒ *miyiwanawana* "for food" (*Japanangkarlu kaju karli yinyi miyiwanawana.* "Japanangka is giving me a boomerang in exchange for food.")

The **comparative** case means "than X". It marks the standard of comparison and it differs from the comparative degree (Section 4.7), which marks the property being compared. It occurs in Dravidian and Northeast-Caucasian languages.

The **equative** case means "X-like, similar to X, same as X". It marks the standard of comparison and it differs from the equative degree (Section 4.7), which marks the property being compared. It occurs in Turkish.

(100) Turkish: *ben* "I" ⇒ *bence* "like me"

4.5.3 Local, Temporal and Directional Cases

Some language families, such as the Uralic languages, have a complex set of fine-grained locational and directional cases. Indo-European languages typically combine polyfunctional cases with prepositions to achieve the same effect. Still, the Balto-Slavic branch of Indo-European, Sanskrit and Latin have cases like locative and ablative, whose prototypical meaning is locational and directional, respectively.

The **locative** case often expresses location in space or time, which gave it its name. It is a static location rather than direction but there is no fine-grained distinction whether the location is *in*, *at* or *on* something; these shades are decided by prepositions. As with other cases, non-locational meanings also exist and they are not rare. And even in languages that have the locative, some location roles may be expressed using other cases (e.g. because those cases are required by a preposition).

In Slavic languages this is the only case that is used exclusively in combination with prepositions; the Russian name of this case translates as "prepositional case" (but such a restriction does not hold in other languages that have locative). Examples:

Location	Static	Dir to	Dir from
in, at, on	locative	lative	(ablative)
inside, in	inessive	illative	elative
vicinity, at	adessive	allative	ablative
surface, on	superessive	sublative	delative

Table 4.2: System of local and directional cases.

(101) Czech: *V červenci jsem byl ve Švédsku.* (lit. *in July.LOC I-have been in Sweden.LOC*) "In July I was in Sweden."

(102) Czech: *Mluvili jsme tam o morfologii.* (lit. *talked we-have there about morphology.LOC*) "We talked there about morphology." (non-local non-temporal example)

(Sylak-Glassman, 2016) use the term **essive** to denote static location without distinguishing whether it is in, at or on something. In that sense their essive is a synonym for the locative described here; however, note that other authors understand the essive case as an equivalent of the English preposition "as", as described above.

While in Slavic languages the locative case indicates position both in space and time, Hungarian has a separate **temporal** case for time specification:

(103) Hungarian: *hétkor* "at seven (o'clock)"

(104) Hungarian: *éjfélkor* "at midnight"

(105) Hungarian: *karácsonykor* "at Christmas"

The **lative** case denotes movement towards/to/into/onto something. A similar case in Basque is called **directional allative** (Spanish *adlativo direccional*). However, lative is typically thought of as a union of allative, illative and sublative (see below), while in Basque it is derived from allative, which also exists independently.

(106) Basque: *behe* "low" ⇒ *beherantz* "down"

The **inessive** case expresses location inside of something.

(107) Hungarian: *ház* "house" ⇒ *házban* "in the house"

(108) Finnish: *talo* "house" ⇒ *talossa* "in the house"

(109) Estonian: *maja* "house" ⇒ *majas* "in the house"

The **illative** case expresses direction into something.

(110) Hungarian: *ház* "house" ⇒ *házba* "into the house"

(111) Finnish: *talo* "house" ⇒ *taloon* "into the house"

(112) Estonian: *maja* "house" ⇒ *majasse* "into the house"

Some scholars in Estonian also distinguish the **additive,** which has the illative meaning and the traditional grammar thus considers it just an alternative form of illative. Forms of this case exist only in singular and not for all nouns. It is recognized by the MULTEXT-EAST Estonian tagset.

(113) Estonian: *riik* "government" ⇒ *riigisse* "to the government" (singular illative); *riiki* "to the government" (singular additive)

The **elative** case expresses direction out of something.

(114) Hungarian: *ház* "house" ⇒ *házból* "from the house"

(115) Finnish: *talo* "house" ⇒ *talos* "from the house"

(116) Estonian: *maja* "house" ⇒ *majas* "from the house"

The **adessive** case expresses location in the vicinity of something, at or on something. The corresponding directional cases are allative (towards something) and ablative (from something).

(117) Hungarian: *pénztár* "cash desk" ⇒ *pénztárnál* "at the cash desk"

(118) Finnish: *pöytä* "table" ⇒ *pöydällä* "on the table"

(119) Estonian: *laud* "table" ⇒ *laual* "on the table"

Note that adessive is used to express location on the surface of something in Finnish and Estonian, but does not carry this meaning in Hungarian.

The **allative** case expresses direction to something (destination is adessive, i.e., at or on that something).

(120) Hungarian: *pénztár* "cash desk" ⇒ *pénztárhoz* "to the cash desk"

(121) Finnish: *pöytä* "table" ⇒ *pöydäll* "onto the table"

The **ablative** case expresses direction from some point.

(122) Hungarian: *barátomtól jövök* (lit. *friend.my.ABL I-come*) "I am coming from my friend"

(123) Finnish: *pöytä* "table" ⇒ *pöydältä* "from the table"

(124) Finnish: *katolta* "from the roof"

(125) Finnish: *rannalta* "from the beach"

The **superessive** case is used, chiefly in Hungarian, to indicate location on top of something or on the surface of something.

(126) Hungarian: *asztal* "table" ⇒ *asztalon* "on the table"

(127) Hungarian: *könyvek* "books" ⇒ *könyveken* "on books"

The **sublative** case is used in Finno-Ugric languages to express the destination of movement, originally to the surface of something (e.g. "to climb a tree"), and, by extension, in other figurative meanings as well (e.g. "to university").

(128) Hungarian: *Belgrádtól 150 kilométerre délnyugatra* (lit. *Belgrade.ABL 150 kilometer.SUB southwest.SUB*) "150 kilometers southwest of Belgrade"

(129) Hungarian: *hajó* "ship" ⇒ *hajóra* "onto the ship"

(130) Hungarian: *bokorra* "on the shrub"

The **delative** case is used, chiefly in Hungarian, to express the movement from the surface of something (like "moved off the table"). Other meanings are possible as well, e.g. "about something".

(131) Hungarian: *asztal* "table" ⇒ *asztalról* "off the table"

(132) Hungarian: *Budapestről jövök* "I am coming from Budapest"

The **perlative** case denotes movement along something. It is used in Warlpiri (Andrews, 2007, p. 161–162). Note that (Sylak-Glassman, 2016) mentions the English preposition "along" in connection with what they call prolative/translative; but we have shown different definitions of those two cases above.

(133) Warlpiri: *yurutu* "road" ⇒ *yurutuwana* "along the road" (*Pirli kalujana yurutuwana yirrarni.* "They are putting stones along the road.")

The **terminative** case specifies where something ends in space or time. Similar case in Basque is called **terminal allative** (Spanish *adlativo terminal*).

(134) Estonian: *jõeni* "down to the river"

(135) Estonian: *kella kuueni* "till six o'clock"

(136) Hungarian: *házig* "up to the house"

(137) Hungarian: *hat óráig* "till six o'clock"

(138) Basque: *erdi* "half" ⇒ *erdiraino* "up to the half"

4.6 Definiteness

Definiteness is typically a feature of nouns, adjectives and articles. Its value distinguishes whether we are talking about something known and concrete, or something general or unknown. It can be marked on definite and indefinite articles, or directly on nouns, adjectives etc. In Semitic languages, definiteness is also called the **state.**

Indefinite entities are unknown to the addressee and they are new to the discourse: we have not talked about them previously. In English, indefiniteness is primarily conveyed by the indefinite article *a, an*; it can be also expressed with an indefinite determiner or pronoun *(some, somebody)*. In Swedish, the indefinite article is supported by nominal morphology, namely by the absence of a definite morpheme on the noun. Spoken Arabic has the indefinite suffix *-n* applied to nouns; in writing it can be represented by diacritical marks but these are usually omitted.

(139) English: *a dog*

(140) Swedish: *en hund* "a dog"

(141) Arabic: كَلبٌ *(kalbun)* "a dog"

Some languages further divide indefinite nouns into **specific** and **non-specific.** This occurs e.g. in Lakota. For instance, with the noun "stick", specific determiner would refer to "a certain stick" while non-specific determiner means "any (one) stick".

Definite entities are known to the addressee, they have been mentioned in the previous discourse, or can be seen or otherwise sensed by both the speaker and the addressee. In English, definiteness is primarily conveyed by the definite article *the*; it can be also expressed with a demonstrative determiner or pronoun *(this, that)*. In Swedish, Bulgarian or Romanian, a definite suffix is attached directly to the noun. Arabic and Hebrew have a definite prefix; some tokenization schemes may cut it off and treat it as a separate definite article.

(142) English: *the dog*

(143) Swedish: *hunden* "the dog"

(144) Arabic: اَلكَلبُ *(al-kalbu)* "the dog"

Definiteness can be morphologically marked also on adjectives. In Arabic, adjectives agree in definiteness with the modified noun, i.e., all adjectives and nouns in the noun phrase bear the definite prefix *al-* (145). In German, the noun itself does not show definiteness and an article must be used like in English. However, adjectives still reflect it and take different forms in definite and indefinite phrases (146) and (147).

(145) Arabic: المملكة الأردنّيّة الهاشميّة *(al-mamlakatu al-ʾurdunnīyatu al-hāšimīyatu)* "Hashemite Kingdom of Jordan"

(146) German: *ein schwarzer Hund* "a black dog"

(147) German: *der schwarze Hund* "the black dog"

The Semitic languages further distinguish a special state of reduced definiteness, called **construct state** (status constructus) (Fischer, 1997, p. 195). If two nouns are in genitive relation, the definiteness of the first one (the 'nomen regens') is reduced. The second noun is the genitive modifier and can be either definite or indefinite. The

reduced form has neither the definite prefix (article), nor the indefinite suffix (nunation).

(148) Arabic: حلوَةٌ *(ḥulwatun)* "a sweet" (indefinite state)

(149) Arabic: الحلوَةُ *(al-ḥulwatu)* "the sweet" (definite state)

(150) Arabic: حلوَةُ *(ḥulwatu)* "sweet of" (construct state)

Another definiteness-related and Arabic-specific phenomenon is **improper annexation.** The genitive construction described above normally consists of two nouns (first reduced, second genitive). That is called proper annexation or *iḍāfa*. If the first member is an adjective or adjectivally used participle and the second member is a definite noun, the construction is called improper annexation or *false iḍāfa*. The result is a compound adjective that is usually used as an attributive adjunct and thus must agree in definiteness with the noun it modifies. Its first part (the adjective or participle) may get again the definite article. Although it may look the same as the definite state, the origin of the form is different (Hajič et al., 2004, p. 3).

(151) Arabic: مُخْتَلِفٌ *(muxtalifun)* "different/various" (active participle, Form VIII)

(152) Arabic: نَوْعٌ ج أَنْوَاعٌ *(nawʾun ja anwāʾun)* "kind"

(153) Arabic: مُخْتَلِفُ الأَنْوَاعِ *(muxtalifu al-anwāʾi)* "of various kinds" (false iḍāfa)

(154) Arabic: مَشَاكِلُ مُخْتَلِفَةُ الأَنْوَاعِ *(mašākilu muxtalifatu al-anwāʾi)* "problems of various kinds" (indefinite phrase)

(155) Arabic: اَلْمَشَاكِلُ الْمُخْتَلِفَةُ الأَنْوَاعِ *(al-mašākilu **al-muxtalifatu** al-anwāʾi)* "the problems of various kinds" (definite phrase)

4.7 Degree of Comparison

Degree of comparison is typically an inflectional feature of some adjectives and adverbs. It allows to express that one entity shows a higher (or same or lower) degree of the property denoted by the adjective than another entity. The other entity is termed **standard of comparison.** Some languages choose to mark morphologically the standard of comparison rather than the adjective—then we have a comparative case (Section 4.5) rather than a comparative degree.

While some languages use inflection to express comparison, in other languages it is a periphrastic construction with the help of function words like "more", "less" or "than". English employs a combination of both approaches: some adjectives can be compared morphologically *(smart – smarter – smartest)* while others only periphrastically *(intelligent – more intelligent – most intelligent)*.

If an adjective has morphological degrees, the first degree is the basic, uncompared form and it is called **positive.** This term is widely accepted but it can be misleading if

the language also has morphological means of changing polarity (Section 4.8). Using the word 'positive' for a negated adjective may be confusing but it is inevitable because negative properties can be compared, too.

(156) English: *young man*

(157) Czech: *mladý muž* "young man"

(158) Czech: *nechutné jídlo* "unpalatable food"

When the quality of one object is compared to the same quality of another object, and the result is that they are identical or similar ("as X as"), the degree of the quality is **equative.** Note that it marks the adjective and it is distinct from the equative case, which marks the standard of comparison.

(159) Estonian: *pikkus* "tall" ⇒ *pikkune* "as tall as"

When the quality of one object is compared to the same quality of another object and the degree of the compared is found higher than the degree of the standard of comparison, the adjective takes the **comparative** form (also called the **second degree** in some grammatical traditions).

(160) English: *The man is **younger** than me.*

(161) Czech: *Ten muž je **mladší** než já.* "The man is **younger** than me."

(162) Czech: *Moje jídlo je **nechutnější** než tvoje.* (lit. *my food is more-unpalatable than yours*) "My food is less palatable than yours."

When the degree of the quality of one object is found higher than the degree of the same quality of all other objects within a set, the adjective takes the **superlative** form (also called the **third degree** in some grammatical traditions).

(163) English: *This is the **youngest** man in our team.*

(164) Czech: *Toto je **nejmladší** muž v našem týmu.* "This is the **youngest** man in our team."

(165) Czech: *Moje jídlo je **nejnechutnější** ze všech.* (lit. *my food is most-unpalatable from all*) "My food is the least palatable of all."

Some languages can express morphologically that the studied quality of the given object is so strong that there is hardly any other object exceeding it. The quality is not actually compared to any particular set of objects. This is called the **absolute superlative.**

(166) Spanish: *guapo* "handsome" ⇒ *guapísimo* "indescribably handsome"

All the above examples demonstrate comparison of adjectives. Some adverbs can be compared as well, especially those manner adverbs that can be related to (or derived from) adjectives. This process can be better observed in Czech than in English:

(167) Czech: *Martin běhá* ***rychle.*** "Martin runs fast." (positive adverb) Cf. adjective: *Martin je* ***rychlý.*** "Martin is fast."

(168) Czech: *Martin běhá* ***rychleji*** *než Petr.* "Martin runs faster than Petr." (comparative adverb) Cf. adjective: *Martin je* ***rychlejší*** *než Petr.* "Martin is faster than Petr."

(169) Czech: *Martin běhá* ***nejrychleji*** *ze všech.* "Martin runs the fastest of all." (superlative adverb) Cf. adjective: *Martin je* ***nejrychlejší*** *ze všech.* "Martin is the fastest of them all."

4.8 Polarity

Polarity is typically a feature of verbs, adjectives, sometimes also adverbs and nouns in languages that negate using bound morphemes. In languages that negate using a function word, that function word can be seen as bearing the feature of (negative) polarity.

Unnegated words are **affirmative;** one could also refer to them as **positive** but then some care is needed so that the feature is not confused with the positive degree of comparison (Section 4.7). Positive polarity (affirmativeness) is rarely, if at all, encoded using overt morphology. If a tagset encodes it explicitly, then it usually means to signal that a lemma has **negative** forms but this particular form is not negative.

For instance, all Czech verbs and adjectives can be negated using the prefix *ne-*. In theory, all nouns can be negated too, with the meaning "anything except the entities denotable by the original noun". However, negated nouns are rare.

In English, verbs are negated using the particle *not* and adjectives are also negated using prefixes, although the process is less productive than in Czech *wise – unwise, probable – improbable*. Nouns can be negated using the prefix *non-*.

Both languages also have negative determiners, pronouns and pronominal adverbs *(no, nobody, nothing, nowhere, never)*. These may be already recognizable by the lexical feature of pronominal type, but they can be additionally also understood and annotated as bearing the negative polarity. (This is similar to the situation with (in)definiteness, which is primarily conveyed through indefinite articles or bound morphemes on nouns, but there is also a class of indefinite pronouns and determiners.)

The polarity feature can be also used to distinguish response particles / interjections *yes* and *no*.

Affirmative examples:

(170) Czech: *přišel* "he came"

(171) Czech: *velký* "big"

(172) English: *yes*

Negative examples:

(173) Czech: *nepřišel* "he did not come"

(174) Czech: *nevelký* "not big"

(175) English: *not*

(176) English: *no*

4.9 Person

Person is typically a feature of personal and possessive pronouns / determiners, and of verbs. On verbs it is in fact an agreement feature that marks the person of the verb's subject (some languages, e.g. Basque, can also mark person of objects). Person marked on verbs makes it unnecessary to always add a personal pronoun as subject and thus subjects are sometimes dropped (pro-drop languages).

In singular, the **first person** refers just to the speaker / author. In plural, it must include the speaker and one or more additional persons. Some languages (e.g. Taiwanese) distinguish inclusive and exclusive first person plural pronouns; see Section 4.10 for details.

(177) English: *I, we*

(178) Czech: *dělám* "I do"

The **second person** singular refers to the addressee of the utterance / text (i.e., the listener / reader). In plural, it may mean several addressees and optionally some third persons too.

(179) English: *you*

(180) Czech: *děláš* "you do"

The **third person** refers to one or more persons that are neither speakers nor addressees.

(181) English: *he, she, it, they*

(182) Czech: *dělá* "he/she/it does"

Grammatical descriptions of some languages also recognize the **fourth person** but the term has different meaning in grammars of different languages. In Finnish it refers to indefinite or generic referents; for that meaning, see the 'zero person' below. In several North American languages, the fourth person can be understood as a third person

argument morphologically distinguished from another third person argument. Alternatively, both arguments can be labeled as third person and an additional feature can be used that marks one of them as **proximate** (more topical third person), the other as **obviative** (less topical third person).

The **zero person** (sometimes also called the fourth person) is used in impersonal statements where the referent is indefinite or generic; the closest English equivalent would be the pronoun *one* (however, when subject-verb agreement is considered, English *one* behaves like a third person subject). Similarly, the terms 'zero person' or 'fourth person' are used for impersonal constructions in Finnish but, again, there is no unique morphology that could be labeled as zero person; instead, the verb takes its third-person singular form with no subject. Yet there is a language that has such overt morphology: Keres, a language isolate of New Mexico (Davis, 1964, p. 75). The fourth (zero) person is used "when the subject of the action is obscure, as when the speaker is telling of something that he himself did not observe. It is also used when the subject of the action is inferior to the object, as when an animal is the subject and a human being the object."

(183) Keres: *gàku* "he (third person) bit him"

(184) Keres: *c̓àku* "he (fourth person) bit him"

4.10 Clusivity

Clusivity is a feature of first-person plural personal pronouns; it can be viewed as an elaboration of the system of persons. It is attested in various language families outside Europe, for instance in the Austronesian languages.

Clusivity is a bipolar distinction between inclusive and exclusive pronouns. The **inclusive** *we* is understood as including the addressee; **exclusive** *we* refers to the speaker and one or more third persons, but not to the listener.

(185) Indonesian: *kita* "we" (inclusive: I + you + maybe they)

(186) Indonesian: *kami* "we" (exclusive: I + they)

While in many languages clusivity is just a lexical feature of pronouns, it can be also reflected (together with person and number) in verbal morphology that cross-references arguments of the verb (Wolvengrey, 2011):

(187) Plains Cree: *niwīcihāw* "I help him"

(188) Plains Cree: *kiwīcihāw* "you.Sing help him"

(189) Plains Cree: *niwīcihānān* "we.Excl help him"

(190) Plains Cree: *kiwīcihānaw* "I+you (we.Incl) help him"

(191) Plains Cree: *kiwīcihāwāw* "you.Plur help him"

4.11 Politeness

Various languages have various means to express politeness or respect; some of the means are morphological. Three to four dimensions of politeness are distinguished in linguistic literature (Brown and Levinson, 1987; Comrie, 1976; Wenger, 1982).

- speaker-referent axis (meant to include the addressee when he happens to be the referent)
- speaker-addressee axis (word forms depend on who is the addressee, although the addressee is not referred to)
- speaker-bystander axis
- speaker-setting axis

Large part of politeness considerations is lexical: certain parts of the vocabulary may not be appropriate in a given situation. This is especially true for the speaker-setting axis, also called style or register. We will look a bit more closely at the first two axes: speaker-referent and speaker-addressee.

Changing pronouns and/or person and/or number of the verb forms when respectable persons are addressed in Indo-European languages belongs to the speaker-referent axis because the honorific pronouns are used to *refer* to the addressee.

In Czech, formal second person has the same form for singular and plural, and is identical to informal second person plural. This involves both the pronoun and the finite verb but not a participle, which has no special formal form (that is, formal singular is identical to informal singular, not to informal plural).

In German, Spanish or Hindi, both number and person are changed (informal third person is used as formal second person) and in addition, special pronouns are used that only occur in the formal register (German *Sie*; Spanish *usted, ustedes*; Hindi आप *(āpa)*).

In Japanese, verbs and other words have polite and informal forms but the polite forms are not referring to the addressee (they are not in second person). They are just used because of who the addressee is, even if the topic does not involve the addressee at all. This kind of polite language is called teineigo (丁寧語) and belongs to the speaker-addressee axis.

If a language distinguishes levels of politeness, then the **informal** register is usually meant for communication with family members and close friends.

(192) Czech: *ty jdeš / vy jdete* "you go.SING/PLUR"

(193) German: *du gehst / ihr geht* "you go.SING/PLUR"

(194) Spanish: *tú vas / vosotros vais* "you go.SING/PLUR"

(195) Japanese: 行かない *(ikanai)* "will not go"

In contrast, the **formal / polite** register is usually meant for communication with strangers and people of higher social status than that of the speaker.

(196) Czech: *vy jdete* "you go.SING/PLUR"

(197) German: *Sie gehen* "you go.SING/PLUR"

(198) Spanish: *usted va / ustedes van* "you go.SING/PLUR"

(199) Japanese: 行きません *(ikimasen)* "will not go"

The formal register in Japanese has several layers. There is a set of honorific forms, called sonkeigo (尊敬語), that elevate the status of the referent; and there are other honorific forms, called kenjōgo (謙譲語), that lower the speaker's status, thereby raising the referent's status by comparison. Both (200) and (201) can be translated as "to do" and both can be combined with the formal suffix -ます *(-masu)*. In addition, (200) elevates the status of the referent. This version of "to do" is suitable when talking about actions of a customer or a superior. (Note that Japanese verbs do not cross-reference the person and number of the subject, hence the elevating register may provide an additional hint as to who we are talking about.) In contrast, (201) is an example of the 'humbling speech' that lowers the status of the speaker; this version of "to do" is suitable when referring to one's own actions or the actions of a group member.

(200) Japanese: なさる *(nasaru)*, なさいます *(nasaimasu)* "to do" (elevating the referent)

(201) Japanese: いたす *(itasu)*, いたします *(itashimasu)* "to do" (humbling the speaker)

4.12 Deixis

We have discussed deixis as a general characteristic of words whose reference cannot be determined solely from their lexical meaning and must be resolved according to the context in which they are used. We have categorized deictic words as pronouns, determiners and pronominal adverbs (among others) and we have treated the terms 'deictic' and 'pronominal' as more or less synonymous (see Section 3.5.5 and the discussion of pronominal types in the beginning of Chapter 4). Such a determination may be seen by some as too broad; in a narrower sense, deixis is associated in particular with demonstratives. These are often further categorized along various dimensions. (Sylak-Glassman, 2016, p. 22–24) lists four: **distance, reference point, visibility** and **verticality.**

Distance is the distinction between English *this* and *that*, *here* and *there*, *now* and *then*. Some languages (e.g. Spanish or Basque) have three values of this feature (instead of the two seen in English) and they can be labeled as **proximate, medial** and **remote.** Reference point is often interrelated with distance and the meaning of the three distance values may correspond to "near the speaker", "near the addressee", and "distant from both", respectively. However, (Sylak-Glassman, 2016) shows that the two aspects are partially independent in Hausa, thus they deserve to be treated separately.

	Unposs	1 Sing	2 Sing	3 Sing	1 Plur	2 Plur	3 Plur
Sing	*ház*	*házam*	*házad*	*háza*	*házunk*	*házatok*	*házuk*
Plur	*házak*	*házaim*	*házaid*	*házai*	*házaink*	*házaitok*	*házaik*

Table 4.3: The possessor-referencing forms of the Hungarian noun *ház* "house". The forms in the first column are unpossessed, the other columns correspond to different possessors. The rows represent the number of the possessed entity.

Several languages also distinguish pronouns depending on whether their referent is visible or not. Finally, it may be also important whether the referent is above the plane of the speaker, below it, or at the same level. Both distinctions can be observed in the masculine singular demonstratives of Khasi (an Austro-Asiatic language of eastern India (Bhat, 2004, p. 133)):

(202) Khasi: *une* "he (near)"

(203) Khasi: *uto* "he (not near, not far)"

(204) Khasi: *utay* "he (far away, visible)"

(205) Khasi: *uto* "he (far away, not visible)"

(206) Khasi: *utey* "he (above)"

(207) Khasi: *uthie* "he (below)"

4.13 Cross-reference of Possessor

While some languages use possessive determiners, adjectives or the genitive case to express possession, in other languages the possessed noun can take affixes that cross-reference the features of the possessor. Typically involved are person (Section 4.9) and number (Section 4.4), sometimes also gender (Section 4.1). Some tagsets represent the possessor reference as a single feature, although it could be decomposed into person and number (and gender, if applicable). If it is decomposed and represented as atomic features, some mechanism is needed to distinguish multiple layers of a feature (see Section 3.2.2 for details). The reason is that the possessed noun typically has its own number value, which must be distinguished from the number of the possessor.

Table 4.3 shows how possessors are cross-referenced by the forms of the Hungarian noun *ház* "house": *házam* means "my house", *házaik* "their houses" and *házatok* "your house" (you is plural, the house is in singular). A case suffix, if needed, will be attached after the possessive suffix: *házban* "in house", *házamban* "in my house", *házukban* "in their house" etc.

	Sing	Plur
1	*tengo*	*tenemos*
2	*tienes*	*tenéis*
3	*tiene*	*tienen*

Table 4.4: Present indicative forms of the Spanish verb *tener* "to have", cross-referencing the person and number of the subject.

	Sing 1	Sing 2I	Sing 2F	Sing 3	Plur 1	Plur 2	Plur 3
Abs	*naiz*	*haiz*	*zara*	*da*	*gara*	*zarete*	*dira*
Dat	*zait*	*zaik*	*zaizu*	*zaio*	*zaigu*	*zaizue*	*zaie*
Erg	*dut*	*duk*	*duzu*	*du*	*dugu*	*duzue*	*dute*

Table 4.5: Present indicative forms of the Basque auxiliary in intransitive clauses, depending on the case of the single argument that is cross-referenced. The second-person singular forms are either informal or formal.

4.14 Cross-reference of Verbal Arguments

Similarly to cross-referencing of possessors on nouns, verbal arguments can be cross-referenced on verbs. This phenomenon is not present in all languages—for example, Chinese or Vietnamese verbs get along with just one form. Most European languages cross-reference only the subject, which is known as the subject-verb agreement. English verb paradigms are quite deficient in this respect, with only the third-person singular present differing from the rest (and with the exception of the verb *to be*). A better example is Spanish where a present indicative verb has 6 forms according to the different combinations of person and number of the subject (Table 4.4). The subject reference is usually annotated as two features of the verb, person and number.

However, there are languages that can cross-reference more than one argument of the verb. A European example is Basque with auxiliary verbs that can cross-reference arguments of three types: an absolutive argument, a dative argument and an ergative argument (though at most two of them are cross-referenced at the same time). Similarly to the number feature of possessed nouns, some mechanism for layered features (Section 3.2.2) is needed in order to distinguish person and number of the different arguments. Alternatively, three language-specific features can be defined for the three types of arguments, and their values will be combinations of person and number. Tables 4.5, 4.6, 4.7 and 4.8 illustrate the argument marking on present indicative forms of Basque auxiliaries.

Abs/Dat	Sing 1	Sing 2I	Sing 2F	Sing 3	Plur 1	Plur 2	Plur 3
Sing 1	–	*natzaik*	*natzaizu*	*natzaio*	–	*natzaizue*	*natzaie*
Sing 2I	*hatzait*	–	–	*hatzaio*	*hatzaigu*	–	*hatzaie*
Sing 2F	*zatzaizkit*	–	–	*zatzaizkio*	*zatzaizkigu*	–	*zatzaizkie*
Sing 3	*zait*	*zaik*	*zaizu*	*zaio*	*zaigu*	*zaizue*	*zaie*
Plur 1	–	*gatzaizkik*	*gatzaizkizu*	*gatzaizkio*	–	*gatzaizkizue*	*gatzaizkie*
Plur 2	*zatzaizkidate*	–	–	*zatzaizkiote*	*zatzaizkigute*	–	*zatzaizkiete*
Plur 3	*zaizkit*	*zaizkik*	*zaizkizu*	*zaizkio*	*zaizkigu*	*zaizkizue*	*zaizkie*

Table 4.6: Present indicative forms of the Basque auxiliary, cross-referencing an absolutive and a dative argument. The forms corresponding to third person singular absolutive are also used if there is just a single dative argument.

Abs/Erg	Sing 1	Sing 2I	Sing 2F	Sing 3	Plur 1	Plur 2	Plur 3
Sing 1	–	*nauk*	*nauzu*	*nau*	–	*nauzue*	*naute*
Sing 2I	*haut*	–	–	*hau*	*haugu*	–	*haute*
Sing 2F	*zaitut*	–	–	*zaitu*	*zaitugu*	–	*zaitute*
Sing 3	*dut*	*duk*	*duzu*	*du*	*dugu*	*duzue*	*dute*
Plur 1	–	*gaituk*	*gaituzu*	*gaitu*	–	*gaituzue*	*gaitute*
Plur 2	*zaituztet*	–	–	*zaituzte*	*zaituztegu*	–	*zaituztete*
Plur 3	*ditut*	*dituk*	*dituzu*	*ditu*	*ditugu*	*dituzue*	*ditute*

Table 4.7: Present indicative forms of the Basque auxiliary, cross-referencing an absolutive and an ergative argument. The forms corresponding to third person singular absolutive are also used if there is just a single ergative argument.

Dat/Erg	Sing 1	Sing 2I	Sing 2F	Sing 3	Plur 1	Plur 2	Plur 3
Sing 1	–	*dizkidak*	*dizkidazu*	*dizkit*	–	*dizkidazue*	*dizkidate*
Sing 2F	*dizkizut*	–	–	*dizkizu*	*dizkizugu*	–	*dizkizute*
Sing 3	*dizkiot*	*dizkiok*	*dizkiozu*	*dizkio*	*dizkiogu*	*dizkiozue*	*dizkiote*
Plur 1	–	*dizkiguk*	*dizkiguzu*	*dizkigu*	–	*dizkiguzue*	*dizkigute*
Plur 2	*dizkizuet*	–	–	*dizkizue*	*dizkizuegu*	–	*dizkizuete*
Plur 3	*dizkiet*	*dizkiek*	*dizkiezu*	*dizkie*	*dizkiegu*	*dizkiezue*	*dizkiete*

Table 4.8: Present indicative forms of the Basque auxiliary, cross-referencing a dative and an ergative argument. These forms are also used in clauses with three arguments, although they do not change for different persons and numbers of the absolutive argument.

4.15 Tense

Tense is typically a feature of verbs. It may also occur with other parts of speech (nouns, adjectives, adverbs), depending on whether borderline word forms such as participles are classified as verbs or as the other category.

Tense is a feature that specifies the time when the action took / takes / will take place, in relation to a reference point. The reference is often the moment of producing the sentence, but it can be also another event in the context. In some languages (e.g. English), some traditionally recognized tenses are actually combinations of tense and aspect (Section 4.16). In other languages (e.g. Czech), aspect and tense are separate, although not completely independent of each other.

Remember that we are defining features that apply to a single word. If a tense is constructed periphrastically (two or more words, e.g. auxiliary verb indicative + participle of the main verb) and none of the participating words are specific to this tense, then the features will probably not directly reveal the tense. For instance, English *I had been there* is past perfect (pluperfect) tense, formed periphrastically by the simple past tense of the auxiliary *to have* and the past participle of the main verb *to be*. The auxiliary will be tagged as past indicative and the participle will be past participle; none of the two will be labeled as pluperfect. On the other hand, Portuguese can form the pluperfect morphologically as just one word, such as *estivera*; the Portuguese tagset may either provide a dedicated tag for pluperfect, or it may combine the features of tense (past) and aspect (perfect).

The **past** tense denotes actions that happened before a reference point. In the prototypical case, the reference point is the moment of producing the sentence and the past event happened before the speaker speaks about it. However, the term 'past' is also used to distinguish past participles from other kinds of participles, and past converbs from other kinds of converbs; in these cases, the reference point may itself be in past or future when compared to the moment of speaking. For instance, the Czech converb *spatřivše* "having seen" in the sentence *spatřivše vojáky, velmi se ulekli* "having seen the soldiers, they got very scared" describes an event that is anterior to the event of getting scared. It also happens to be anterior to the moment of speaking, but that fact is not encoded in the converb itself, it is rather a consequence of "getting scared" being in the past tense.

Traditional grammars use various other terms in connection with the past tense, such as **preterite** or **aorist.** These terms have different meaning in different traditions and languages. Both preterite and aorist are sometimes described as aspect-neutral, which contrasts to the imperfect tense (see below).

Turkish has two past tenses; one of them, the *-miş* past, is dubbed the 'narrative' past. Tagsets may use Turkish-specific labels and treat them as two tenses; but in fact, the difference between them is a difference in a separate feature of evidentiality (Section 4.19).

In the following examples, the highlighted words deserve to be labeled as past tense:

(208) English: *he **went** home*

(209) English: *he has **gone** home*

(210) Portuguese: *afirmou que os sequestradores já **ligaram*** "he said that the kidnappers had already called" (morphological pluperfect)

The **imperfect** tense is a special type of the past tense. It exists in several Romance languages and in a few Slavic languages, such as Bulgarian. The cleanest way of annotating it would be to decompose it into two features, the past tense and the imperfective aspect. Unfortunately, this does not work well in Bulgarian where every verb has a lexical aspect (see Section 4.16), inherent in the verbal lemma, and it does not always match the grammatical aspect. In main clauses, imperfective verbs have the imperfect tense and perfective verbs have perfect tense. However, both rules can be violated in embedded clauses.

(211) Bulgarian: *тя оставаше, където той и да **отиде*** *(tja ostavaše, kădeto toj i da **otideše**)* "she remained where he left her" (perfective verb in the imperfect tense)

The **present** tense denotes actions that are in progress (or states that are valid) at a reference point; it may also describe events that usually happen. In the prototypical case, the reference point is the moment of producing the sentence; however, the term 'present' is also used to distinguish present participles from other kinds of participles, and present converbs from other kinds of converbs. In these cases, the reference point may be in past or future when compared to the moment of speaking. For instance, the English present participle may be used to form a past progressive tense: *he was watching TV when I arrived.*

(212) English: *he **goes** home*

(213) English: *he was **going** home*

The **future** tense denotes actions that will happen after a reference point; in the prototypical case, the reference point is the moment of producing the sentence. English does not have morphological future tense; if anything in English deserves the label at all, then it is the modal verb *will*. However, other languages have dedicated future forms of finite verbs:

(214) Spanish: ***irá** a la casa* "he/she/it will go home"

The three-way distinction of present, past and future is sometimes further divided on the basis of the distance between the reference point and the time of the event. (Sylak-Glassman, 2016) refers to (Comrie, 1985) and provides a survey of the distinctions made in various languages. The first step is splitting the past, the future or

both into **recent** and **remote.** Some languages will also have special verb forms for events that happened / will happen today **(hodiernal)**, yesterday **(hesternal)** or tomorrow. The following examples are from Ngiemboon, a Grassfields Bantu language of Cameroon. All are given in the perfective aspect and the tenses are relative, i.e., the reference point is provided by the larger context (Nurse and Philippson, 2003, p. 246):

(215) Ngiemboon: *à **là nzá** mbáb* "he cut the meat (some time ago)"

(216) Ngiemboon: *à **kà zàʔ** mbàb* "he cut the meat (yesterday)"

(217) Ngiemboon: *à **tó záʔ** mbàb* "he will cut the meat (tomorrow)"

(218) Ngiemboon: *à **lù záʔ** mbàb* "he will cut the meat (some time from now)"

4.16 Aspect

Aspect is typically a feature of verbs. It may also occur with other parts of speech (nouns, adjectives, adverbs), depending on whether borderline word forms such as gerunds and participles are classified as verbs or as the other category.

Aspect is a feature that specifies duration of the action in time, whether the action has been completed etc. In some languages (e.g. English), some tenses (Section 4.15) are actually combinations of tense and aspect. In addition, aspect in English is defined by periphrastic constructions and there are not many opportunities to annotate it as a morphological feature—perhaps with the exception of the participles. In other languages (e.g. Czech), aspect and tense are separate, although not completely independent of each other. In yet other languages, aspect is said to be much more important in the verbal paradigm than tense.

In Czech and other Slavic languages, aspect is a lexical feature. Pairs of imperfective and perfective verbs exist and are often morphologically related but the space is highly irregular and the verbs are considered to belong to separate lemmas. In Bulgarian this leads to a conflict between lexical and grammatical aspect (the imperfect tense—see Section 4.15).

The **iperfect(ive)** aspect means that the action took / takes / will take some time span and there is no information whether and when it was / will be completed.

(219) Czech: *péci* "to bake.IMP" ⇒ ***pekl** chleba* "he baked / was baking a bread"

The **perfect(ive)** aspect means that the action has been / will have been completed. Since there is emphasis on one point on the time scale (the point of completion), this aspect does not work well with the present tense. For example, Czech morphology can create present forms of perfective verbs but these actually have a future meaning.

(220) Czech: *upéci* "to bake.PERF" ⇒ ***upekl** chleba* "he baked / has baked a bread"

The **prospective** aspect can be described as relative future: the action is / was / will be expected to take place at a moment that follows the reference point; the

reference point itself can be in past, present or future. In the English sentence *When I got home yesterday, John called and said he would arrive soon,* the last clause *(he would arrive soon)* is in the prospective aspect. Nevertheless, English does not have overt affixal morphemes dedicated to the prospective aspect, and we do not need the label in English. But other languages do; the *-ko* suffix in Basque is an example.

(221) Basque: *Liburua **irakurriko** behar du.* (lit. *book-a read.PROSP must he-does*) "He must go to read a book."

English progressive tenses *(I am eating, I have been doing …)* can serve as examples of the **progressive** aspect. They are constructed analytically (auxiliary + present participle) but the *-ing* participle is so bound to the progressive meaning that it seems a good idea to annotate it with this feature (we have to distinguish it from the past / perfect participle somehow; we may use both the tense and aspect features to mark the difference).

In languages other than English, the progressive meaning may be expressed by morphemes bound to the main verb, which makes this value even more justified. Example is Turkish with its two distinct progressive morphemes, *-yor* and *-mekte*.

(222) Turkish: *eve **gidiyor*** "she is going home (now)"

(223) Turkish: *eve **gitmekte*** "she is going home (now)"

(224) Turkish: *eve **gidiyordu*** "she was going home (when I saw her)"

(225) Turkish: *eve **gimekteydi*** "she was going home (when I saw her)"

The **habitual** aspect is used to describe events that usually happen but we do not want to point at one particular event and anchor it in time. The English simple present tense implies the habitual aspect in most contexts.

(226) English: *He **speaks** five languages.*

The **iterative** or **frequentative** aspect denotes repeated action. It is attested e.g. in Hungarian. Iteratives also exist in Czech with this name but their meaning is rather habitual. They can be formed only from imperfective verbs and they are usually not classified as a separate aspect; they are considered imperfective.

(227) Hungarian: *üt* "hit" ⇒ ***ütöget*** "hit several times"

4.17 Voice

Voice is typically a feature of verbs. It may also occur with other parts of speech (nouns, adjectives, adverbs), depending on whether borderline word forms such as gerunds and participles are classified as verbs or as the other category.

For Indo-European speakers, voice means mainly the active-passive distinction. In other languages, other shades of verb meaning are categorized as voice. Their common denominator is that by transforming the verb into another voice, its valency frame is altered. For example, passivization transforms a transitive verb into intransitive by demoting the original subject (recoding it as an oblique argument or removing it completely), and promoting the original object into the subject position.

English passive voice is constructed completely periphrastically. Neither the auxiliary verb nor the participle is passive in isolation, hence there is no reason to distinguish the active and passive voices as morphological features. Czech passive voice is also periphrastic but the participle is specifically passive and the voice feature can distinguish it from other participles, which are active. And in Swedish, even finite verbs have passive forms.

Identifying grammatical relations such as subject and object with semantic roles such as agent and patient is not always reliable, yet the prototypical interpretation of the **active** voice is that the subject of the verb is the doer of the action (agent) and the object is the undergoer affected by the action (patient).

(228) Czech: ***Napadli*** *jsme nepřítele.* "We **attacked** the enemy." (the active participle *napadli* can be used to form either past tense or conditional mood; here it forms the past tense)

(229) Swedish: *Hunden* ***jagade*** *katten.* "The dog chased the cat."

In contrast, in the **passive** voice the subject of the verb is the undergoer affected by the action (patient). The doer (agent) is either unexpressed or it appears as an oblique argument of the verb.

(230) Czech: *Jsme* ***napadeni*** *nepřítelem.* "We are **attacked** by the enemy." (the passive participle *napadeni* is used to form the passive in all tenses; here it forms the present passive)

(231) Swedish: *Katten* ***jagades*** *av hunden.* "The cat was chased by the dog."

While English distinguishes active and passive voice, Ancient Greek distinguishes active and **middle** voice. As the name suggests, its meaning is halfway between active and what is normally understood as passive. The subject is both doer and undergoer in a sense: he is acting upon himself. Some intransitive verbs that are active in English actually only occur in the middle voice in Greek. For example, *to go* and *to come* are actions that always affect the person who initiates them.[4]

(232) Ancient Greek: *λύει τὸν ἵππον μου (luei ton hippon mou)* "he frees my horse" (active)

[4] Examples from `https://ancientgreek.pressbooks.com/chapter/21/`.

(233) Ancient Greek: *λύομαι τὸν ἵππον (luomai ton hippon)* "I free (my own) horse" (middle)

Some grammatical descriptions define the **reflexive** voice, which is very similar in nature to the middle voice. Reflexivity is a reference from a non-subject argument (or adjunct) in a clause to the subject of the clause. In English it is achieved by a reflexive pronoun: *John saw himself in the mirror.* In other languages, the form of the verb may indicate that the object will not be expressed overtly because it is identical with the subject. In Slavic languages, the clitic form of the reflexive pronoun has taken a number of additional functions, some of which are not unlike the function of the middle voice in Ancient Greek. In East Slavic languages, the clitic has become a suffix of the verb *(-ся, -сь / -sja, -s')* and it could be regarded as a middle voice form.

(234) Russian: *Я **вернy** эти книги в библиотеку. (Ja **vernu** èti knigi v biblioteku.)* "I will return the books to the library." (active)

(235) Russian: *Я **вернусь** в библиотеку. (Ja **vernus'** v biblioteku.)* "I will return to the library." (middle/reflexive)

The Turkish grammar distinguishes a **reciprocal** voice (the *-iş* morpheme), indicating that the two arguments are both actors and patients at the same time, each acting upon the other. Note that in Russian the reflexive form of selected verbs can convey reciprocal meaning as well and typically the same label (middle voice) will be used for both functions.

(236) Turkish: *Barış Filiz'i öptü.* "Barış kissed Filiz." (active)

(237) Turkish: *Filiz ve Barış **öpüştüler.*** "Filiz and Barış kissed." (reciprocal)

(238) Russian: *Иван поцеловал Надю. (Ivan poceloval Nadju.)* "Ivan kissed Nadja." (active)

(239) Russian: *Иван и Надя поцеловались. (Ivan i Nadja pocelovalis'.)* "Ivan and Nadja kissed." (middle with reciprocal meaning)

Some ergative-absolutive languages (see Section 4.5 for the description of ergative and absolutive), have the **antipassive** voice, which is both similar and different from the passive in nominative-accusative languages. Example is Yidiny, a Pama-Nyungan language of Australia (Andrews, 2007, p. 193–197), (Dixon, 1977). When the active verb form is replaced by the antipassive, the former object is transformed to the dative or locative case, hence it is no longer a direct object and becomes and oblique argument. The subject is still subject, but now of an intransitive clause. The purpose of the transformation is that certain grammatical constructions require the target argument to be either a subject of an intransitive verb, or an object, but not a subject of a transitive verb; after the transformation, our subject can be used in such constructions. It is analogous to the passive in English, which is sometimes needed

because certain constructions target the subject only (as with the control verb in *I want her to be admired by everyone*).

(240) Yidiny: *Waguɖaŋgu guda:ga* ***wawa:l***. "The man saw the dog." (active)

(241) Yidiny: *Wagu:ɖa gudaganda* ***wawa:ɖiɲu***. "The man saw the dog." (antipassive)

In a large group of Austronesian languages (e.g. Indonesian), the voice system is characterized as symmetric. When the verb is transformed to what could be considered a passive voice, its valency is not reduced and the former agent is not removed or demoted to an oblique position. Both the agent and the patient stay in the sentence, just their marking and roles are swapped. The reasons that lead to choosing one or the other voice may be pragmatic, distinguishing between topic and focus of the sentence. Therefore the literature sometimes avoids using the terms active and passive in these languages and speaks about the **actor voice** and **undergoer voice,** respectively. A selected noun phrase, called pivot, is marked in a unique way (by word order, function words or morphology), which may have some pragmatic meaning such as "this is the topic". The voice of the verb then indicates whether the pivot is to be interpreted as the actor or the undergoer.

An extreme version of a pragmatic voice system has developed in the Philippine-type languages. Noun phrases playing many different semantic roles can be turned into pivots (topic), and verbs may have up to seven different voices (Sylak-Glassman, 2016, p. 58) to account for that. The following examples are from Tagalog (Andrews, 2007, p. 203); in this language, the pivot is preceded by the phrase marker *ang*.

(242) Tagalog: ***Magaalis*** *ang babae ng bigas sa sako para sa bata.* "The woman will take some rice out of a/the sack for a/the child." (agent voice)

(243) Tagalog: ***Aalisin*** *ng babae ang bigas sa sako para sa bata.* "A/the woman will take the rice out of a/the sack for a/the child." (patient voice)

(244) Tagalog: ***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." (locative voice)

(245) Tagalog: ***Ipagaalis*** *ng babae ng bigas sa sako ang bata.* "A/the woman will take some rice out of a/the sack for the child." (benefactive voice)

Another type of a symmetric voice system is attested in the Algonquian languages of North America. These languages work with a salience hierarchy where the more salient, higher types of arguments are by default expected to be actors / doers, and the less salient arguments are expected to be patients / undergoers. For example, second person arguments may be rated higher than third person arguments, animate nouns are higher than inanimates. Third-person arguments may be morphologically marked as proximate or obviative (see Section 4.9) in order to distinguish arguments that would be otherwise at the same level.

In this system, the default voice of the verb is called **direct.** If we want to say that the argument that is lower in the hierarchy is actually the actor, we must put the verb in the other voice, which is called **inverse.** The following examples are from Plains Cree, an Algonquian language of Canada (Wolvengrey, 2011). The person and number of both arguments is cross-referenced through the verbal morphology, hence no overt noun phrases are required.

(246) Plains Cree: *Niwīcih**ā**nānak.* "We help them." (direct)

(247) Plains Cree: *Niwīcih**iko**nānak.* "They help us." (inverse)

The **causative** voice adds a new argument (causer) and transforms an intransitive clause into a transitive, or a transitive clause into a ditransitive. In English, causative constructions are periphrastic, as in *She made him clean up.* In other languages however, the causative is a morphological transformation of the verb. In Basque, an active transitive verb is replaced by its causative form, the original subject is transformed to a dative object and denotes the causee. A new ergative subject is inserted and denotes the causer (Zúñiga and Fernández, 2014).

(248) Basque: *Arazo hau **ikusi** genuen.* "We have seen this problem." (active)

(249) Basque: *Arazo hau **ikuserazi** digute.* "They have made us see this problem." (causative)

4.18 Mood

Mood is a feature that expresses modality and subclassifies finite verb forms. Similarly to the case of nouns, there is a wide variety of meanings that can be encoded as moods, although every language has only a small subset of them.

The **indicative** can be considered the default mood. A verb in indicative merely states that something happens, has happened or will happen, without adding any attitude of the speaker.

(250) Czech: ***Studuješ** na univerzitě.* "You **study** at the university."

(251) German: *Du **studierst** an der Universität.* "You **study** at the university."

(252) French: *Tu le **fais.*** "You **do** it."

(253) Turkish: *Eve **gidiyor.*** "She **is going** home."

(254) Turkish: *Eve **gitti.*** "She **went** home."

(255) Estonian: *Sa ei **tule.*** "You are not coming."

(256) Albanian: *Ti **flet** shqip.* "You speak Albanian."

The **imperative** is used by the speaker to order or ask the addressee to do the action of the verb.

(257) Czech: ***Studuj*** *na univerzitě!* "**Study** at the university!"

(258) German: ***Studiere*** *an der Universität!* "**Study** at the university!"

(259) Turkish: *Eve* ***git!*** "**Go** home!"

(260) Turkish: *Eve* ***gidin!*** "**Go** home!" (plural)

(261) Turkish: *Eve* ***gitsin!*** "[Let him] **go** home!" (3rd person imperative)

(262) Sanskrit: ब्रूहि राजः *(**brūhi** rājaḥ)* "**Tell** the king"

The **conditional** mood is used to express actions that would have taken place under some circumstances but they actually did not / do not happen. Grammars of some languages may classify conditional as tense (rather than mood) but e.g. in Czech it combines with two different tenses (past and present), and examples (264) to (267) show four tense-aspect combinations with the conditional in Turkish. The English conditional is constructed periphrastically with the help of the modal auxiliaries *would, could, should, might*; if anything should be tagged with the feature in English, then these four words. Similarly the Slavic conditional (263) is constructed using special forms of the auxiliary "to be", in some languages reduced to a single form *(by)*. Like in English, only the auxiliary is conditional-specific and deserves to be annotated as such; the participle with which it combines is more prototypically associated with the past tense indicative.

(263) Czech: *Kdy**bych** byl chytrý, studoval **bych** na univerzitě.* "If I were smart I **would** study at the university."

(264) Turkish: *eve* ***gittiyse*** "if she went home"

(265) Turkish: *eve* ***gidiyorsa*** "if she is going home"

(266) Turkish: *eve* ***giderse*** "if she goes home"

(267) Turkish: *eve* ***gidecekdiyse*** "if she was going to go home"

The **potential** mood indicates that the action of the verb is possible but not certain. This mood corresponds to the modal verbs *can, might, be able to* and is used e.g. in Finnish or Turkish. See also the optative below.

(268) Turkish: *Eve* ***gidebilir.*** "She can go home."

(269) Turkish: *Eve* ***gidemeyebilir.*** "She may not be able to go home."

The **subjunctive** (also called **conjunctive**) mood is used under certain circumstances in subordinate clauses or reported speech, typically for actions that are subjective or otherwise uncertain. In German, it may be also used to convey the conditional meaning: in (272), the past subjunctive *(Konjunktiv II)* of *to be* is used in the adverbial clause both in German and in the English translation; in the main clause, the subjunctive of the auxiliary *werden* "to become" is used to construct the periphrastic conditional.

(270) French: *Je veux que tu le **fasses.*** (lit. *I want that you it do.*) "I want you to do it."

(271) German: *Die Mutter **sei** wieder dabeigewesen.* (lit. *the mother is.SUB again present.been*) "The mother was reportedly present again." (subjunctive I or present subjunctive)

(272) German: *Wenn ich klug **wäre, würde** ich an der Universität studieren.* "If I **were** smart I would study at the university." (subjunctive II or past subjunctive)

The **jussive** mood expresses the desire that the action happens; it is thus close to both imperative and optative. Unlike in desiderative, it is the speaker, not the subject who wishes that it happens. It is also used for negative commands. The jussive is attested in Arabic. Sanskrit has a similar mood that is called **injunctive** there.

(273) Sanskrit: मैवं वोचः *(maivam **vocaḥ**)* "Do not **speak** this way"

The **purposive** mood occurs in Amazonian languages and its meaning can be paraphrased by English "in order to".

The **quotative** mood is used e.g. in Estonian to denote reported speech.

(274) Estonian: *Sa ei **tulevat.*** "You are reportedly not coming."

The **optative** mood expresses exclamations like "May you have a long life!" or "If only I were rich!" In Turkish it also expresses suggestions. In Sanskrit it may express possibility (cf. the potential mood in other languages).

(275) Turkish: *Eve **gidelim.*** "Let's go home."

(276) Sanskrit: अप्रधानः प्रधानः स्यात् *(apradhānaḥ pradhānaḥ **syāt**)* "the unimportant person **may be (become)** important"

The **desiderative** mood corresponds to the modal verb *to want to*. Unlike the jussive, it is the subject (actor) and not the speaker who wishes that the action happens.

(277) Japanese: 食べたい *(**tabetai**)* "want to eat"

The **necessitative** mood expresses necessity and corresponds to the modal verbs *must, should, have to*:

(278) Turkish: *Eve **gitmeli.*** "She should go home."

(279) Turkish: *Eve **gitmeliydi.*** "She should have gone home."

The **admirative** mood expresses surprise, irony or doubt. Occurs in Albanian, other Balkan languages, and in Caddo (a Native American language from Oklahoma). In some contexts, it can be translated using English "apparently".

(280) Albanian: *Ti **fliske** shqip!* "You (surprisingly) speak Albanian!"

4.19 Evidentiality

Evidentiality (Aikhenvald, 2004) is the morphological marking of a speaker's source of information. It is sometimes viewed as a category of mood and modality.

Many different values are attested in the world's languages. Perhaps the most common distinction is whether the information conveyed is **firsthand** or **non-firsthand.** For instance, Turkish has two past tenses that differ in this feature. The normal past tense (also definite past tense, seen past tense) is used to convey firsthand information, something that the speaker has personally witnessed. In contrast, the so-called *miş*-past (non-firsthand, renarrative, indefinite, heard past tense) is used for information that the speaker acquired indirectly.

(281) Turkish: *Eve* ***geldi.*** "She came home." (and I was there and saw her coming)

(282) Turkish: *Eve* ***gelmiş.*** "She came home." (I did not witness her coming but I know it because someone told me or because I see that she is there now)

Aikhenvald also distinguishes **reported** evidentiality, occurring in Estonian and Latvian, among others. We have listed it as the quotative mood (Section 4.18).

Chapter 5

Dependency Trees

The syntactic structure of a sentence can be annotated and visualized in various ways, depending on the underlying theory and available tools. However, most syntactic frameworks make use of a data structure called **tree.** We will be focusing on one particular type of trees, namely rooted directed dependency trees, whose usage in linguistics dates back to the seminal work of (Tesnière, 1959).

We have seen a few trees in the previous chapters, without a formal definition. A tree consists of two types of elements. The first are *nodes,* also called *vertices;* in our situation, nodes (mostly) correspond to words or tokens. Nodes are connected by *edges* (also called *arcs, relations* or *dependencies*); this is the second type of elements. Edges typically have labels on them that further specify the type of the relation. One of the nodes is designated as the *root.* Edges are directed, i.e., they can be depicted as arrows showing the direction on the path from the root to the outer nodes. Every node has exactly one *incoming edge* (except for the root, which has none); the number of *outgoing edges* is not limited. In consequence, there is always just one path from the root to any node. A node that has no outgoing edges is called *leaf.* Sometimes the node at the origin of an edge is called *parent, governor* or *head* and the node where the edge ends is called *child* or *dependent.*

Trees can be depicted either two-dimensionally, where the x-axis corresponds to the word order and the y-axis defines layers of nodes that have the same distance from the root; or linearly, where the sentence is written on one line and edges are shown as arcs that connect the words. The latter style is used throughout this book. For instance, Figure 5.1 shows a sentence with two alternative trees: one above the words, the other beneath them.

There are many different ways how to encode the syntax of a sentence in a dependency tree. Like with morphological tags, we will try to show the existing approaches to annotation of various syntactic constructions and discuss their advantages and drawbacks. We will especially compare two influential and quite different annota-

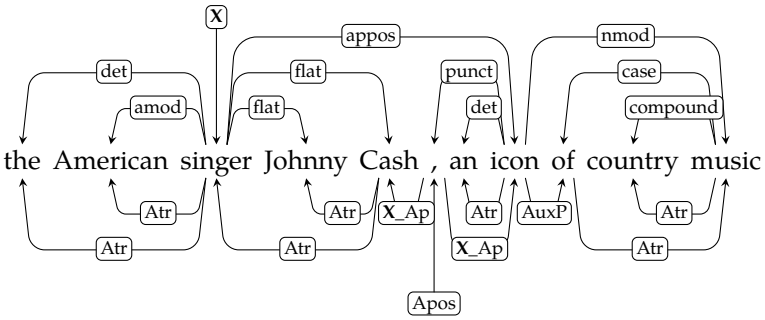


Figure 5.1: An English noun phrase in the UD (above) and PD (below) styles.

tion styles: the Prague style, and Universal Dependencies. The former is grounded in the Functional Generative Description (FGD) (Sgall et al., 1986) and tries to apply the theory to real data; the latter aspires to be as theory-neutral as possible, while highlighting parallelism between different languages. In addition, we will draw upon various other treebanks when applicable. For instance, we will occasionally refer to the "Paninian style" (Bharati et al., 2006a), which is an annotation scheme based on the traditional Indian linguistic school founded by Pāṇini, so far applied to several Indo-Aryan and Dravidian languages of the Indian subcontinent.

The Prague annotation style is exemplified in the Prague family of treebanks, which include the Prague Dependency Treebank of Czech (Hajič et al., 2000), the Prague Czech-English Dependency Treebank (Hajič et al., 2011) or the Prague Arabic Dependency Treebank (Smrž et al., 2008). It has been applied to a number of other languages such as Slovak (Šimková and Garabík, 2006), Slovenian (Džeroski et al., 2006), Croatian (Tadić, 2007), Lithuanian (Bielinskienė et al., 2016), Latin (Passarotti and Dell'Orletta, 2010), Ancient Greek (Bamman and Crane, 2011), Modern Greek (Prokopidis et al., 2005) and Tamil (Ramasamy and Žabokrtský, 2012). The Prague style is also one of the two annotation styles used in HamleDT, a collection of 30 treebanks with harmonized annotation (Zeman et al., 2014). Some of the Prague-style treebanks have two layers of syntactic annotation, which are called analytical (or surface-syntactic) and tectogrammatical (deep-syntactic). Unless explicitly noted otherwise, we will be looking at the analytical layer. We will refer to the Prague-style trees as **PD** (Prague Dependencies).

Universal Dependencies (Nivre et al., 2016)[1] is a community effort to define an annotation style applicable to all natural languages, and to provide annotated data in as many languages as possible. The syntactic annotation evolved from another popular framework, the Stanford Dependencies **(SD)** (de Marneffe et al., 2014). So

[1] http://universaldependencies.org/

far it has been applied to over 70 languages from 17 language families. We refer to Universal Dependencies as **UD.**

We only discuss dependency relation labels where it makes a difference, and we do not provide detailed description of all the labels that appear in our examples. However, in order to illustrate the different granularity, and also because we have many examples from PD and UD, we show the complete label sets of these two frameworks in Tables 5.1 and 5.2, respectively. The UD labels are meant to fit all languages, possibly with the help of additional subtypes of the universally defined relations. UD also tries to fit all domains, including spontaneous speech. The PD labels we show were defined specifically for the original Czech treebank, while some PD-style treebanks for other languages define a few extra labels. Unlike UD, PD does not distinguish whether the dependent is a clause, a noun phrase or another modifier word.

Here are some objectives that lead to different annotation styles:

- To model how the language system works. More specifically, 'constituents' stick together and the direction of an edge corresponds to syntactic government, which can be defined in terms of *reduction analysis* (Lopatková et al., 2005). UD is known for *not* following this principle of government in direction of edges (but it usually preserves the necessary information in the labels of the edges).
- To maximize parallelism across languages (as in UD). This means that content words are put closer to the root than function words, whenever possible.
- To model semantics and help language-understanding applications.
- To provide structure that can be easily modeled by machine-learning algorithms, that is, to maximize computational "learnability" (Rosa, 2015; Silveira and Manning, 2015; Schwartz et al., 2012).

5.1 Simple Noun Phrases

The noun phrase is the main building block of a sentence. It is normally headed by a noun or a pronoun, which may be optionally modified by an adjective, a numeral, a determiner, another noun phrase, or a combination of the above. A special case of a noun phrase is the adpositional phrase, which contains an adposition. For the moment we leave open the question whether the noun or the adposition should be the head of an adpositional phrase.

Figure 5.1 illustrates a number of modifier types that can occur in a noun phrase, and their alternative annotations in UD and PD. One striking difference is that UD treats apposition as a dependency of the second constituent on the first one *(singer → icon)* while in PD both constituents appear on the same level and the comma is (ab)used as a neutral head. This approach is obviously inspired by the PD treatment of coordination and we will discuss its pros and cons when we discuss coordination (Section 5.7).

Another difference is in labels. In PD, modifiers of nouns are simply `Atr` (attribute). UD has a more detailed relation taxonomy, with labels partially differentiated by the

Adv	adverbial modifier (adverb, adpositional phrase or clause)
AdvAtr	adverbial or attribute, ambiguity not semantic
Apos	head node of apposition (typically function word or punctuation)
Atr	attribute of a nominal (det., adjective, numeral, NP or clause)
AtrAdv	attribute or adverbial, ambiguity not semantic
AtrAtr	attribute, can depend on any preceding nominal
AtrObj	attribute or object, ambiguity not semantic
Atv	verbal attribute attached to a nominal
AtvV	verbal attribute attached to a verb
AuxC	subordinating conjunction
AuxG	unspecified punctuation
AuxK	sentence-terminating punctuation
AuxO	semantically redundant particle (often a pronoun)
AuxP	adposition
AuxR	reflexive marker of middle/passive and impersonal constructions
AuxS	artificial root node
AuxT	reflexive marker of an inherently reflexive verb
AuxV	auxiliary verb
AuxX	comma
AuxY	other conjunction, adverb or particle
AuxZ	adverbial emphasizing a noun phrase
Coord	head node of coordination (typically conjunction or punctuation)
ExD	ex-dependent; real parent has been elided
Obj	object (valency-licensed noun phrase, infinitive or clause)
ObjAtr	object or attribute, ambiguity not semantic
Pnom	non-verbal predicate attached to copula
Pred	predicate (root verb of the sentence or of a parenthetical)
Sb	subject (noun phrase, infinitive or clause)

Table 5.1: Dependency types ('analytical functions') of the Prague Dependency Treebank.

`acl`	adnominal clause
`advcl`	adverbial clause
`advmod`	adverbial modifier (adverb)
`amod`	adnominal modifier (adjective)
`appos`	apposition
`aux`	auxiliary verb or particle
`case`	case marker
`cc`	coordinating conjunction
`ccomp`	complement clause
`clf`	classifier
`compound`	dependent part of compound
`conj`	non-first conjunct
`cop`	copula
`csubj`	subject clause
`dep`	unknown dependency
`det`	determiner
`discourse`	discourse particle
`dislocated`	dislocated nominal
`expl`	expletive
`fixed`	part of a fixed functional expression
`flat`	flat constituent with external head
`goeswith`	non-first part of broken word
`iobj`	indirect object
`list`	list item
`mark`	subordinating marker
`nmod`	nominal modifying another nominal
`nsubj`	nominal subject
`nummod`	definite cardinal numeral
`obj`	object (direct)
`obl`	oblique argument or adjunct
`orphan`	orphan after an elided head node
`parataxis`	loosely connected phrase
`punct`	punctuation
`reparandum`	reparation in spontaneous speech
`root`	independent node attached to the artificial root
`vocative`	vocative nominal
`xcomp`	controlled complement clause

Table 5.2: Universal dependency relation types in UD v2 guidelines. There are also subtypes (including language-specific) such as `acl:relcl` or `obl:arg`.

part of speech of the dependent: `det` for determiners, `amod` for adjectives, `nummod` for cardinal numerals and `nmod` for nested noun phrases. UD also distinguishes compounds *(country ← music)* from general noun-noun modification *(icon → of music)*.

Finally, UD uses the `flat` relation to signal that in a sequence of nouns that together denote an entity, none of them can be said to clearly head the others. If any of them is selected and the others removed, the sentence stays grammatical. This is typically the case with names, titles and occupations, as in *singer Johnny Cash*. UD always selects the first node in a flat cluster as the artificial head but the label `flat` says that this is not a dependency in the usual linguistic sense. In contrast, the PD guidelines say that titles and occupations depend on names, and given names depend on family names. This rule is obviously semantic rather than syntactic (note that it does not mean that the dependency goes always right-to-left: the guidelines specifically warn the annotators that in Chinese names such as *Deng Xiaoping*, the first syllable is the family name and the rest is a given name).

Apart from these differences, most dependency treebanks (including those that are neither PD nor UD) share the view that nouns are heads and determiners, adjectives or genitives are their dependents. There is at least one exception though: the Danish Dependency Treebank (Kromann et al., 2004). In DDT, noun phrases seem to be inside out: nouns often depend on determiners, numerals or on genitive / possessive modifiers (see Figure 5.2). This approach has actually some support among linguists; but even (Hudson, 2004), one of its strongest advocates in English dependency grammar, admits that there is evidence for both the determiner and the noun being the head. In the context of natural language processing, and especially multilingual natural language processing, putting determiners to head positions does not seem very advantageous. Noun phrases in some languages have more than one 'determiner'; in other languages, no determiner is needed at all (even English allows determiner-less noun phrases). For applications like syntactic and semantic parsing, it would be beneficial to see a direct link between the lexical noun and its lexical head, such as a verb whose valency the noun satisfies. In the specific implementation taken in DDT, even the link between a noun and its modifying adjectives is obscured, as the adjectives are also attached to the determiner. Finally, I would argue (although it is hard to prove) that it is more intuitive to view *this book* as a specific instance of *book*, rather than a specific instance of *this*.

Some nouns have arguments whose form and semantic role they define; this is especially true of deverbal nouns. For instance, the English verb *to deny* takes an object *(he denied all accusations)*, and the derived noun *denial* inherits the valency, although the form of the argument changes and requires the preposition *of* when used with the noun *(his denial of the accusations)*. Most treebanks agree that complementation of nouns is different from complementation of verbs and should not be modeled using the same type of relation. In PD and UD, the *accusations* from the above example will be attached to *denial* as `Atr` and `nmod` respectively, not as an 'object'. Nevertheless, there are treebanks that treat arguments of nouns the same way as arguments

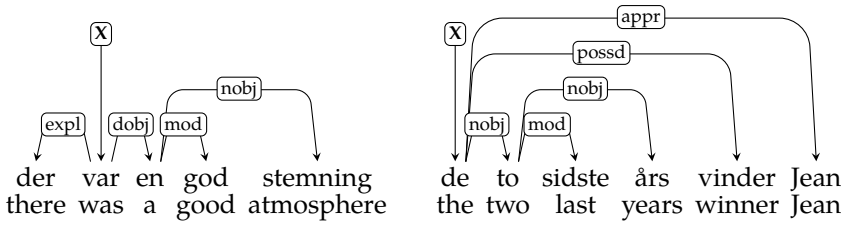


Figure 5.2: Two fragments from the Danish Dependency Treebank show how determiners, numerals and genitives / possessives govern noun phrases.

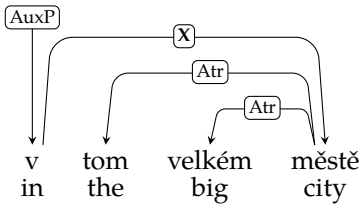


Figure 5.3: Prepositional phrase in PDT (Czech).

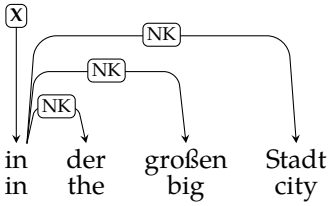


Figure 5.4: Prepositional phrase in TIGER (German).

of verbs. One example is the Polish Dependency Treebank, Składnica zależnościowa (Wróblewska and Woliński, 2012). In this treebank, the attachment of optional modifiers is labeled `adjunct` regardless whether they are adjectives under nouns or adverbials under verbs. Analogically, valency-conditioned arguments are labeled `subj`, `obj`, `obj_th` and `comp` regardless whether their parent is a verb or not.

Adpositions (Figures 5.3–5.5) can either govern their noun phrase or they can be attached to the head of the NP. Similarly to determiners, it is not completely clear which approach is better. The adposition typically cannot stay in the sentence with-

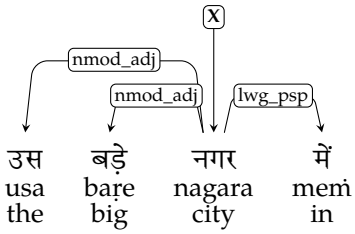


Figure 5.5: Postpositional phrase in the Hindi Treebank.

out the noun. The noun may be able to exist independently (unless it is in a form that is used exclusively with adpositions) but probably not in the context of the given sentence. In some languages, the adposition governs the case of the noun, which supports the analysis that the noun is an argument of the adposition and fills it valency. Such analysis is in favor of making the adposition the head. PD, SD and most other pre-UD treebanks had adpositions as heads (a notable exception is the Paninian framework of the Hindi treebank, shown in Figure 5.5). UD turned them to leaves with the aim of improving parallelism across languages. What is encoded as an adposition in one language may be just a case affix in another; that justifies viewing adpositions as features of the nouns. As a result, the direction of edges in UD does not always correspond to the linguistic notion of government and subordination. This is an aspect of UD which we will observe in several other parts of the grammar; it may complicate certain aspects of linguistic research, such as figuring out whether a language is prevailingly head-initial or head-final. Nevertheless, it is not a fatal problem because relations in UD are labeled with types. Not the direction of the edge alone, but the direction together with the label inform us about the nature of the relation and about the consequences for syntactic dominance.

When the adposition heads the NP, other modifiers of the main noun are usually attached to the noun but they can also be attached to the preposition, which is the case in the TIGER Treebank (Brants et al., 2002) (Figure 5.4). The label of the relation of the adpositional phrase to its parent (we label it **X** in the example trees) can sometimes be found at the adposition, while the relation between the adposition and the noun is labeled as a 'prepositional argument' or 'object'. Elsewhere the adposition, despite serving as head, gets an auxiliary label (such as `AuxP` in PD) and the real label is found at the noun. This can be justified in linguistic and semantic terms (the noun bears the largest portion of the meaning of the NP), but it complicates computational processing because types of outgoing relations of a parent node cannot be collected simply by examining all its child nodes.

5.2 Quantifiers and Classifiers

English cardinal numerals are similar to adjectives. They modify the noun and may or may not be accompanied by a determiner *((the) **two** monkeys)*. Indefinite quantifiers like *few* and *many* are themselves determiners and do not tolerate other determiners in the same NP. In either case, English quantifiers are typically analyzed as dependents of nouns. This is not necessarily the case in other languages.

In Czech, only the numerals *jeden* "one" to *čtyři* "four" can be compared to adjectives. They agree with the counted noun in morphological case (*one* and *two* agree also in gender and number). From *pět* "five" up and for all pronominal quantifiers, the pattern changes and depends on the required case of the entire quantified phrase (quantifier + NP; let us denote it QP). This external case follows from the function of the QP in the sentence, for example, from a verb's valency. If the required case is genitive, dative, locative or instrumental, both the numeral and the NP take the same case form and we can still view the numeral as a modifier. However, if the QP is in nominative, accusative or vocative, only the numeral takes the form of the external case,[2] while the NP is forced to the genitive. In these cases, the QP is syntactically very similar to nouns with genitive modifiers that have a partitive interpretation: cf. *dvacet mužů* "twenty [of] men" vs. *skupina mužů* "group of men". In addition, if the QP is in subject position, the verb no longer agrees in gender and number with the noun phrase; instead, it switches to the default form, which is neuter singular. The numeral can thus be said to govern the noun because it governs its case and it prevents the noun from being cross-referenced by the verb. Czech PD closely follows the government rules and makes the numeral the head of the QP when the external case is nominative, accusative or vocative; the noun is the head otherwise. UD attaches all numerals as dependents of counted nouns and defines a language-specific relation `nummod:gov` that marks places where the numeral is the governor (Zeman, 2015). This relation is currently used in 7 Slavic languages and Sanskrit (see Figure 5.6 for a Russian example). It is a rare situation where UD prefers semantic parallelism over syntactic evidence; in other difficult constructions, UD often gives precedence to syntax. Both solutions have their up- and downsides. For semantically oriented applications such as relation extraction, it is probably advantageous to have the noun higher in the tree and closer to the verb. Whether it also helps syntactic parsers is questionable: on the one hand a UD parser does not have to learn the intricacies of finding out when the numeral behaves differently, on the other hand it will now see subjects and direct objects in the genitive case (rather than nominative resp. accusative).[3]

[2] In fact, high-value numerals show heavy case syncretism and have only two distinct forms: one for nominative-accusative-vocative, and one for the other four cases.

[3] There is a third possible solution: to make the numeral always the head. The counted noun will either agree with it, or will be assigned the genitive case. Such approach could work well in Czech but it would break the cross-lingual parallelism in UD.

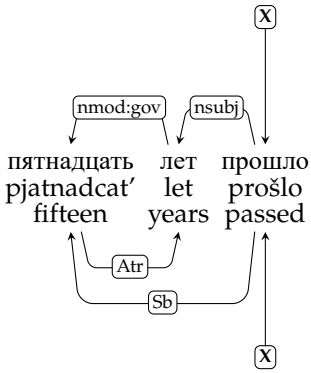


Figure 5.6: A Russian quantified phrase in UD (above) and PD (below).

A marginal question related to quantified phrases is what to do with compound numerals such as *five thousand three hundred sixty-seven*. The UD guidelines propose a `flat` analysis similar to personal names in Figure 5.1. In PD, the component numerals would be connected via `Atr`. Another marginal question is whether numbers should be attached to nouns the same way even if they indicate neither quantity *(**seven** windows)* nor rank *(the **seventh** window; Louis **XIV**)* but just a version number in a product name *(Windows **7**)*. Some treebanks distinguish all the three functions of numbers but most ignore the different status of the last one, merging it either with ranks or with quantities. Within UD, only Russian currently uses a relation subtype called `nummod:entity`. Note that in PD the difference would not lead to a different label, as both quantities and sub-names are labeled `Atr`.

Some languages, especially in Asia, require numerals to be accompanied by classifiers—noun-like words that semantically correlate with the counted noun. We introduced classifiers in Section 3.5.1 and illustrated them on an example from Chinese; we show a UD tree of the example in Figure 5.7. Chinese treebanks generally agree on joining first the classifier with the numeral, then the numeral with the counted noun. Note that the UD guidelines specifically say that this analysis is not intended for languages like English, where classifier structures are not highly grammaticalized, although there are somewhat similar constructions with measure words and units, such as *three cups of coffee*. Such constructions receive a traditional analysis in these languages, that is, *cups* will be treated as an ordinary plural noun, and *of coffee* as its genitive modifier.

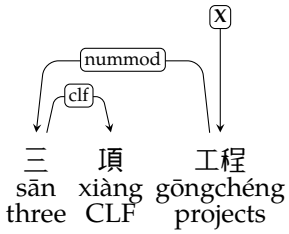


Figure 5.7: A Chinese phrase with classifier in UD.

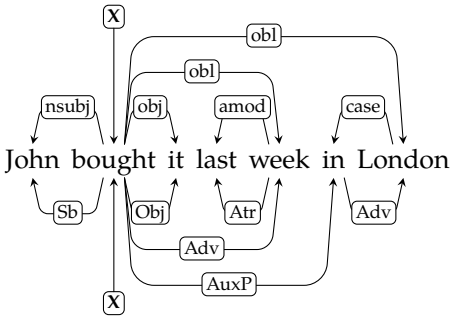


Figure 5.8: A simple English transitive clause in UD (above) and PD (below).

5.3 Simple Clauses

A clause is typically built around a verb (though we will discuss non-verbal clauses later). The verb may have one or more **arguments** and there may be some optional **adjuncts** too. Prototypical arguments are noun phrases; adjuncts are either noun phrases (often adpositional) or adverbs. Both types of dependents can be also realized as subordinate clauses but we will ignore them for the moment.

Figure 5.8 illustrates a simple transitive clause with two adjuncts and its alternative annotations in UD and PD. The two trees are largely parallel, except for the reversed internal structure of the adpositional phrase *in London*. Following the European grammatical tradition, adjuncts in PD are labeled as *adverbial modifiers*, regardless whether they are realized as adverbs or noun phrases. In UD, they are labeled as *oblique dependents*.

What the example does not reveal is the different borderline in the two styles between objects and other dependents. An object in PD is essentially any non-subject argument defined by the verb's valency frame. If it is a noun phrase (with or without

adposition), its form (such as adposition or morphological case) is determined by the verb, and if the verb also assigns it a semantic role, then it is an object. This, too, is a traditional approach in many European grammars. An object in PD thus matches the usual definition of argument (e.g., (Andrews, 2007)) with the additional constraint that it is not subject, and the `Obj-Adv` distinction in PD is in fact the argument-adjunct distinction. Various other non-UD treebanks take similar approaches, although their label sets may be more fine-grained and distinguish the accusative (direct) object, the dative (indirect) object, objects in other cases and prepositional objects. (Wróblewska and Woliński, 2012) go a step further in Polish and also distinguish adverbials required by the verb (such as the target location in *to put something somewhere*; labeled `comp`) from free adjuncts (labeled `adjunct`).

Unfortunately, distinguishing arguments from adjuncts is not always trivial. One has to determine the valency frame of every verb sense consistently. Even then the line may be fuzzy for certain verbs and argument types. For example, the Prague Dependency Treebank labels the bare dative noun phrase in Czech as an object if it denotes the recipient role of *to give* or a similar verb; but if the role is "just" a beneficiary, it is considered an adjunct that can appear with most verbs and determines its own semantics autonomously. But which datives are recipients and which are beneficiaries? Apparently it was not easy for the annotators, and there are many verbs whose dative dependents appear with both labels: in PDT 3.0, *koupit někomu něco* "to buy [for] somebody something" has 40 × `Obj` and 10 × `Adv`; *otevřít někomu něco* "to open something [for] somebody" has 13 × `Obj` and 11 × `Adv`. Datives of verbs with negative effects seem to be more likely to be understood as "beneficiaries": *zlomit někomu něco* "to break somebody's something" has 13 × `Adv` and no `Obj`.

The difficulty of defining and recognizing the borderline between arguments and adjuncts led to the rejection of this distinction in Universal Dependencies. Instead, the main dividing line in UD goes between **core arguments** and **oblique dependents,** which is a common category for non-core arguments and adjuncts. UD treebanks can still choose to distinguish oblique arguments from adjuncts, but this distinction is secondary and optional: the relation subtype `obl:arg` is used for oblique arguments while adjuncts use plain `obl` (Figure 5.9).

Recognizing core arguments is supposed to be easier than recognizing arguments in general, as it can usually be based on surface features such as word order, adpositions, morphological case and agreement with the verb. The exact signs are specific for each language and cannot be defined universally. Therefore the difficult part is establishing what exactly should count as a core argument in a particular language. (Andrews, 2007) suggests looking at primary transitive verbs such as *to kill somebody* and describe the coding strategy and behavior of the argument that has the "P" function, i.e., the argument that is more patient-like. Then one should look at other two-argument verbs and decide whether their arguments are core; if they are, the verb will be considered *transitive* (although not necessarily primary transitive). The rule of thumb is: "If an NP is serving as an argument of a two-argument verb, and receiving

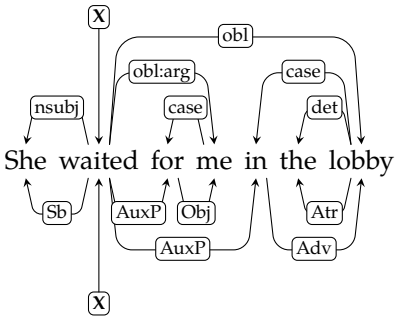


Figure 5.9: English: UD can optionally distinguish oblique arguments from adjuncts using `obl:arg`. In most UD treebanks both use just `obl`. In PD, oblique arguments are objects and adjuncts are adverbials.

a morphological and syntactic treatment normally accorded to the patient of a primary transitive verb, we shall say that it has the grammatical function P" (Andrews, 2007, p. 138). In the UD context it directly follows that it will be labeled as the direct object, `obj`. A potential problem here is the part saying 'normally accorded', as well as determining which verbs qualify as primary transitive. What if there are a few two-argument verbs that could possibly pass for primary transitives but their patient-like argument has different properties than the mainstream, such as a different morphological case? Should we say that there are two different 'normal treatments' in the language that both denote the function P? Or should we require that the other case is frequent enough to be considered 'normal'? And how much is enough? Traditional grammar will not help, core argument is a term from language typology. It is thus not surprising that several UD treebanks, especially those converted from other annotation styles, have changed their definitions of core arguments between UD releases, and there is still work to be done before cross-linguistic consistency is achieved.

The distinction is relatively easy in English. Core arguments must be bare noun phrases (without prepositions) and appear close to the verb. The argument that appears before the verb and agrees with the verb in person and number is subject. The argument that appears after the verb and can be promoted via passivization is object. Some care has to be taken to exclude bare nominal adjuncts such as *last week* in *He died last week*; apart from that, prepositions are the most reliable indicators of obliqueness.

Many other Indo-European languages are in a more difficult situation: they distinguish multiple morphological cases. It seems correct to say that their prepositional objects are not core arguments but it is not clear which bare NPs (in what cases) are or are not core. The literature is not unanimous either: for instance, (Andrews, 2007) discusses the possible coreness of the dative in German (without giving a definite

answer), while (Foley, 2007)'s position in the same volume is that German dative is oblique. (Zeman, 2017) examines several case-marking Indo-European languages and comes to a tentative conclusion that bare noun phrases in all cases could be treated as core arguments, provided they are really arguments, i.e., there is no evidence that they are actually adjuncts. It usually means filtering out the equivalents of the *last week*-type adjunct in English, which appears as a bare accusative in several case-marking languages. But it also means that many more borderline situations have to be considered in the other cases. We have discussed the recipient-beneficiary distinction in Czech datives and we can expect even more adjuncts when it comes to the instrumental case. More fuzzy cases lead to more annotation errors and less consistency; hence we now lean towards stipulating that only nominatives and accusatives are generally core, unless there is evidence for a specific exception in a particular language. One such exception could be the genitive case of quantified subject phrases (cf. Section 5.2). Some authors also propose 'logical dative subjects' of certain predicates, such as experiencing verbs (Czech *líbí se* ***mi*** *to* "it is pleasing **to me**") but these phrases are not really syntactic subjects, and they may not even be core objects in the light of what we just said about the dative. On the other hand, (Andrews, 2007) discusses dative subjects in Icelandic and demonstrates that these are really syntactic subjects, triggering verb agreement and other rules that normally target subjects.

Head-marking languages like Basque cross-reference more than one argument in verbal morphology (cf. Section 4.14). Cross-referenced arguments are always core—therefore, Basque has three core morphological cases: absolutive, ergative and dative. The two arguments in Basque transitive clauses are either ergative + absolutive, or dative + absolutive, or ergative + dative (Zúñiga and Fernández, 2014). Whenever there is absolutive, it denotes the patient; whenever there is ergative, it denotes the actor; and the dative is either actor or patient depending on the other argument's case. UD annotation will then mark the actor as the subject and the patient as the object. Intransitive clauses have a single argument that is always subject, even though different verbs will force it to different morphological forms, viz. absolutive, ergative or dative. In contrast, ditransitive clauses have all three arguments and here the ergative is subject, absolutive is object and dative is indirect object (Figure 5.10).

The Philippine languages are famous for controversies about whether they have an argument that could be termed 'subject'. One candidate is the topic-marked pivot (cf. Section 4.17), another possibility is to identify the subject with the semantic actor. Neither of them is a perfect match for the tests of subjecthood that work in other languages (see (Andrews, 2007, p. 202–211) for a detailed discussion). In respect to annotation consistency it seems easier to pick the pivot as the subject, considering that it is conveniently marked by a specific function word, such as *ang* in Tagalog. Any other core arguments (marked with *ng* in Tagalog) are objects. Such an approach will also mean that the subject has the semantic role identified as topic by the verb voice. So in the benefactive voice the subject is the beneficiary; this is an analogy to the passive of Indo-European languages, where the subject is the patient. Indeed, this

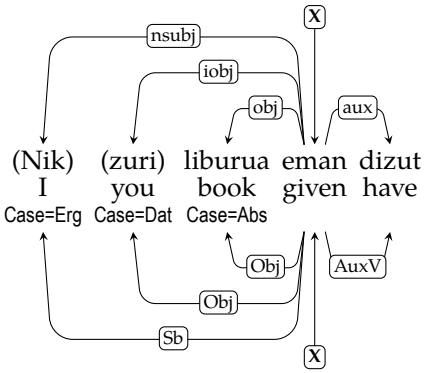


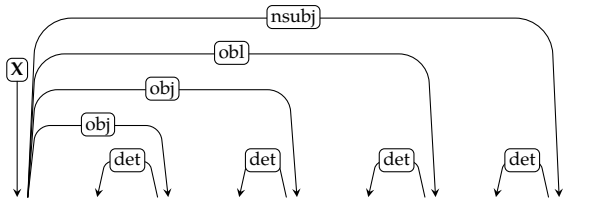
Figure 5.10: A Basque ditransitive clause in UD (above) and PD (below). The 1st- and 2nd-person pronouns can be omitted because the information about the person and number of these arguments is already conveyed by the auxiliary verb.

is the approach taken in the Tagalog dataset in UD 2.2, the only existing treebank of a Philippine language we are aware of. Figure 5.11 shows the UD annotation of the benefactive voice example (245).

A similar question about subjecthood arises also in the Algonquian family with its direct-inverse voice system (Section 4.17). (Wolvengrey, 2011) even advices against using the subject-object terminology, pointing out that it carries a European bias that is not suitable for Plains Cree and related languages. Nevertheless, for comparative projects like Universal Dependencies it is useful to simplify the situation and reuse the traditional labels, with the necessary grain of salt: labels are used cross-linguistically, but their exact definition is always language-particular. With that in mind, and in an analogy to the other languages, we can assign the UD `nsubj` relation to the NP that is more salient according to the language-internal hierarchy. As a result, the more actor-like argument will be the subject if the verb is in direct voice, and the more patient-like argument will be subject in the inverse voice. Note that unlike in some other languages, it is easy to decide whether an argument is core. Plains Cree is a head-marking language and core arguments are those that are cross-referenced by the verb.

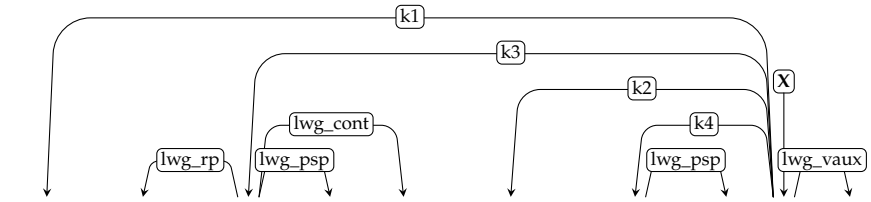
Let us now turn to yet another possible approach to classifying nominal arguments, which is different from both UD and PD:[4] The Paninian syntax (PanD) that is the basis for annotation in some treebanks of Indian languages (Bharati et al., 2006a). This framework defines so-called **karaka relations,** which lie half-way between syn-

[4] It is different from the analytical layer of the Prague Dependencies, but it is relatively close to the tectogrammatical layer.



"A/the woman will take some rice out of a/the sack for the child."

Figure 5.11: The Tagalog clause in the benefactive voice (245) in UD.



"We have only sent the matter through the President to the Commission."

Figure 5.12: A Hindi sentence with the first four karaka relations.

tax and semantics. On the one hand, they do not distinguish subject and object in the usual sense, as they rather operate along the actor-patient axis and abstract from different realizations of the roles in active and passive clauses. On the other hand, karaka relations are very coarse-grained and do not directly correspond to semantic roles. For example, the relation `k1` *karta* is the most independent participant in the event and it often corresponds to the actor, but there are clauses in which the karta is not what other theories would want to describe as actor. Thus in (283) the karta is *the boy*; but it is *the key* in (284) and *the lock* in (285).

(283) *The boy opened the lock.*

(284) *The key opened the lock.*

(285) *The lock opened.*

k1	karta	doer / agent / subject
k2	karma	patient / object
k3	karana	instrument
k4	sampradaana	recipient / beneficiary
k5	apaadaana	source
k7	adhikarana	location in space or time

Table 5.3: The six karaka relations of the Paninian syntax.

Table 5.3 briefly introduces the six main karaka relations and their labels in PanD treebanks. Figure 5.12 shows an example sentence from the Hindi treebank with four karakas.

5.4 Verb Groups

Various sorts of verbal groups include periphrastic verb forms (such as auxiliary + participle), modal verbs with infinitives and similar constructions. Dependency relations, both internal (between group elements) and external (either incoming from the parent of the group or outgoing to the modifiers), may be defined according to various criteria: content verb vs. auxiliary, finite form vs. infinitive, subject-verb agreement (typically holds for finite verbs and participles but not for infinitives). Participles often govern auxiliaries as in the PD-style Slovenian Dependency Treebank (Džeroski et al., 2006). That is also true in UD—not because they are participles but because they are content verbs, which are always placed higher than function words in UD (Figure 5.13). Elsewhere the finite verb is the head as in the Russian Dependency Treebank (Boguslavsky et al., 2000) (Figure 5.14) or both approaches are possible based on semantic criteria, as in the Prague Dependency Treebank. In the treebanks of Hindi (Husain et al., 2010) and Tamil (Ramasamy and Žabokrtský, 2012), the content verb (which could be a participle or a bare verb stem) is the head and auxiliaries (finite or participles) are attached to it.

The head typically bears the label describing the relation of the group to its parent. As for child nodes, subject and negative particle (if any) are often attached to the head, especially if it is the finite element while the arguments (objects) are attached to the content element whose valency slot they fill (often participle or infinitive). Sometimes even the subject or the negative particle can be attached to the non-head content element, as in the Alpino treebank of Dutch (van der Beek et al., 2002) (Figure 5.15).

Various infinitive-marking particles (English *to*, Swedish *att*, Bulgarian *да*) can be treated similarly to subordinating conjunctions, i.e., they can govern the infinitive as in the Danish (Kromann et al., 2004) and Bulgarian (Simov and Osenova, 2005) treebanks, or be attached as children of the infinitive, as in German (Brants et al., 2002) or

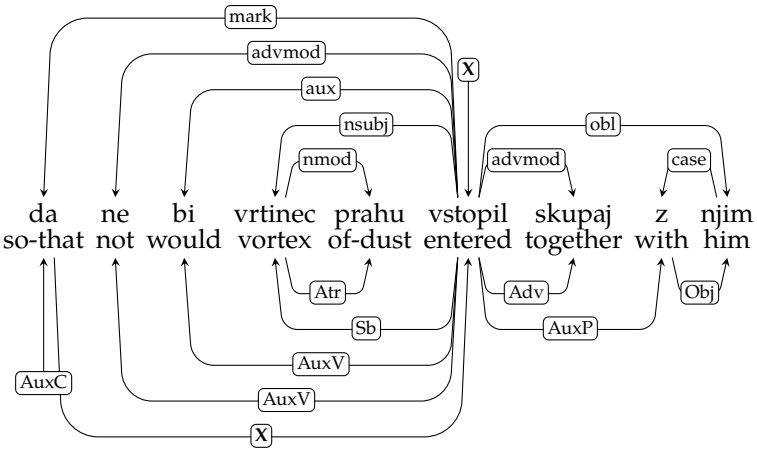


Figure 5.13: UD-style (above) and PD-style (below) negated conditional construction in Slovenian, meaning "so that the vortex of dust would not enter together with him." Past participle of content verb *(vstopil)* is the head in both styles; the negative particle *(ne)* and the auxiliary *(bi)* depend on it.

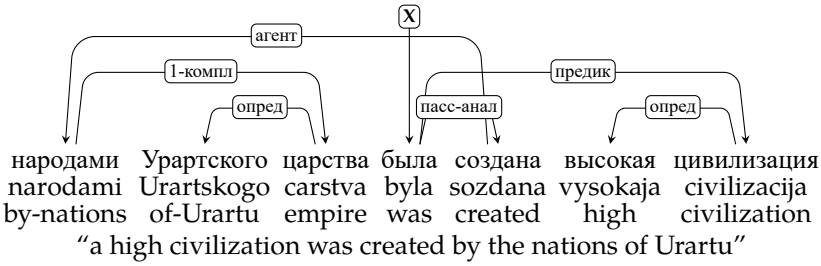


Figure 5.14: Passive construction in Russian: Finite auxiliary verb *(была)* is the head, passive participle *(создана)* depends on it. As a result, the agent *(народами)* is attached non-projectively to the participle *(создана)*.

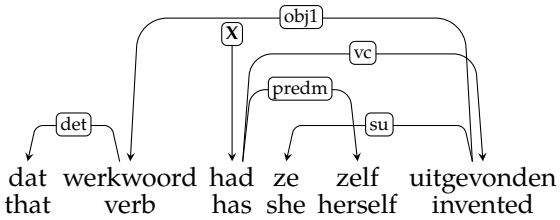


Figure 5.15: A Dutch example from the Alpino treebank. Unlike in other treebanks, even the subject *(ze)* is attached to the non-head participle *(uitgevonden)*.

Swedish (Einarsson, 1976). In the Bosque treebank of Portuguese (Afonso et al., 2002), prepositions used between the main verb and the infinitive (*estão a usufruir* "are enjoying") are attached to the finite verb (Figure 5.16). In Bulgarian, all modifiers of the verb including the subject are attached to *да* instead of the verb below (Figure 5.17).

Modal verbs (*can, must, should* etc.) are often represented as head nodes that govern infinitives of lexical verbs (in other words: their valency requires an infinitive argument). Figure 5.18 shows a German example (from TIGER treebank) annotated this way. Languages often have other verbs that subcategorize for an infinitive, although they are not classified as modal. Verbs of control *(he wants to do it)* and raising *(he seems to be doing it)* belong here.

Some modals in some languages (including English) can be subsumed under auxiliary verbs. It makes sense to view modal verbs as function words; after all, some languages use morphological mood inflection to achieve the same meaning shift. However, it is difficult to delimit them only semantically. English has tests that can reveal whether a verb is auxiliary; for instance, they do not need the auxiliary *do* when they are negated (see Section 3.5.2 and (Huddleston and Pullum, 2002, p. 92)). Such tests may not be available in other languages or they may lead to different sets of verbs. Hence German *wollen* "to want" is traditionally considered modal, while English *to want* is a verb of control but it is neither modal nor auxiliary. All these nuances impact the UD annotation style, where multiple auxiliaries are attached to the same lexical verb (Figure 5.19). Cross-linguistic parallelism is achieved only for verbs that are auxiliary in both languages. English UD thus differs from Slavic languages, where modals are not treated as auxiliaries. Furthermore, modality can be expressed by words that are not verbs at all. Russian regularly uses modal adjectives and even in English we can encounter them:

(286) Russian: *Мне нужно учиться. (Mne nužno učit'sja.)* (lit. *to-me necessary to-study*) "I have to study."

(287) English: *It is possible to do it better.*

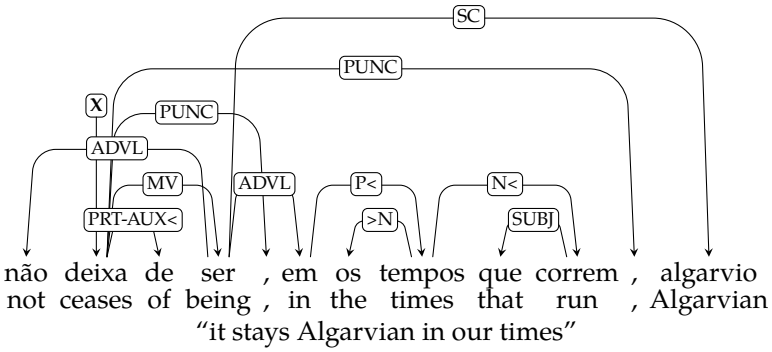


Figure 5.16: Infinitive with preposition in Portuguese. Unlike prepositions that govern noun phrases, *(de)* is not attached between the phase verb *(deixa)* and the infinitive *(ser)*. The negative particle *(não)* is attached non-projectively to the non-head verb *(ser)*. Moreover, the commas around the parenthesis *(em os tempos que correm)* are also non-projective.

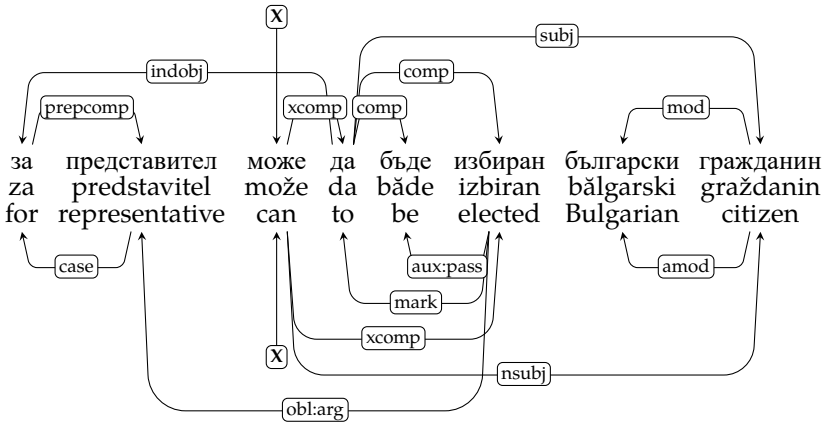


Figure 5.17: Modal passive construction in Bulgarian, meaning “only a Bulgarian citizen can be elected representative.” Above the sentence is the original BulTreeBank annotation while the tree below illustrates the UD style. Finite modal verb *(може)* is the head, infinitive particle *(да)* is the second-level head. Infinitive auxiliary *(бъде)* is attached to *да*, as is the passive participle of the content verb *(избиран)* and the two arguments of the content verb—one of them *(за представител)* non-projectively.

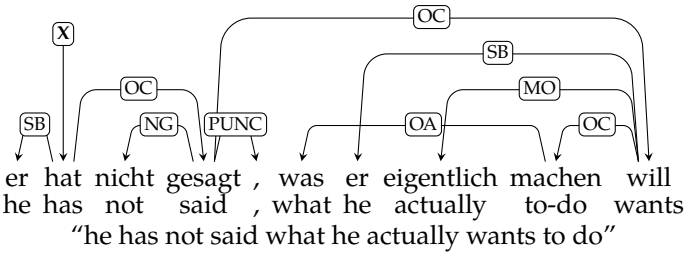


Figure 5.18: Combination of perfect tense, modal verb and infinitive in German. Infinitives are attached to modals as their objects in many treebanks. Finite auxiliary verb *(hat)* is the head of the perfect tense, the participle *(gesagt)* depends on it. Subject *(er)* is attached to the finite verb *(hat)* while the object clause *(was er eigentlich machen will)* is attached to the content verb *(gesagt)*.

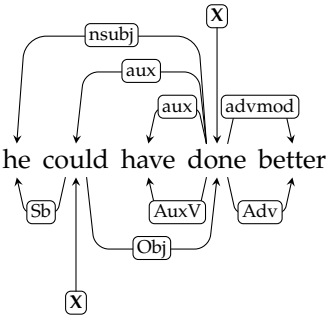


Figure 5.19: English modal auxiliary in UD (above) and PD (below).

(288) English: *I am able to do it better.*

(289) English: *I can do it better.*

The UD trees of the above examples will thus be quite different, as only in (289) the modal element is an auxiliary verb.

Preserving the syntactic hierarchy of modal and (other) auxiliary verbs that appear in one group may also be useful for modeling scope: in English *you must not do it*, the verb *do* is first negated, then the negative complex becomes an argument of *must*. In contrast, the apparent German counterpart *du musst es nicht machen* has a different meaning: the modal verb *müssen* is negated, yielding "you don't have to"; then the affirmative *machen* "to do" becomes the argument of the negated modality ("you don't have to do it").

To summarize, there is a large number of annotation approaches that different treebanks take to groups of verbs. While none of them is perfect, we can still come up with some recommendations:

1. Attach subject as high as possible and let the lower verbs inherit it. If the finite verb is higher, we also get the subject-verb agreement directly represented. If the lexical verb is higher, we get better representation of meaning and potentially better cross-linguistic parallelism. Unfortunately, it is a common pattern that the finite verb is auxiliary and the lexical verb is nonfinite, which puts these two objectives in conflict.
2. If an argument (object) fits in the valency frame of the lower verb but not of the higher verb, attach it to the lower verb.
3. Varying attachment of free modifiers (adjuncts) may lead to different interpretation: e.g. *[he persuaded me at the meeting] to do it* has not necessarily the same meaning as *he persuaded me [to do it at the meeting]*. If there is no evidence that a particular attachment is preferred, attach the modifier to the highest verb possible.
4. Preserving hierarchy of the verbs in the group may be useful for modeling scope. However, it would mean that either we cannot attach all auxiliaries to the same lexical verb, or we have to exclude modal verbs from the set of auxiliaries.

5.5 Clauses with Non-Verbal Predicates

Adjectives, nouns or even adpositional phrases may occasionally act as predicates instead of verbs. In English they are accompanied by a **copula**—the auxiliary verb *to be* that supplies the clause with verbal features such as tense, aspect and mood. In Russian, the copula is used only in the past and future tenses, while lack of copula is interpreted as the present tense. Other languages (e.g., Polish) may use a pronoun as the copula (*Krystyna **to** Marty koleżanka* "Krystyna is Marta's colleague"), despite the fact that it does not provide any verbal features, it just links the subject with the predicate.

Figure 5.20 illustrates the annotation of nonverbal clauses in UD and PD. The original PD for Czech assumes that copula is mostly present, and because it is a verb, it is made the head of the sentence. The nominal part of the predicate (noun or adjective) is attached as a child of the copula via a `Pnom` relation. PD consideres the second sentence an existential statement with a locative modification, therefore the `Pnom` relation is not used there; in contrast, UD allows locations and other adverbials to be nonverbal predicates. The 'copula-head' approach of PD leads to difficulties in languages that omit the copula. Figure 5.21 shows a PD tree of an Arabic non-verbal sentence: a new label, `PredP`, had to be introduced so that the preposition could be analyzed as the head of the predicate, and thus of the entire sentence.

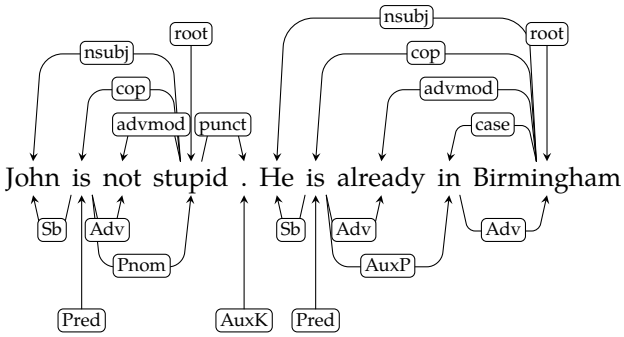


Figure 5.20: English: Two nonverbal clauses in UD (above) and PD (below).

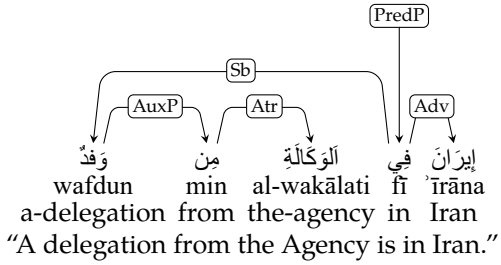


Figure 5.21: An Arabic non-verbal clause from the Prague Arabic Dependency Treebank (PD style). Words are organized left-to-right, contrary to the Arabic orthography.

UD takes a different approach. It always attaches the nonverbal predicate higher. The copula, if present, is attached as a child of the nonverbal part. This way the Russian present tense (Figure 5.22) receives a tree that is similar to the past tense (Figure 5.23) and also to English. On the other hand, trees with copula are no longer parallel to trees with a verb that functions almost like a copula but it also adds another shade of meaning, hence it should be treated as a lexical verb. In UD-style annotation of (290), *teacher* will be the root. In (291), the root will be *became*.

(290) English: *Ivan is a teacher.*

(291) English: *Ivan became a teacher.*

Also note that UD annotation of English nonverbal clauses (292) is different from existential clauses (293). Even though the same verb *to be* is used in both constructions,

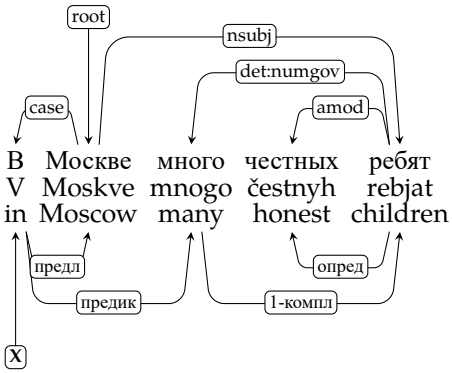


Figure 5.22: A Russian non-verbal clause from SynTagRus meaning "There are many honest children in Moscow." The original annotation is shown below the sentence, UD is above.

the syntactic structure is different and the verb is not treated as a copula; instead, it is made the head of the clause. In contrast, Czech versions of the two constructions are syntactically identical; note that the word order in Czech is flexible, driven only pragmatically. Both (294) and (295) can be analyzed as copular sentences. Only in (296) there is no location that could serve as the predicate, so the verb *je* must be promoted as the head. Finally, in languages like German (297, 298), existential constructions use a different verb and there is no need to make them parallel with copulas.

(292) English: *The food is in the refrigerator.*

(293) English: *There is food in the refrigerator.*

(294) Czech: *Jídlo je v ledničce.* "The/some food is in the refrigerator."

(295) Czech: *V ledničce je jídlo.* "There is food in the refrigerator." or also: "The stuff that is in the refrigerator is food."

(296) Czech: *Jídlo je.* (lit. *Food is.*) "There is food. Food exists."

(297) German: *Das Essen ist im Kühlschrank.* "The food is in the refrigerator."

(298) German: *Es gibt Essen im Kühlschrank.* (lit. *It gives food in-the refrigerator.*) "There is food in the refrigerator."

If the nonverbal predicate is a noun (and not an adjective), it may be difficult to decide which noun is the subject and which is the predicate. In the rigid word order of English, it seems natural to say that the noun before the copula is subject. Hence *Ivan* is the subject in (299) but predicate in (300). Czech has a free word order but sometimes the issue is resolved by case marking: the predicative noun may take the instrumental

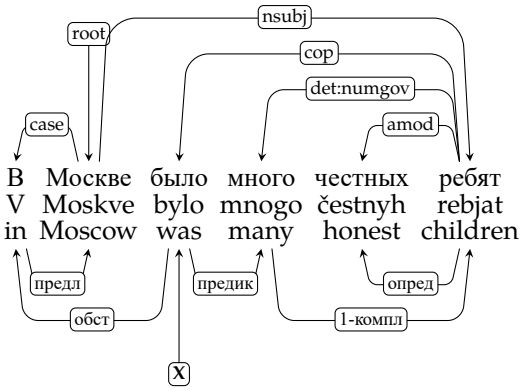


Figure 5.23: A Russian non-verbal clause with copula meaning "There were many honest children in Moscow." The SynTagRus annotation is shown below the sentence, UD is above.

form while the subject is in the nominative (301). Unfortunately, the instrumental is not required and the predicate can be nominative as well. Then only semantic or pragmatic factors can help the annotator, for example, the more descriptive nominal is the predicate (*dítě* "child" in (302)) while the more identifying nominal (pronoun or proper noun) is labeled as the subject (*bratr Pavel* "brother Pavel").

(299) English: *Ivan is the best dancer.*

(300) English: *The best dancer is Ivan.*

(301) Czech: *Ona dárcem být nemůže.* (lit. *she.NOM donor.INS to-be cannot*) "She cannot be the donor."

(302) Czech: *To dítě byl můj bratr Pavel.* (lit. *the.NOM child.NOM was my.NOM brother.NOM Pavel.NOM*) "My brother Pavel was the child."

Besides adjectives and nouns, the nonverbal predicate can be also a nested clause. Even if the nested clause is headed by a verb, it is a nonverbal predicate (and it needs a copula) because the predicate is in fact the entire clause, not just its head verb. Sometimes we can say that the nested clause functions as a subject rather than a predicate (following the arguments discussed above), as in (303). However, this solution is out of reach if both the subject and the predicate are nested clauses, as in (304), where two copular clauses are joined by another copula. Making a verb the root of both the inner and the outer clause may result in two subjects being attached to it, which would make the tree structure quite confusing. Here even the UD guidelines give up and rule that the main copula should be made the head. See Figure 5.24 for the dependency tree of (304).

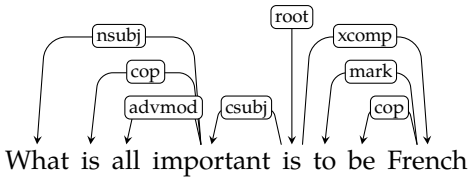


Figure 5.24: An English copular sentence with nested copular clauses as subject and predicate.

(303) Czech: *Problém je, že zmizel.* (lit. *problem is that he-disappeared*) "The problem is that he disappeared."

(304) English: *What is all important is to be French.*

To summarize, copulas and nonverbal predicates represent a difficult construction and there is no single good solution to their annotation. If the annotation targets only languages that always use an overt verbal copula, it is probably a good idea to make it the head of the clause. However, if the language sometimes uses a copula and sometimes not, or if cross-linguistic parallelism is desirable, every solution has unfortunate consequences in certain situations. That includes the UD v2 approach: it is clearly aimed at improving parallelism across languages but it stays questionable whether the positives actually outweigh the negatives.

5.6 Subordinate Clauses

Various types of constituents can be realized as subordinate clauses, finite or nonfinite. Nominals can be modified by adnominal (attributive) clauses. In (305), the relative clause *that I have ever tasted* modifies the noun *curry*. In (306), the infinitive *to say* modifies *things*. In (307), *nice and hot* is considered a secondary predication about "them", which at least in UD means that it is treated as a reduced clause: the primary predicate is "they were delivered", the secondary predicate is "they were nice and hot" but it is analyzed as a modification of the nominal ("they [who were] nice and hot").

Clauses that replace core arguments are either subjects, as *whether that unity endures* in (308), or objects, as *where he was* in (309). Special type of clausal complements are controlled clauses, whose subject is coreferential with either subject or object of the matrix clause, as *feel good* in (310). Adverbial clauses describe various circumstances of the matrix predicate: *when we get back* in (311) is a temporal adverbial. Finally, the clause *on how Europe responds* in (308) corresponds to an oblique argument; therefore UD will treat it similarly to adjuncts (adverbial clauses), while PD will label it as an object clause.

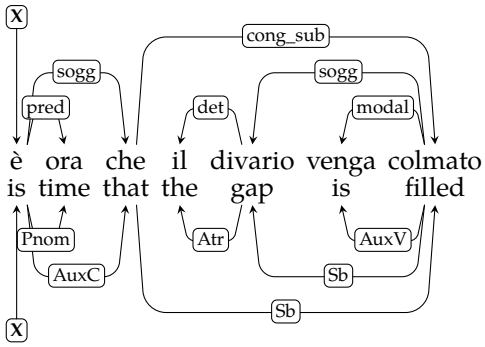


Figure 5.25: Subject subordinate clause in Italian, meaning "it is time to fill the gap". The original annotation is shown above the sentence, a PD equivalent below.

(305) English: *Best yellow curry that I have ever tasted.*

(306) English: *I have fantastic things to say.*

(307) English: *They were delivered nice and hot.*

(308) English: *Whether that unity endures will depend on how Europe responds.*

(309) English: *He knew where he was.*

(310) English: *It made me feel good.*

(311) English: *You can buy me dinner when we get back.*

Roots (predicates) of adnominal clauses are always attached to the noun they modify, hence in (305) we have either *curry* → *tasted* or *curry* → *have*, depending on what is the head of the subordinate clause in the particular annotation style. If there is a relative pronoun, it is considered a part of the subordinate clause. It represents the modified noun there and it is not uncommon that it fills an argument slot of the subordinate verb (*tasted* → *that*).

Finite complement and adverbial clauses often use a subordinating conjunction or complementizer; similarly, infinitive clauses may take an infinitive marker or adposition. Note that the conjunction should not be confused with a relative adverb or pronoun. In an analogy to adpositional phrases, some treebanks make the subordinating marker the head of the subordinate clause, attach the main predicate as its child, and the predicate of the matrix clause becomes its parent. In PD the analogy includes labeling the attachment of the marker with just an auxiliary label (`AuxC`) and putting the real function of the subordinate clause (`Sb`, `Obj` or `Adv`) to the verb. Other styles, e.g. the Italian treebank (Montemagni et al., 2003), reveal the function right away on the marker (Figure 5.25). Another common approach is that the predicate

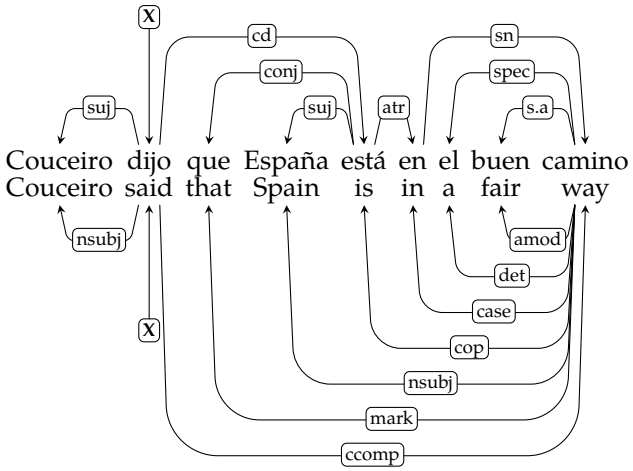


Figure 5.26: Complement clause in Spanish. The original annotation is shown above the sentence, a UD equivalent below.

of the nested clause is directly attached to the predicate of the matrix clause. It may be the only approach in cases where there is no subordinator (309). Other treebanks take this approach systematically regardless whether there is a subordinator or not; if it is there, it is attached to the head of the subordinate clause. The AnCora treebanks of Catalan and Spanish (Taulé et al., 2008) can serve as an example; this approach is also adopted in UD (Figure 5.26).

A few treebanks treat relative pronouns like subordinating conjunctions, cut them off the subordinate clause (including adpositions, if any) and put them on the path between the matrix verb and the subordinate verb. The CoNLL 2006 annotation of Alpino (van der Beek et al., 2002) is an example: in (312), we get *krijg* → *wat* → *verdient*.

(312) Dutch: *dan krijg je wat je verdient* "then you get what you earn"

The subordinating conjunction can also be attached as a sibling of the relative clause, as in the Szeged Treebank (Figure 5.27).

As a conclusion of this chapter, we want to advocate for several advantages of the UD approach to subordinate clauses over the other approaches:

- A clause is represented (headed) by its predicate, just like main clauses. Its attachment to the superordinate word is labeled differently from non-clausal de-

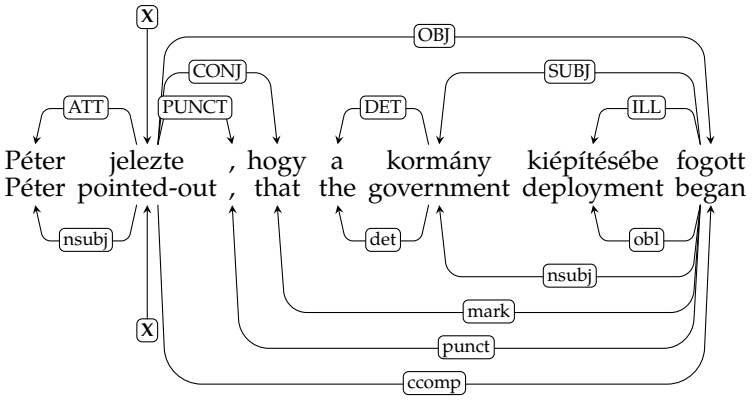


Figure 5.27: Subordinate clause in Hungarian, with the original Szeged Treebank annotation above and UD annotation below the sentence.

pendents, which makes it easier to account for different requirements placed on the dependents by the parent.[5]

- Not representing the clause by a subordinating conjunction makes the annotation more parallel to situations where the subordinator is not present.
- All words specific to the subordinate clause, including conjunctions and relative pronouns, are part of the subtree of the clause. This is especially important with relative pronouns that may fill a valency slot of the subordinate predicate.

5.7 Coordination

Coordination[6] is a construction where multiple constituents of a similar type **(conjuncts)** are put together and share one function in the sentence. There are coordinate clauses, coordinate noun phrases and even modifiers and function words. A coordinating conjunction is typically found in coordination but its presence is not required. In *asyndetic coordination*, conjuncts are simply juxtaposed (in writing they may be delimited by commas or other punctuation symbols).

Capturing coordination in a dependency framework is difficult because the conjuncts are at the same level and the relation between them is not a dependency in the usual sense. A construction whose members are equal is sometimes termed **paratactic** (while subordination is **hypotactic**).

[5] We do not claim that the concrete labels used in UD v2 are optimal. We only point out that distinguishing clausal and non-clausal dependents has some advantages.

[6] This section is based on joint work with Martin Popel, David Mareček, Jan Štěpánek and Zdeněk Žabokrtský (Popel et al., 2013).

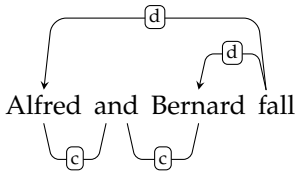


Figure 5.28: Coordinate subject and its analysis following (Tesnière, 1959), with two types of relations: subordinating dependencies `d` ('connections') and symmetric coordinating relations `c` ('junctions').

The dominating solution in treebank design is to introduce artificial rules for encoding coordinate structures (CS) within dependency trees, by the same means that expresses dependencies, i.e., by edges and their labels. Obviously, any tree-shaped representation of coordination must be perceived only as a 'shortcut', since the relations within a CS form an undirected cycle, as illustrated already by (Tesnière, 1959). For example, if a verb has two coordinate subjects, there is a (symmetric) coordination relation between the two conjuncts and two (asymmetric) dependency relations between the conjuncts and the verb (Figure 5.28).

However, as there is no obvious linguistic intuition on which tree-shaped CS encoding is better and as the degree of freedom has several dimensions, one can find a number of distinct conventions introduced in particular dependency treebanks. Variations both in topology (the shape of the tree) and in labeling are possible. Many different models have been proposed since (Tesnière, 1959), out of which the following are the most frequently used ones:

- **Mel'čuk**-style coordination as proposed in the Meaning-Text Theory (MTT). Conjuncts and conjunction(s) are connected in a chain, with the first conjunct as the head (Mel'čuk, 1988) (Figures 5.29 and 5.30). (Popel et al., 2013) refer to chain-like annotation styles as the **Moscow** family.
- **Prague** style as used in PD treebanks: a coordinating conjunction is picked as the artificial head, all conjuncts are attached as its children (Hajič et al., 2000) (Figures 5.31 and 5.32). (Popel et al., 2013) refer to annotation frameworks with this approach as the Prague family. Since they refer only to coordination topology, the family includes also treebanks whose annotation style substantially differs from PD in other aspects.
- **Stanford** style originally refers to the output of the Stanford parser, later to the Stanford Dependencies (SD) (de Marneffe et al., 2014): the first conjunct is the head and the remaining conjuncts (as well as conjunctions) are attached as its children (Figure 5.33). A slightly modified version of this style is used in **UD** (Nivre et al., 2016) (Figure 5.34). (Popel et al., 2013) refer to all approaches where one conjunct is the head and other conjuncts its children as the Stanford family.

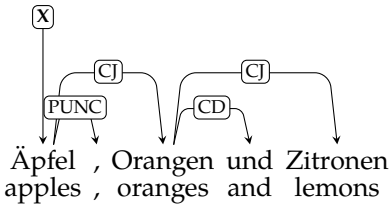


Figure 5.29: Coordination in the Tiger treebank of German (Brants et al., 2002) takes an approach inspired by Mel'čuk, except that the conjunction is not included in the chain.

One can find various arguments supporting the particular choices. MTT possesses a complex set of linguistic criteria for identifying the governor of a relation (see (Mazziotta, 2011) for an overview), leading to the Mel'čuk style. The PD approach is advocated by (Štěpánek, 2006) who claims that it can represent shared modifiers using a single additional binary attribute, while Mel'čuk would require a more complex coindexing attribute for that. An argumentation of (Tratz and Hovy, 2011) follows a similar direction: *We would like to change our [Mel'čuk] handling of coordinating conjunctions to treat the coordinating conjunction as the head [Prague] because this has fewer ambiguities…*

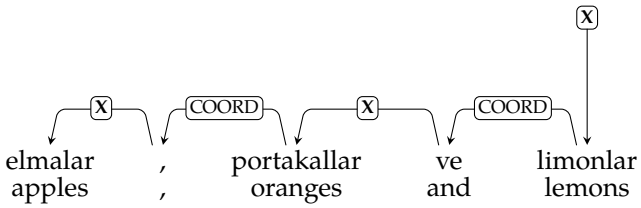


Figure 5.30: Coordination in the METU Treebank (Atalay et al., 2003) follows the Mel'čuk approach but the chain goes right-to-left because Turkish is generally a head-final language.

Formal representation of CSs is further complexified by the following facts:

- It is not uncommon that a CS has more than two conjuncts.
- Besides private dependents of individual conjuncts, there are dependents shared by all conjuncts, such as the subject in *Mary came and cried* (Figure 5.36).
- Shared modifiers can be coordinated too: *big and cheap apples and oranges*.

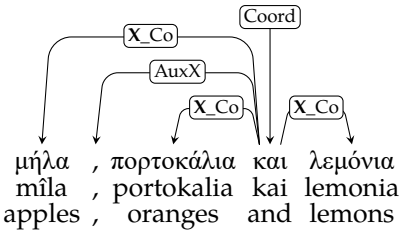


Figure 5.31: Prague-style coordination in the Greek Dependency Treebank. The type of the relation to the parent of the coordination (**X**) is indicated on the edges that connect the coordination head with the conjuncts, while the edge that connects the head with the parent is labeled just `Coord`.

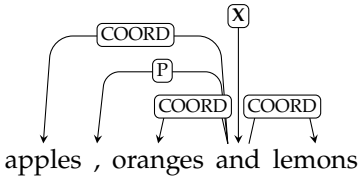


Figure 5.32: Prague-style coordination in English (Surdeanu et al., 2008). Here the relation of the CS to its parent is indicated directly on the incoming edge.

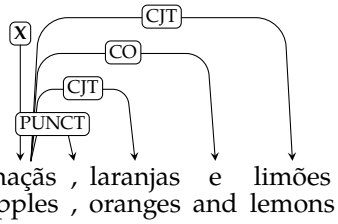


Figure 5.33: Stanford-style coordination in Portuguese (Afonso et al., 2002).

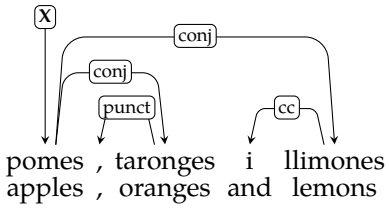


Figure 5.34: The AnCora treebanks of Catalan and Spanish (Taulé et al., 2008) originally use the Stanford style. This Catalan tree has been modified for UD.

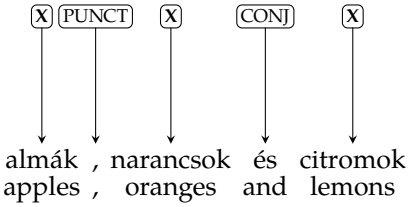


Figure 5.35: The Szeged Treebank of Hungarian (Csendes et al., 2005) gets close to the analysis proposed by Tesnière: All participating nodes are attached directly to the parent of the coordination. However, there are no 'junction' links between conjuncts.

- Nested (embedded) coordinations are possible: *John and Mary or Sam and Lisa* (Figure 5.37). A token may be conjunct or shared modifier of the outer CS, and at the same time conjunction or conjunct of the inner CS.
- Punctuation (commas, semicolons, three dots) is frequently used in CSs, mostly with multi-conjunct coordination or juxtapositions which can be interpreted as asyndetic coordination *(Don't worry, be happy!)*. In many languages, comma or other punctuation mark can play the role of the main coordinating conjunction.
- The coordinating conjunction itself can be a multiword expression *(as well as)*.
- There are deficient CSs consisting of a single conjunct.
- Abbreviations like *etc.* comprise both the conjunction and the last conjunct.
- Coordination combined with ellipsis forms an intricate structure. For example, a conjunct can be elided while its arguments remain in the sentence: *I gave the books to Mary and the records to Sue.*
- The border between paratactic and hypotactic means for expressing coordination-like semantic relations is fuzzy. Some languages can use enclitics instead of conjunctions/prepositions, e.g. Latin *Senatus Populus**que** Romanus* "The Senate **and**

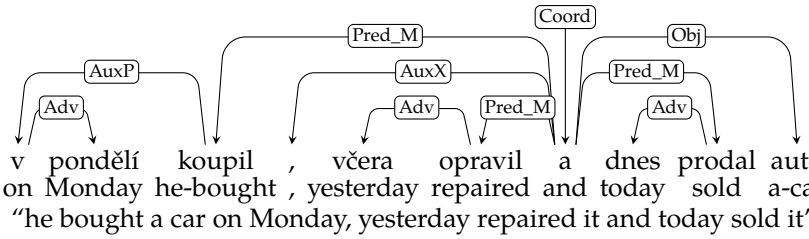


Figure 5.36: Shared and private dependents in the Prague style (Czech): *auto* "car" is the object shared by all three verbs while the adverbials ("on Monday, yesterday, today") are private. The whole structure is in predicate relation to its parent (which is probably the artificial root node): `X = Pred. _M` denotes *members* of coordination, i.e. conjuncts.

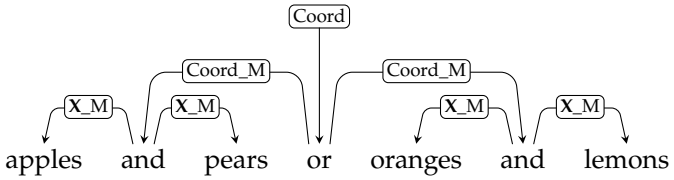


Figure 5.37: English: Nested coordination in the Prague style. `X` represents the relation of the whole structure to its parent.

People of Rome". Purely hypotactic means such as the preposition in *John with Mary* occur too.

- Careful semantic analysis of CSs discloses additional complications: if a node is modified by a CS, it might happen that it is the node itself (and not its modifiers) what should be semantically considered as a conjunct. Note the difference between *red and white wine* (which is synonymous to *red wine and white wine*) and *red and white flag of Poland*. Similarly, *five dogs and cats* has different meaning than *five dogs and five cats*.

The Prague style provides for nested coordinations, as in *apples and pears or oranges and lemons* (see Figure 5.37). The asymmetric treatment of conjuncts in the other styles makes nested coordination difficult to read, and in some situations even impossible to capture. For instance, in a left-to-right Mel'čuk style, the nested CS *A and (B or C)* cannot be distinguished from the flat three-conjunct CS *A and B and C*.

The Prague style also distinguishes between shared dependents such as the subject in *Mary came and cried*, from private dependents of individual conjuncts, as in *John*

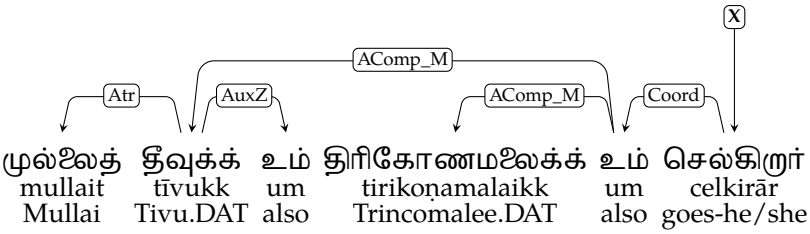


Figure 5.38: Coordination in Tamil. The morphological suffixes *um* have the coordinating function. They had to be made separate nodes during tokenization because the Tamil treebank uses the Prague style and no other coordination head was available except these morphological indicators.

came and Mary cried (see Figure 5.36). This distinction is not done in non-Prague-style treebanks. When one conjunct serves as the head, its own private dependents could be distinguished only by different labels, but we are not aware of a treebank trying to do that.

Although the Prague style seems to be the most expressive one, it has some downsides, too. First, its head node (conjunction or punctuation) is not representative of the type of the conjuncts. A CS of noun phrases will itself function as a noun phrase and could thus be expected to be headed by a noun, but there is a function word instead. Moreover, it may happen that there is no function node that could serve as the head. Fortunately, such cases are rare at least in written data, where punctuation is often available. Figure 5.38 shows an example from Tamil where punctuation is not available and the coordinating function is encoded by a clitic that is written together with the preceding conjunct. To provide the head node, the clitics had to be recognized as separate tokens.

In addition to the main family, (Popel et al., 2013) distinguish several other dimensions along which CS annotation varies:

Choice of head – leftmost or rightmost. In the Prague family, the head can be either the leftmost or the rightmost conjunction or punctuation. In the other two families, the head can be either the leftmost or the rightmost conjunct. The Persian Dependency Treebank (Rasooli et al., 2011) takes a mixed approach and picks the rightmost[7] head in coordination of verbs, leftmost otherwise.

Attachment of shared modifiers. Are they attached to the head of the CS, or to the nearest conjunct (which may or may not be the same)?

[7] Here we use the words 'left' and 'right' in the logical sense. Persian is actually written right-to-left, hence if we say 'rightmost', it will be rendered as leftmost in writing.

Attachment of coordinating conjunctions and punctuation. In the Moscow family, conjunctions can be either part of the chain of conjuncts, or they may be put aside the chain and attached to the previous or following conjunct. In the Stanford family, conjunctions can be either attached to the CS head, or they may be attached to the previous or following conjunct. The same set of options are available for punctuation, which often receives different treatment than conjunction words (in addition, a comma can be also attached to a conjunction).

Edge labels help encode information that cannot be inferred from the tree topology alone: we need to distinguish conjuncts, conjunctions, shared and private modifiers, nested CS, and the type of the relation to the parent.

The **Prague family** can be divided to the PD treebanks and the rest. PD treebanks (Figure 5.31) attach the head of the CS to its parent via a `Coord` edge, and the real function of the CS is indicated on the edges that connect the conjunct to the head. Labels of these edges are further extended with a suffix `_Co` or `_M`[8] in order to distinguish them from shared modifiers of the CS, which are also attached to the head. In the non-PD part of the Prague family, the main incoming edge indicates the relation to the parent, while the head-conjunct edges use a CS-specific label (Figure 5.32).

The PD approach is unwieldy when one wants to query functions of all children of a node, and one of the children happens to be a CS. It has one little advantage though. It can encode hybrid coordination where the relation between the parent and conjunct A requires a different label than relation between the parent and conjunct B. This may be needed in sentences like *Who and why did it?* where *who* should be labeled as subject and *why* as an adverbial modifier.

In the **Stanford** and **Moscow families,** one of the conjuncts is the head. In practice, it is never labeled as a conjunct because its conjunctness can be deduced from the fact that there are conjuncts among its children. Usually, the other conjuncts are labeled as conjuncts; conjunctions and punctuation also have special labels. A rare exception is the METU treebank (Atalay et al., 2003) where all conjuncts in the Moscow chain bear the label indicating the parent relation, while their conjunctness follows from the `COORDINATION` labels of the intervening conjunction and punctuation nodes.

To represent shared modifiers in the Stanford and Moscow families, an additional label is needed again to distinguish between private and shared modifiers, since they cannot be distinguished topologically. Moreover, if nested CSs are allowed, the label cannot be just binary (i.e. 'shared' versus 'private') because it also has to indicate what conjuncts the shared modifier belongs to.

In addition to all the options mentioned above, there are numerous proposals to capture coordination using means that go beyond a rooted dependency tree. Tesnière himself designed first such approach, as shown in Figure 5.35. Other examples would be the full annotation of the Tiger corpus (Brants et al., 2002) with its secondary

[8] Different PD treebanks use slightly different suffixes; in fact, the Prague Dependency Treebank version 3 encodes 'conjunctness' as an attribute independent from the functional label.

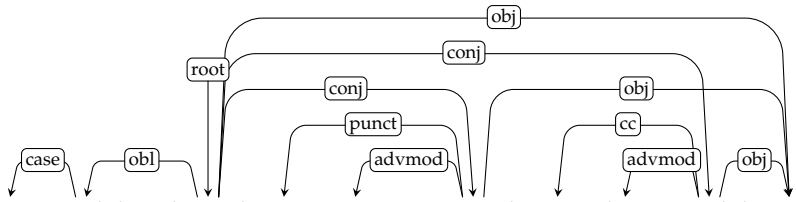


Figure 5.39: Czech: Shared and private dependents from Figure 5.36 reannotated in enhanced UD. The structure is a directed graph but not a tree because the shared object *auto* "car" has three incoming edges from the three coordinate verbs.

(tree-crossing) edges, or the 'bubble trees' suggested for the Mel'čuk style by (Kahane, 1997). In Universal Dependencies the so-called *enhanced representation* defines several situations where edges can be added or removed, yielding a general directed graph structure. One of the use cases is propagating parent-child dependencies to/from all conjuncts, reflecting the distinction between shared and private dependents (Figure 5.39). However, even the enhanced UD in its current form cannot capture nested coordination.

Chapter 6

Some Concluding Tokens

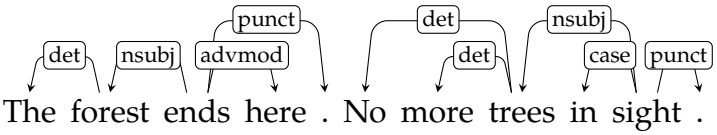
We have attempted to present a cross-linguistic survey of morphological and dependency-based syntactic annotation. We looked at

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