

Parts of Speech

Daniel Zeman

October 9, 2020



EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated

- Vague definitions, criteria of mixed nature
- **Loong tradition...** (difficult to change)
 - Traditional linguistics:
 - Classification differs cross-linguistically!
 - (Even among established classes, not just endemic minor parts of speech.)
 - Computational linguistics (tagsets):
 - Dozens of classes and subclasses
 - Significant differences even within one language

- 4th century BC: Sanskrit
- European tradition (prevailing in modern linguistics): Ancient Greek
 - Plato (4th century BC): sentence consists of nouns and verbs
 - Aristotle added “conjunctions” (included conjunctions, pronouns, and articles)
 - End of 2nd century BC: classification stabilized at 8 categories (Διονύσιος ὁ Θρᾶξ: Τέχνη Γραμματική / Dionysios o Thrax: Art of Grammar)

Ancient Greek Word Classes

- **Noun** (ὄνομα *onoma*)
 - inflected for case, signifying a concrete or abstract entity
- **Verb** (ῥῆμα *rēma*)
 - without case inflection, but inflected for tense, person and number, signifying an activity or process performed or undergone
- **Participle** (μετοχή *metochē*)
 - sharing the features of the verb and the noun
- **Interjection** (ἄρθρον *arthron*)
 - expressing emotion alone
- **Pronoun** (ἀντωνυμία *antōnymia*)
 - substitutable for a noun and marked for person
- **Preposition** (πρόθεσις *prothesis*)
 - placed before other words in composition and in syntax
- **Adverb** (ἐπίρρημα *epirrēma*)
 - without inflection, in modification or in addition to a verb
- **Conjunction** (σύνδεσμος *syndesmos*)
 - binding together the discourse and filling gaps in its interpretation

Where Are Adjectives?

- The best matching Ancient Greek definition is that of nouns, and perhaps participles.
- Adjectives are a relatively new (1767) invention from France:
 - Nicolas Beauzée: *Grammaire générale, ou exposition raisonnée des éléments nécessaires du langage*. Paris, 1767

Traditional English Parts of Speech

- 1 Noun
- 2 Verb
- 3 Adjective
- 4 Adverb
- 5 Pronoun
- 6 Preposition
- 7 Conjunction
- 8 Interjection

“Traditional” means: taught in elementary schools, marked in dictionaries.

Linguists (and especially computational linguists) may see other categories, e.g., determiners.


Traditional Czech Parts of Speech

- 1 Noun (*podstatné jméno, substantivum*)
- 2 Adjective (*přídavné jméno, adjektivum*)
- 3 Pronoun (*zájmeno*)
- 4 Numeral (*číslovka*)
- 5 Verb (*sloveso*)
- 6 Adverb (*příslovce, adverbium*)
- 7 Preposition (*předložka*)
- 8 Conjunction (*spojka*)
- 9 Particle (*částice*)
- 10 Interjection (*citoslovce*)


A Mixture of Criteria

- Parts of speech are defined on the basis of morphological, syntactic, and semantic criteria
- In many cases they are just rough approximation
- Because of long *tradition* in some languages, it is difficult to redesign the system
- Sets of POS tags strive to
 - keep reasonable consistency with tradition
 - partition the word space systematically


Morphological Criteria

- By definition language-dependent. In  Czech (simplified):
- **Nouns:** (gender), number, case. Include some pronouns (*někdo* “somebody”) and numerals (*pět* “five”, *tisíc* “thousand”, *sedmero* “seven”, *polovina* “half”)


Morphological Criteria

- By definition language-dependent. In  Czech (simplified):
- **Nouns:** (gender), number, case. Include some pronouns (*někdo* “somebody”) and numerals (*pět* “five”, *tisíc* “thousand”, *sedmero* “seven”, *polovina* “half”)
- **Adjectives:** gender, number, case, sometimes degree; agr. with N. Include some pronouns (*který* “which”, *žádný* “no”) and numerals (*první* “first”, *čtverý* “four kinds”)


Morphological Criteria

- By definition language-dependent. In  Czech (simplified):
- **Nouns:** (gender), number, case. Include some pronouns (*někdo* “somebody”) and numerals (*pět* “five”, *tisíc* “thousand”, *sedmero* “seven”, *polovina* “half”)
- **Adjectives:** gender, number, case, sometimes degree; agr. with N. Include some pronouns (*který* “which”, *žádný* “no”) and numerals (*první* “first”, *čtverý* “four kinds”)
- **Personal pronouns:** person, gender, number, case


Morphological Criteria

- By definition language-dependent. In  Czech (simplified):
- **Nouns:** (gender), number, case. Include some pronouns (*někdo* “somebody”) and numerals (*pět* “five”, *tisíc* “thousand”, *sedmero* “seven”, *polovina* “half”)
- **Adjectives:** gender, number, case, sometimes degree; agr. with N. Include some pronouns (*který* “which”, *žádný* “no”) and numerals (*první* “first”, *čtverý* “four kinds”)
- **Personal pronouns:** person, gender, number, case
- **Possessive pronouns:** possessor's person, gender, number; possessed gender and number

Morphological Criteria

- By definition language-dependent. In  Czech (simplified):
- **Nouns:** (gender), number, case. Include some pronouns (*někdo* “somebody”) and numerals (*pět* “five”, *tisíc* “thousand”, *sedmero* “seven”, *polovina* “half”)
- **Adjectives:** gender, number, case, sometimes degree; agr. with N. Include some pronouns (*který* “which”, *žádný* “no”) and numerals (*první* “first”, *čtverý* “four kinds”)
- **Personal pronouns:** person, gender, number, case
- **Possessive pronouns:** possessor’s person, gender, number; possessed gender and number
- **Verbs:**
 - infinitive
 - finite: mood (indicative/imperative), tense (present/future), person, number
 - participle: voice (active/passive), tense, gender, number, (case)
 - converb (gerund, transgressive): tense (present/past), gender, number
 - verbal noun: aspect, gender, number, case

Morphological Criteria

- By definition language-dependent. In  Czech (simplified):
 - **Nouns:** (gender), number, case. Include some pronouns (*někdo* “somebody”) and numerals (*pět* “five”, *tisíc* “thousand”, *sedmero* “seven”, *polovina* “half”)
 - **Adjectives:** gender, number, case, sometimes degree; agr. with N. Include some pronouns (*který* “which”, *žádný* “no”) and numerals (*první* “first”, *čtverý* “four kinds”)
 - **Personal pronouns:** person, gender, number, case
 - **Possessive pronouns:** possessor’s person, gender, number; possessed gender and number
 - **Verbs:**
 - infinitive
 - finite: mood (indicative/imperative), tense (present/future), person, number
 - participle: voice (active/passive), tense, gender, number, (case)
 - converb (gerund, transgressive): tense (present/past), gender, number
 - verbal noun: aspect, gender, number, case
 - **Non-inflectional words**

Syntactic / Distributional Criteria

- Slightly less language-dependent
- **Nouns:** arguments of verbs (subject, object), nominal predicate (*he is a teacher*) etc. Also attribute of other nouns. Include personal pronouns (*I, you*), some numerals in some languages.

Syntactic / Distributional Criteria

- Slightly less language-dependent
- **Nouns:** arguments of verbs (subject, object), nominal predicate (*he is a teacher*) etc. Also attribute of other nouns. Include personal pronouns (*I, you*), some numerals in some languages.
- **Adjectives:** modify noun phrases; also predicates (*he is smart*).

Syntactic / Distributional Criteria

- Slightly less language-dependent
- **Nouns:** arguments of verbs (subject, object), nominal predicate (*he is a teacher*) etc. Also attribute of other nouns. Include personal pronouns (*I, you*), some numerals in some languages.
- **Adjectives:** modify noun phrases; also predicates (*he is smart*).
- **Verbs:** predicates of clauses.

Syntactic / Distributional Criteria

- Slightly less language-dependent
- **Nouns:** arguments of verbs (subject, object), nominal predicate (*he is a teacher*) etc. Also attribute of other nouns. Include personal pronouns (*I, you*), some numerals in some languages.
- **Adjectives:** modify noun phrases; also predicates (*he is smart*).
- **Verbs:** predicates of clauses.
- **Adverbs:** modify verbs, usually as adjuncts (optional).

Syntactic / Distributional Criteria

- Slightly less language-dependent
- **Nouns:** arguments of verbs (subject, object), nominal predicate (*he is a teacher*) etc. Also attribute of other nouns. Include personal pronouns (*I, you*), some numerals in some languages.
- **Adjectives:** modify noun phrases; also predicates (*he is smart*).
- **Verbs:** predicates of clauses.
- **Adverbs:** modify verbs, usually as adjuncts (optional).
- **Prepositions:** govern noun phrases and their case, semantically modify their relation to verbs or other nouns.

Syntactic / Distributional Criteria

- Slightly less language-dependent
- **Nouns:** arguments of verbs (subject, object), nominal predicate (*he is a teacher*) etc. Also attribute of other nouns. Include personal pronouns (*I, you*), some numerals in some languages.
- **Adjectives:** modify noun phrases; also predicates (*he is smart*).
- **Verbs:** predicates of clauses.
- **Adverbs:** modify verbs, usually as adjuncts (optional).
- **Prepositions:** govern noun phrases and their case, semantically modify their relation to verbs or other nouns.
- **Coordinating conjunctions** (*and, or, but*).





Syntactic / Distributional Criteria

- Slightly less language-dependent
- **Nouns:** arguments of verbs (subject, object), nominal predicate (*he is a teacher*) etc. Also attribute of other nouns. Include personal pronouns (*I, you*), some numerals in some languages.
- **Adjectives:** modify noun phrases; also predicates (*he is smart*).
- **Verbs:** predicates of clauses.
- **Adverbs:** modify verbs, usually as adjuncts (optional).
- **Prepositions:** govern noun phrases and their case, semantically modify their relation to verbs or other nouns.
- **Coordinating conjunctions** (*and, or, but*).
- **Subordinating conjunctions** (*that*): join dependent to main clause.


Syntactic / Distributional Criteria

- Slightly less language-dependent
- **Nouns:** arguments of verbs (subject, object), nominal predicate (*he is a teacher*) etc. Also attribute of other nouns. Include personal pronouns (*I, you*), some numerals in some languages.
- **Adjectives:** modify noun phrases; also predicates (*he is smart*).
- **Verbs:** predicates of clauses.
- **Adverbs:** modify verbs, usually as adjuncts (optional).
- **Prepositions:** govern noun phrases and their case, semantically modify their relation to verbs or other nouns.
- **Coordinating conjunctions** (*and, or, but*).
- **Subordinating conjunctions** (*that*): join dependent to main clause.
- **Relative (not interrogative) pronouns** (*which*): merger of nouns/adjectives and subordinating conjunctions.

Syntactic Nouns

- Arguments of verbs (subject, object)
- Nominal predicates (*he is a teacher*)
- Attributes of other nouns ( cs: *auto prezidenta* “president’s car”)
 -  en: *Christmas present*: is *Christmas* a syntactic adjective or noun?
 - It would be syntactic adjective in  Czech (*vánoční*).
 - Even if definitions are purely syntactic, consensus across languages is not guaranteed because every language has its own set of syntactic constructions.
- Including
 - Pronouns: personal (*I, you, he, we*), indefinite (*somebody*), negative (*nothing*), total (*everyone*), some demonstratives (*this* in *this is ridiculous*)
 -  cs: some numerals in some cases (*pět* “five”, *deset* “ten”, *tisíc* “thousand”, *miliarda* “billion”, *třetina* “one third”, *sedmínásobek* “septuple”, *desatero* “ten kinds”)




Syntactic Adjectives

- Modify a noun phrase, may agree with it in gender, number, and case. Include:
 - Possessive **pronouns (determiners?)** (*my, your, his, our*)
 - Demonstrative pronouns in some contexts (**this** *apple is sweet*)
 - Some indefinite and other pronouns in some languages ( cs: *nějaký* “some”, *každý* “every”, *žádný* “no”) (in other languages these may not be traditionally considered pronouns)
 - Cardinal numerals (but see next slide) (*one, two, three*)
 - Adjectival ordinal numerals (*first, second, third*)
 - Adjectivally used participles (**traveling** *salesman*, **mixed** *feelings*)
 - Possibly even “adjectivally” used nouns in compounds (**Christmas** *present*, **car** *repair*, **New York Times** *advisory board member*)







Syntactic Behavior of Czech Cardinal Numerals





- *jeden* “one”, *dva* “two”, *tři* “three”, *čtyři* “four” are syntactic adjectives. They agree in case (and also gender and number) with the counted noun.
- *pět* “five” and higher may behave as syntactic nouns.
 - whole phrase in nominative / accusative / vocative: the numeral governs the counted noun, forces it to genitive: *pět.Nom židlí.Gen* “five chairs”, not *pět *židle.Nom* ⇒ *pět* is a **syntactic noun**
 - whole phrase in other genitive / dative / locative / instrumental: the numeral agrees in case with the counted noun ⇒ it modifies the noun: *k pěti.Dat židlím.Dat* “to five chairs” ⇒ *pěti* is a **syntactic adjective**
- *tisíc* “thousand”, *milión* “million”, *miliarda* “billion” in both Czech and English can be used as
 - nouns (morphologically and syntactically): *z banky zmizely milióny* “millions vanished from the bank”
 - traditional numerals, syntactic nouns: *dluží mi milión dolarů* “he owes me one million dollars”

- Predicates of main clauses (especially finite verbs)
- Predicates of dependent clauses (especially nonfinite forms)
- Auxiliary verbs, modal verbs or other parts of complex verb forms:
 -  en: *would have been willing (to) keep smiling*
 -  cs: *bych byl býval mohl chtít udělat* “(I) could have wanted to do”
- Copulas in nominal predicates:
 -  en: *he **is** a teacher*

Syntactic Adverbs




- Modify verbs, optionally specify circumstances such as location, time, manner, extent, cause...
- Modify adjectives, specify degree or manner (*very large*)
- Modify other adverbs, specify degree or manner (*very well*)
- Including:
 - some ordinal numerals:  cs: *poprvé* “for the first time”
 - multiplicative numerals:  cs: *dvakrát* “twice”, *pěťasedmdesátkrát* “seventy-five times”
 - converbs (gerunds, transgressives):  cs: *čekajíc na autobus všimla si ho* “she noticed him **while waiting** for a bus”;  hi: दरवाज़ा खोलकर वह कमरे में आई (*darvāzā kholkar vah kamre mẽ āī*) “**having opened** the door she came in”


Conjunctions

- Coordinating conjunctions join phrases of same or similar type or even whole clauses (independent)
 - single coordinators:
 -  en: *Peter **and** Paul; today **or** tomorrow; he wanted to go **but** she didn't like the idea*
 - paired coordinators:
 -  en: ***neither** here **nor** there; **the** sooner **the** better; **as** soon **as** possible*
- Subordinating conjunctions join dependent clauses or phrases to the governing node, specifying their function
 - single subordinators:
 -  en: *that; if; whether; because*
 - paired subordinators:
 -  hi: जब मैं कहूँगा तब आना (***jab** maĩ kahũgā **tab** ānā*) (lit. ***when** I tell **then** come*) “come when I tell you”



Relative Pronouns, Determiners, Numerals, and Adverbs




- Merge properties of syntactic nouns / adjectives / adverbs and of subordinating conjunctions
 - relative syntactic noun: *those **who** know; a car **that** never breaks; the man **whom** I met; who knows **what** you find*
 - relative syntactic adjective: *the man **whose** son is this boy; you decide from **what** time on you work; ... **which** color you like*
 - cs: relative numeral: *pověz mi, **kolik** máš peněz* “tell me **how much** money you have”; ... ***kolikátý** jsi byl* (lit. ***how-many-th** you were*) “where did you rank”
 - relative syntactic adverb: *I don't know **when** she came; ... **where** it is; ... **how** to say; ... **why** he's here*
- Interrogative pronouns (adverbs etc.) may have same form (in some languages) but not the same joining function.





- Govern syntactic noun (dictate its case marking), specify its role as argument of
 - a verb (*believe **in** something*)
 - another noun (*lack **of** something*)
 - an adjective (*acceptable **for** me*)
- Appear before, after or around the noun phrase:
 - Preposition:  en: ***in** the house; **under** the table; **beyond** this point*
 - Postposition:  hi: कमरे में (*kamre **mē***) (lit. *room **in***) “in the room”
 - Circumposition:  de: ***von** diesem Zeitpunkt **an*** “**from** this moment **on**”

- **Semantic noun:** a concrete or abstract entity
 -  cs: *otcův* “father's” is traditionally a possessive adjective but could be regarded as a form of the semantic noun *otec* “father”; not to confuse with genitive case *otce/otců*





Semantic / Notional Criteria

- **Semantic noun:** a concrete or abstract entity
 -  cs: *otcův* “father's” is traditionally a possessive adjective but could be regarded as a form of the semantic noun *otec* “father”; not to confuse with genitive case *otce/otců*
- **Semantic adjective:** a quality, property
 -  en: *cleverly* could be regarded as a form of the semantic adjective *clever*
 - How far should we go? Is *cleverness* an adjective, too? What purpose would such classification serve?

- **Semantic noun:** a concrete or abstract entity
 -  cs: *otcův* “father's” is traditionally a possessive adjective but could be regarded as a form of the semantic noun *otec* “father”; not to confuse with genitive case *otce/otců*
- **Semantic adjective:** a quality, property
 -  en: *cleverly* could be regarded as a form of the semantic adjective *clever*
 - How far should we go? Is *cleverness* an adjective, too? What purpose would such classification serve?
- **Semantic adverb:** a circumstance (location, time, manner)
 -  cs: traditional adjective *zířejší* could be regarded as a form of the semantic adverb *zítra* “tomorrow”

- **Semantic noun:** a concrete or abstract entity
 -  cs: *otcův* “father's” is traditionally a possessive adjective but could be regarded as a form of the semantic noun *otec* “father”; not to confuse with genitive case *otce/otců*
- **Semantic adjective:** a quality, property
 -  en: *cleverly* could be regarded as a form of the semantic adjective *clever*
 - How far should we go? Is *cleverness* an adjective, too? What purpose would such classification serve?
- **Semantic adverb:** a circumstance (location, time, manner)
 -  cs: traditional adjective *zířejší* could be regarded as a form of the semantic adverb *zítra* “tomorrow”
- **Semantic verb:** an action (or a state?)
 -  cs: deverbative nouns (*dělání* “doing”) and adjectives (*dělající* “doing”; *udělavší* “the one that did”; *udělaný* “done”) could be regarded as forms of the semantic verb

Semantic / Notional Criteria

- **Semantic noun:** a concrete or abstract entity
 -  cs: *otcův* “father's” is traditionally a possessive adjective but could be regarded as a form of the semantic noun *otec* “father”; not to confuse with genitive case *otce/otců*
- **Semantic adjective:** a quality, property
 -  en: *cleverly* could be regarded as a form of the semantic adjective *clever*
 - How far should we go? Is *cleverness* an adjective, too? What purpose would such classification serve?
- **Semantic adverb:** a circumstance (location, time, manner)
 -  cs: traditional adjective *zířejší* could be regarded as a form of the semantic adverb *zítra* “tomorrow”
- **Semantic verb:** an action (or a state?)
 -  cs: deverbative nouns (*dělán* “doing”) and adjectives (*dělající* “doing”; *udělavší* “the one that did”; *udělaný* “done”) could be regarded as forms of the semantic verb
- **Numeral:** a number, amount
 - *one, two, three; first, second, third; once, twice, thrice; twofold; pair, triple, quadruple*

- **Pronoun:** any referential word (trad. pronoun, determiner, numeral, adverb / personal, possessive, indefinite, universal, negative, interrogative, relative, demonstrative)

- **Pronoun:** any referential word (trad. pronoun, determiner, numeral, adverb / personal, possessive, indefinite, universal, negative, interrogative, relative, demonstrative)
- **Function word:** adposition / conjunction / particle / auxiliary


Summary

- Category labels are “universal”
- Definitions are language-specific
- Ideal: cross-linguistic similarity of categories
- 100% match is not possible

<https://ufal.cz/courses/npf1094>

Openness vs. Closeness

Content vs. Function Words

- Open classes (take new words)
 - verbs (non-auxiliary), nouns, adjectives, adjectival adverbs, interjections
 - word formation (derivation) across classes
- Closed classes (words can be enumerated)
 - pronouns / determiners, adpositions, conjunctions, particles
 - pronominal adverbs
 - auxiliary and modal verbs / particles
 - numerals (mathematically infinite, linguistically closed)
 - typically they are not base for derivation
- Even closed classes evolve but over longer period of time
 -  es: *Vuestra Merced* “Your Mercy, Your Grace” ⇒ *usted* (new singular second person pronoun in formal/honorific register)
 - ⇒ new plural *ustedes*

The Big Four

- Nouns
 - Proper nouns
- Verbs
 - Verbal nouns, infinitives
 - Participles (between verbs and adjectives)
 - Converbs (between verbs and adverbs)
- Adjectives
 - Modify nouns
- Adverbs
 - Modify verbs, adjectives or adverbs

Common Minors

- Adpositions
 - Prepositions
 - Postpositions
 - Circumpositions
- Conjunctions
 - Subordinators
 - Coordinators
- Interjections
- Particles
 - Often just “garbage can category”

Pronouns vs. Determiners

- In some tagsets clear (but context-dependent) definition:
 - Pronouns *replace* noun phrases
 - *I, you, he, she, it, we, they, who, something...*
 - *This is unbelievable!*
 - *Yours is better.*
 - Determiners *modify* noun phrases (so they include traditional possessive “pronouns”)
 - *my, your, his, her, its, our, their, whose, some...*
 - *This book is John's.*
 - *Your book is better.*
- Some traditional grammars (and tagsets) refer to all the above as pronouns (e.g. Czech)



BulTreeBank Tagset (bg): The Broadest Sense of Pronouns

- Subcategories:
 - **personal:** аз “I”, ти “you”, той “he”
 - **possessive:** мой “my”, моя “my”, твой “your”, негов “his”
 - **demonstrative:** този “this”, това “this”
 - **interrogative:** кой “who”, коя “who”, кое “what”
 - **relative:** който “who”, що “that”
 - **universal:** всеки “each”, всякой “any”
 - **indefinite:** един “one”, някой “somebody”
 - **negative:** никой “nobody”, никакъв “none”
- Referential type:
 - **entity:** кой “who”, коя “who”, кое “what”
 - **attribute:** какъв “what”, каква “what”, какво “what”
 - **possession:** чий “whose”, чия “whose”, чие “whose”
 - **quantity:** колко “how much”, доколко “how much”
 - **location:** къде “where”, где “where”, докъде “where to”
 - **time:** кога “when”, докога “till when”, откога “since when”
 - **manner:** как “how”
 - **cause:** защо “why”











Prague Dependency Treebank (cs): The Greatest Variety of Numerals

- **Cardinal:** *jeden, dva, tři, čtyři, pět*
- **Cardinal for sets:** *jedny, dvojce, troje*
- **Cardinal / N-tuple:** *dvé, tré, čtvero*
 - But noun n-tuple: *dvojice, trojice*
- **Adjectival ordinal:** *první, druhý, třetí*
- **Adverbial ordinal:** *poprvé, podruhé*
- **Adjectival multiplicative (N sorts of):** *dvojí, trojí, čtverý, paterý*
- **Adverbial multiplicative:** *jedenkrát, dvakrát, třikrát, čtyřikrát, pětkrát*
- **Fraction:** *polovina, třetina, čtvrtina*
- **Number Arabic digits:** *1, 2, 3, 4, 5*
- **Number Roman:** *I, II, III, IV, V*

Pronominal quantifiers:

- **Interrogative / relative:** *kolik, kolikery, kolikero, kolikátý, pokolikáté, kolikerý, kolikrát*
- **Indefinite:** *několik, několikery, několikero, několikátý, poněkolikáté, několikery, několikrát, mnoho, málo*
- **Demonstrative:** *tolik, tolikery, tolikero, tolikátý, potolikáté, tolikery, tolikrát*







Some Endemic Classes

- Existential *there* in  English
- Infinitival marker:  English *to*,  German *zu*,  Swedish *att*
- Predeterminer:  English *both the boys*, *all the people*
- Response particles:  English *yes*, *no*, *thanks*
- Negative particles:  English *not*, *n't*,  Arabic لا (*lā*)
- Question particles:  Polish *czy*,  Hindi क्या (*kyā*)
- Separable verbal prefix:  German *vorstellen* \Rightarrow *stellen Sie sich vor* “imagine”
- Adjectival particle:  German *am besten*, *zu groß*
- Classifier:  Chinese 一個人 (*yí gè rén*) “one (piece) man”






Various Other Classes

- Foreign words (foreign-language quotations, names of books etc.; not loanwords!)
 - *The police confiscated illegal copies of the banned **Mein Kampf** by Adolf Hitler.*
 - Could be subclassified as foreign nouns, verbs etc.
 - POS and features need not be the same as in the source language!
 - German *Burg* “castle” is feminine. If embedded in Czech it will be treated as masculine.
- Abbreviations
 - Could be subclassified as abbreviated nouns, verbs etc.
- Parts of multi-token idioms
- Numbers (123)
- Symbols (\$, €)
- Punctuation (, . - “ ”)

Clitics and Contractions

- Clitic is a
 - Syntactically independent word
 - Phonologically / orthographically dependent morpheme
-  es: *despiértate* “wake yourself”; *démelo* “give me it”
-  ru: *защищатьсЯ* (*zaščiščat'sja*) “to defend oneself”
-  de: *zum* = *zu dem* “to the”; *am* = *an dem* “on the”
-  fr: *du* = *de le* “of the”
-  cs: *proň* = *pro něj* “for him”; *oč* = *o co* “for what”; *tys* = *ty jsi* “you have”; *žes* = *že jsi* “that you have”; *scvrnks* = *scvrnkl jsi* “you flicked off”; *přišelť* = *neboť přišel* “because he came”
-  ar: *وبالفالوجة* (*wabiālfālūjah* = *wa.CONJ bi.PREP AlfAlwjp.NOUN_PROP*) “and in al-Falujah”

Features of Nouns and Adjectives


- Gender / animacy (lexical for nouns, agreement inflection for adjectives) or class (Bantu languages)
- Number (singular, dual, plural, trial, paucal)
- Case ( en: 2 for pronouns;  cs: 7;  fi: 14)
- Definiteness ( ro: *poiană* “a meadow”, *poiana* “the meadow”)
- Polarity ( cs: *schopný* “able”, *neschopný* “unable”)
- Degree of comparison (positive, comparative, superlative, absolute superlative)








Noun Classes in Swahili

Class	Sing	Plur	Gloss
1 (humans)	m + <i>tu</i>	wa + <i>tu</i>	person
3 (thin objects)	m + <i>ti</i>	mi + <i>ti</i>	tree
5 (paired things)	ji + <i>cho</i>	ma + <i>cho</i>	eye
7 (instrument)	ki + <i>tu</i>	vi + <i>tu</i>	thing
11 (extended body parts)	u + <i>limi</i>	n + <i>dimi</i>	tongue

Features of Verbs

- Form: infinitive, verbal noun, participle, converb, supine, finite
- Mood: indicative, imperative, subjunctive, jussive, conditional, potential, optative, necessitative...
- Tense / aspect: present, past, future; continuous; aorist, imperfect, perfect, pluperfect
- Evidentiality: did I witness it myself?
- Voice: active, middle, passive, causative
- Person: 1st, 2nd, 3rd, 4th, 0, honorific registers
- Number: singular, dual, plural
- Gender: masculine, feminine, neuter
- Polarity:  cs: *dělat* “to do”, *nedělat* “not to do”



- Case of adpositions (subcategorization, not inflection)
 - What case must the governed noun phrase be in?
- Possessor's gender and number
 -  cs: *jejímu* *psovi* “to her dog” – feminine possessor, masculine possessed
 -  cs: *její* *krávě* “to her cow” – feminine possessor, feminine possessed
 -  cs: *jehož* *kráva* “whose (‘of which guy’) cow” – singular masculine possessor, singular feminine possessed
 -  cs: *jejíž* *kráva* “whose (‘of which woman’) cow” – singular feminine possessor, singular feminine possessed
 -  cs: *jejichž* *kráva* “whose (‘of which people’) cow” – plural possessor, singular possessed

- Tag as a set of feature (category) values ... (k_1, k_2, \dots, k_n)
- Simple list of tags

$$T = \{t_i\}_{i=1..n}$$

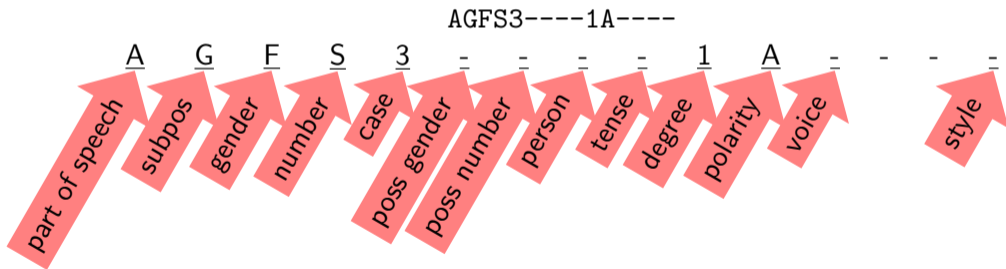
- 1-1 mapping between tags and feature-value space

$$T \leftrightarrow (K_1, K_2, \dots, K_n)$$

-  English
 - Penn Treebank (45 tags), Brown Corpus (87), Claws c5 (62), London-Lund (197)
-  Czech
 - Prague Dependency Treebank (4294; positional), Multext-East (1458; Orwell 1984 parallel corpus), Majka / Desam (MU Brno), Prague Spoken Corpus (over 10000!)
 - `<i1>0</i1><i2>1</i2><i3>1</i3><i4>1</i4><i5>1</i5><i6>_</i6><i7>_</i7><i8>_</i8><i9>_</i9><i10>_</i10><i11></i11>`



Czech Positional Tags of PDT





Parts of Speech in PDT

- N noun (*podstatné jméno*)
- A adjective (*přídavné jméno*)
- P pronoun (*zájmeno*)
- C numeral (*číslovka*)
- V verb (*sloveso*)
- D adverb (*příslowce*)
- R preposition (*předložka*)
- J conjunction (*spojka*)
- T particle (*částice*)
- I interjection (*citoslovce*)
- Z special (e.g. punctuation) (*zvláštní, např. interpunkce*)
- X unknown word (*neznámé slovo*)



Gender in PDT

M	masculine animate	Y	M or I
I	masculine inanimate	T	I or F
F	feminine	W	I or N
N	neuter	H, Q	F or N
X	unknown	Z	M, I or N



Number in PDT

S	singular (<i>jednotné</i>)
D	dual (<i>dvojné</i>)
P	plural (<i>množné</i>)
X	unknown (<i>neznámé</i>)



Case in PDT

1	nominative
2	genitive
3	dative
4	accusative
5	vocative
6	locative
7	instrumental
X	unknown



Degree, Polarity, and Person

- Degree of comparison of adjectives and adverbs:
 - 1 (positive), 2 (comparative), 3 (superlative)
- Polarity of verbs, adjectives, adverbs, and nouns:
 - A (affirmative), N (negative)
- Person of pronouns and verbs:
 - 1, 2, 3



Mood, Tense, and Voice

- Changes relevance of other categories (such as person and number) \Rightarrow in a sense, these are subparts of speech
- Tense:
 - P (present), M (past), F (future)
- Voice:
 - A (active), P (passive)
- Mood:
 - N (indicative), R (imperative), C (conditional)



Style and/or Variant

1	other variant, less frequent
2	other variant, very rare, archaic or literary
3	very archaic or colloquial variant
5	colloquial, tolerated both in spoken and in written discourse
6	colloquial, inappropriate in written discourse
7	colloquial like 6 but less preferred by speakers
9	special usage (e.g. after some prepositions)



The Penn Treebank Tagset

- ① **CC** coordinating conjunction
- ② **CD** cardinal number
- ③ **DT** determiner
- ④ **EX** existential *there*
- ⑤ **FW** foreign word
- ⑥ **IN** preposition or subordinating conjunction
- ⑦ **JJ** adjective
- ⑧ **JJR** adjective, comparative
- ⑨ **JJS** adjective, superlative
- ⑩ **LS** list item marker
- ⑪ **MD** modal
- ⑫ **NN** noun, singular/mass
- ⑬ **NNS** noun, plural
- ⑭ **NNP** proper noun, singular
- ⑮ **NNPS** proper noun, plural
- ⑯ **PDT** predeterminer
- ⑰ **POS** possessive ending
- ⑱ **PRP** personal pronoun
- ⑲ **PRP\$** possessive pronoun



The Penn Treebank Tagset

- 20 **RB** adverb
- 21 **RBR** adverb, comparative
- 22 **RBS** adverb, superlative
- 23 **RP** particle
- 24 **SYM** symbol
- 25 **TO** *to*
- 26 **UH** interjection
- 27 **VB** verb, base (*do*)
- 28 **VBD** verb, past (*did*)
- 29 **VBG** verb, gerund or present participle (*doing*)
- 30 **VBN** verb, past participle (*done*)
- 31 **VBP** verb, non-3rd person singular present (*do*)
- 32 **VBZ** verb, 3rd person singular present (*does*)
- 33 **WDT** wh-determiner (*which*)
- 34 **WP** wh-pronoun (*who*)
- 35 **WP\$** possessive wh-pronoun (*whose*)
- 36 **WRB** wh-adverb (*where*)
- 37 **.** period...

Universal POS Tags

<http://universaldependencies.org/u/pos/index.html>

- **NOUN**
- **PROPN** (proper noun)
- **VERB**
- **ADJ** (adjective)
- **ADV** (adverb)
- **INTJ** (interjection)
- **PRON** (pronoun)
- **DET** (determiner)
- **AUX** (auxiliary)
- **NUM** (numeral)
- **ADP** (adposition)
- **SCONJ** (subordinating conjunction)
- **CCONJ** (coordinating conjunction)
- **PART** (particle)
- **PUNCT** (punctuation)
- **SYM** (symbol)
- **X** (unknown)

Universal Features

<http://universaldependencies.org/u/feat/index.html>

- **PronType** (*druh zájmena*)
- **NumType** (*druh číslovky*)
- **Poss** (*přivlastňovací*)
- **Reflex** (*zvratné*)
- **Foreign** (*cizí slovo*)
- **Abbr** (*zkratka*)
- **Typo** (*překlep*)
- **Gender** (*rod*)
- **Animacy** (*životnost*)
- **NounClass** (*jmenná třída*)
- **Number** (*číslo*)
- **Case** (*pád*)
- **Definite(ness)** (*určitost*)
- **Degree** (*stupeň*)
- **VerbForm** (*slovesný tvar*)
- **Mood** (*způsob*)
- **Tense** (*čas*)
- **Aspect** (*vid*)
- **Voice** (*slovesný rod*)
- **Evident(iality)** (*zjevnost*)
- **Polarity** (*zápor*)
- **Person** (*osoba*)
- **Polite(ness)** (*zdvořilost*)
- **Clusivity** (*kluzivita*)