Caution

• No class on

October 20

&

October 27
Parts of Speech

Daniel Zeman

http://ufal.mff.cuni.cz/daniel-zeman/
zeman@ufal.mff.cuni.cz
Part of Speech

- Vague definitions, criteria of mixed nature
- **Looong 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
History

- 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** (Ουσιαστικό *ousiastiko*)
  - inflected for case, signifying a concrete or abstract entity
- **Verb** (Ρήµα *rîma*)
  - without case inflection, but inflected for tense, person and number, signifying an activity or process performed or undergone
- **Participle** (Μετοχή *metohî*)
  - sharing the features of the verb and the noun
- **Interjection** (Επιφώνηµα *epifônîma*)
  - expressing emotion alone
- **Pronoun** (Αντωνυµία *antônymia*)
  - substitutable for a noun and marked for person
- **Preposition** (Πρόθεση *prothesî*)
  - placed before other words in composition and in syntax
- **Adverb** (Επίρρηµα *epirrîma*)
  - without inflection, in modification of 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 Beuzée: *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, adverbiu*
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) and numerals (pět, tisíc, sedmero, polovina)
Morphological Criteria

• By definition language-dependent. In Czech (simplified):
  – Nouns: (gender), number, case. Include some pronouns (někdo) and numerals (pět, tisíc, sedmero, polovina)
  – Adjectives: gender, number, case, sometimes degree; agr. with N. Include some pronouns (který, žádný) and numerals (první, druhý, čtverý)
Morphological Criteria

- By definition language-dependent. In Czech (simplified):
  - Nouns: (gender), number, case. Include some pronouns (*někdo*) and numerals (*pět, tisíc, sedmero, polovina*)
  - Adjectives: gender, number, case, sometimes degree; agr. with N. Include some pronouns (*který, žádný*) and numerals (*první, druhý, čtverý*)
  - Personal pronouns: person, gender, number, case
Morphological Criteria

- By definition language-dependent. In Czech (simplified):
  - Nouns: (gender), number, case. Include some pronouns (někdo) and numerals (pět, tisíc, sedmero, polovina)
  - Adjectives: gender, number, case, sometimes degree; agr. with N. Include some pronouns (který, žádný) and numerals (první, druhý, čtverý)
  - Personal pronouns: person, gender, number, case
  - Possessive pronouns: possessor’s person, gender & number; possessed gender & number
Morphological Criteria

• By definition language-dependent. In Czech (simplified):
  – Nouns: (gender), number, case. Include some pronouns (někdo) and numerals (pět, tisíc, sedmero, polovina)
  – Adjectives: gender, number, case, sometimes degree; agr. with N. Include some pronouns (který, žádný) and numerals (první, druhý, čtverý)
  – Personal pronouns: person, gender, number, case
  – Possessive pronouns: possessor’s person, gender & number; possessed gender & number
  – Verbs:
    • infinitive
    • finite: mood (indicative/imperative), tense (present/future), person, number
    • participle: voice (active/passive), gender, number
    • transgressive: tense (present/past), gender, number
Morphological Criteria

- By definition language-dependent. In Czech (simplified):
  - Nouns: (gender), number, case. Include some pronouns (někdo) and numerals (pět, tisíc, sedmero, polovina)
  - Adjectives: gender, number, case, sometimes degree; agr. with N. Include some pronouns (který, žádný) and numerals (první, druhý, čtverý)
  - Personal pronouns: person, gender, number, case
  - Possessive pronouns: possessor’s person, gender & number; possessed gender & number
  - Verbs:
    - infinitive
    - finite: mood (indicative/imperative), tense (present/future), person, number
    - participle: voice (active/passive), gender, number
    - transgressive: tense (present/past), gender, number
  - 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.
  – Adjectives: modify noun phrases.
  – Verbs: predicates of clauses.
  – Adverbs: modify verbs, usually as adjuncts (non-obligatory).
  – Prepositions: govern noun phrases, dictate 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 predicate (*he is a teacher*) etc.
• Attributes of other nouns (CS: *auto prezidenta* = *president’s car*)
  – EN: *Christmas present*: is *Christmas* a syntactic adjective or noun?
  – 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*), totality (*everyone*), some demonstratives (*this in this is ridiculous*)
  – CS: some numerals in some cases (*pět, deset, tisíc, miliarda, třetina, sedminásobek, desatero*)
Syntactic Adjectives

• Modify a noun phrase, typically 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 *(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 (five chairs)/*gen, not *pět *židle /*nom ⇒ pět is syntactic noun
  - whole phrase in other cases: 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 a bank*
  - traditional numerals, syntactic nouns: *dluží mi milión dolarů = he owes me one million dollars*
Syntactic Verbs

- Predicate of a main clause
- Predicate of a dependent clause
- Auxiliary verb, modal verb or another part of a complex verb form:
  - en: would have been willing (to) keep smiling 😊
  - cs: bych byl býval mohl chtít udělat
    (= (I) could have wanted to do)
- Copula in nominal predicates:
  - en: he is a teacher
Syntactic Adverbs

• Modify verbs, optionally specify circumstances such as location, time, manner, extent, cause…
• Can also modify adjectives *(very large)* or other adverbs *(very well)*
• Including:
  – some ordinal numerals: cs: *poprvé* *(for the first time)*
  – multiplicative numerals: cs: *dvakrát* *(twice)*, *pětasedmdesátkrát* *(seventy-five times)*
  – converses (transgressives): cs: *čekajíc na autobus všimla si ho* *(she noticed him while waiting for a bus)*; hi: दरवाजा खोलकर वह कमरे में आई* *darvāzā kholkar vah kamre mē 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:
    • *Peter and Paul; today or tomorrow; he wanted to go* but *she didn’t like the idea*
  – paired coordinators:
    • *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:
    • *that; so; if; whether; because*
  – paired subordinators:
    • *hi:जब मैं कहूँगा तब आना* jab maĩ kahūɡā tab ānā (lit: when I tell then come)
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 numerals: *pověz mi, kolik máš peněz* (tell me how much money you have); ...*kolikátý jsi byl* (where did you rank; lit. how-many-th you were)
  - 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.
Adpositions

• Govern syntactic noun (dictate its case marking), specify its role as argument of
  – a verb (*believe in something*)
  – another noun (*lack of something*)
  – or adjective (*acceptable for me*)

• Appear before, after or around the noun phrase:
  – Preposition: *in the house; under the table; beyond this point*
  – Postposition: *है: कमरे में kamre mē (lit. room in)*
  – Circumposition: *de: von diesem Zeitpunkt an (from this moment on)*
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 / 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 / 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 / Notional Criteria

- Semantic noun: a concrete or abstract entity
  - \textit{cs: otcův} (father’s) is traditionally a possessive adjective but could be regarded as a form of the semantic noun \textit{otec} (father); not to confuse with genitive case \textit{otce/otců}

- Semantic adjective: a quality, property
  - \textit{en: cleverly} could be regarded as a form of the semantic adjective \textit{clever}
  - How far should we go? Is \textit{cleverness} an adjective, too? What purpose would such classification serve?

- Semantic adverb: a circumstance (location, time, manner)
  - \textit{cs:} traditional adjective \textit{zířejší} could be regarded as a form of the semantic adverb \textit{zítra} (tomorrow)

- Semantic verb: a state or an action
  - \textit{cs:} deverbal nouns (\textit{dělání = the doing}) and adjectives (\textit{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: a state or an action
  - **cs**: deverbative nouns (_dělání = the doing_ and adjectives (_dělající = doing; udělavší = the one that did; udělaný = done_) could be regarded as forms of the semantic verb

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

• 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: a state or an action
  – **cs**: deverbative nouns (*dělání = the doing*) and adjectives (*dělající = doing; udělavší = the one that did; udělaný = done*) could be regarded as forms of the semantic verb

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

• Numeral: a number, amount (*one, two, three; first, second, third; once, twice, thrice; twofold; pair, triple, quadruple*)
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 oteř/otec

• 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: a state or an action
  – **cs**: deverbal nouns (*dělání = the doing*) and adjectives (*dělající = doing; udělavší = the one that did; udělaný = done*) could be regarded as forms of the semantic verb

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

• Numeral: a number, amount (*one, two, three; first, second, third; once, twice, thrice; twofold; pair, triple, quadruple*)

• Adpositions + conjunctions + particles + auxiliaries (glue material)
Summary

- Category labels are “universal”
- Definitions are language-specific
- Ideal: cross-linguistic similarity of categories
- 100% match is not possible
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
  – 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 2nd person pronoun in formal/honorific register)
The Big Four

- **Nouns**
  - Proper nouns

- **Verbs**
  - Participles (between verbs and nouns / adjectives / adverbs)

- **Adjectives**
  - Modify nouns

- **Adverbs**
  - Modify verbs, adjectives or adverbs
Common Minors

• Adpositions
  – Prepositions
  – Postpositions
  – Circumpositions
• Conjunctions
  – Subordinators
  – Coordinators
• Interjections
• Particles (often “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”)
    - *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, ти, твой
  - possessive: мой, моя, твой, негов
  - demonstrative: този, тоя
  - interrogative: кой, коя, кое
  - relative: който, що
  - collective: всеки, всякой
  - indefinite: един, някой
  - negative: никой, никакъв

- **Referential type:**
  - entity: кой, коя, кое
  - attribute: какъв, каква, какво
  - possession: чий, чия, чие
  - quantity: колко, доколко
  - location: къде, где, докъде
  - time: кога, докога, откога
  - manner: как
  - cause: защо
Numerals vs. Adjectives

• Many tagsets distinguish *cardinal numbers*
  – While some (Danish) take them as special class of adjectives

• Ordinal numbers
  – Sometimes separate POS
  – Sometimes special class of adjectives
  – Sometimes “normal” adjectives (undistinguished)
Prague Dependency Treebank (CS): The Greatest Variety of Numerals

- Cardinal: *jeden, dva, tři, čtyři, pět*
- Adjectival ordinal: *první, druhý, třetí, čtvrtý, pátý*
- Adverbial ordinal: *poprvé, podruhé*
- Multiplicative: *jedenkrát, dvakrát*
- Generic (N sets of): *jedny, dvoje, troje, čtyvery, patery*
- Generic (N sorts of): *dvojí, trojí, čtverý, paterý*
- Generic (N-tuple): *dvé, tré, čtvero, patero*
  - But noun n-tuple: *dvojice, trojice*
- Fraction: *polovina, třetina, čtvrtina*
- Number Arabic digits: *1, 2, 3, 4, 5*
- Number Roman: *I, II, III, IV, V*

Pronominal quantifiers:

- Interrogative / relative: *kolik, kolikátý, pokolikáte, kolikrát, kolikery, kolikerý, kolikero*
- Indefinite: *několik, několikátý, poněkolikáte, několikrát, několikery, několikerý, několikero, mnoho, málo*
- Demonstrative: *tolik, tolikátý, potolikáte, tolikrát, tolikery, tolikery, tolikero*
Some Endemic Classes

• Existential *there* in English
• Infinitival marker: English *to*, German *zu*, Swedish *att*
• Predeterminer: English *both the boys, all the people*
• Response particle: *yes, no, thanks*
• Negative particle: *not, n’t, Arabic َّ lā*
• Question particle: Polish *czy*, Hindi *क्या  kyā*
• Separable verbal prefix: German *vorstellen ⇒ stellen Sie sich vor*
• 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* is feminine. If embedded in Czech it will be treated as masc.

- Abbreviations
  - Could be subclassified as abbreviated nouns, verbs etc.

- Parts of multi-token idioms

- Numbers *(123)*

- Symbols *(\$, €)*

- Punctuation *(, . – “ ” )*
Clitics

- 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; scvrnkls = scvrnkl jsi = you flicked off;
  přišelť = nebot’ přišel = because he came

• ar: وبالقالوجة = wabiāl fālūjah = wa/CONJ + bi/PREP +
  AlfAlwjp/NOUN_PROP = and in al-Falujah
Features of Nouns and Adjectives

- Gender / animateness (lexical for nouns, agreemental 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; schopnost = ability, neschopnost = inability)
- Degree of comparison (positive, comparative, superlative, absolutive)
Noun Classes in Swahili

<table>
<thead>
<tr>
<th>Class</th>
<th>SG</th>
<th>PL</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (humans)</td>
<td>m + tu</td>
<td>wa + tu</td>
<td>person</td>
</tr>
<tr>
<td>3 (thin objects)</td>
<td>m + ti</td>
<td>mi + ti</td>
<td>tree</td>
</tr>
<tr>
<td>5 (paired things)</td>
<td>ji + cho</td>
<td>ma + cho</td>
<td>eye</td>
</tr>
<tr>
<td>7 (instrument)</td>
<td>ki + tu</td>
<td>vi + tu</td>
<td>thing</td>
</tr>
<tr>
<td>11 (extended body parts)</td>
<td>u + limi</td>
<td>n + dimi</td>
<td>tongue</td>
</tr>
</tbody>
</table>
Features of Verbs

- Form: infinitive, participle, gerund, transgressive, 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: 1\textsuperscript{st}, 2\textsuperscript{nd}, 3\textsuperscript{rd}, 4\textsuperscript{th}, 0, honorific registers
- Number: singular, dual, plural
- Gender of participles: masculine, feminine, neuter
- Polarity: dělat = to do, nedělat = not to do
Other Features

• 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**: *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
Tagsets

- Tag as a set of feature (category) values \( \ldots \) \((k_1, k_2, \ldots, k_n)\)
- Simple list of tags
  \[ T = \{t_i\}_{i=1..n} \]
- 1-1 mapping between tags and feature value space
  \[ T \leftrightarrow (K_1, K_2, \ldots, K_n) \]
- English
  - Penn TreeBank (45 tags), Brown Corpus (87), Claws c5 (62), London-Lund (197)
- Czech
  - Prague Dependency Treebank (4294; new positional, old compact), Multext-East (1458; Orwell 1984 parallel corpus), Ajka/Desam (MU Brno), Prague Spoken Corpus (over 10000!)
Czech Positional Tags of PDT
Parts of Speech in PDT

- **N** noun (*podstatné jméno*)
- **A** adjective (*přídavné jméno*)
- **P** pronoun (*zájmeno*)
- **C** numeral (*číselná událost*)
- **V** verb (*sloveso*)
- **D** adverb (*příslovce*)
- **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

<table>
<thead>
<tr>
<th></th>
<th>Masculine Animate</th>
<th>Neuter</th>
<th>Feminine</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>masculine animate</td>
<td>Y</td>
<td>M or I</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>masculine inanimate</td>
<td>T</td>
<td>I or F</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>feminine</td>
<td>W</td>
<td>I or N</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>neuter</td>
<td>H, Q</td>
<td>F or N</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>unknown</td>
<td>Z</td>
<td>M, I or N</td>
<td></td>
</tr>
</tbody>
</table>
## Number in PDT

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S</strong></td>
<td>singular (<em>jednotné</em>)</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>dual (<em>dvojné</em>)</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>plural (<em>množné</em>)</td>
</tr>
<tr>
<td><strong>X</strong></td>
<td>unknown (<em>neznámé</em>)</td>
</tr>
</tbody>
</table>
## Case in PDT

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nominative</td>
</tr>
<tr>
<td>2</td>
<td>genitive</td>
</tr>
<tr>
<td>3</td>
<td>dative</td>
</tr>
<tr>
<td>4</td>
<td>accusative</td>
</tr>
<tr>
<td>5</td>
<td>vocative</td>
</tr>
<tr>
<td>6</td>
<td>locative</td>
</tr>
<tr>
<td>7</td>
<td>instrumental</td>
</tr>
<tr>
<td>X</td>
<td>unknown</td>
</tr>
</tbody>
</table>
Degree, Negativeness and Person

• Degree of comparison of adjectives and adverbs:
  – 1 (positive), 2 (comparative), 3 (superlative)

• Negativeness (for V, A, D and N):
  – A (affirmative), N (negative)

• Person of verbs and pronouns:
  – 1, 2, 3
Mood, Tense and Voice

• Changes relevance of other categories (such as person and number) ⇒ in a sense, these are (sub-)parts of speech
• Tense: present (P), past (M), future (F)
• Voice: active (A), passive (P)
• Mood: indicative (N), imperative (R), conditional (C – e.g. bych)
## Style and/or Variant

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>other variant, less frequent</td>
</tr>
<tr>
<td>2</td>
<td>other variant, very rare, archaic or literary</td>
</tr>
<tr>
<td>3</td>
<td>very archaic or colloquial variant</td>
</tr>
<tr>
<td>5</td>
<td>colloquial, tolerated in both spoken and written discourse</td>
</tr>
<tr>
<td>6</td>
<td>colloquial, inappropriate in written discourse</td>
</tr>
<tr>
<td>7</td>
<td>colloquial like 6 but less preferred by speakers</td>
</tr>
<tr>
<td>9</td>
<td>special usage (e.g. after some prepositions)</td>
</tr>
</tbody>
</table>
The Penn Treebank Tagset

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

20. **RB** adverb
21. **RBR** adv, comparative
22. **RBS** adv, 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 pres. participle *(doing)*
30. **VBN** verb, past participle *(done)*
31. **VBZ** verb, 3rd pers. sing. present *(does)*
32. **WDT** wh-det. *(which?)*
33. **WP** wh-pronoun *(who)*
34. **WP$** possessive wh-pronoun *(whose)*
35. **WRB** wh-adv. *(where)*
36. **.** period...
Universal POS Tags

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

- NOUN
- PROP (proper noun)
- VERB
- ADJ (adjective)
- ADV (adverb)
- INTJ (interjection)
- PRON (pronoun)
- DET (determiner)
- AUX (auxiliary)
- NUM (numeral)
- ADP (adposition)
- SCONJ (subordinating conj.)
- CONJ (coordinating conj.)
- 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)
- Gender (rod)
- Animacy (životnost)
- Number (číslo)
- Case (pád)
- Definite(ness) (určitost)
- Degree (stupeň)

- VerbForm (slovesný tvar)
- Mood (způsob)
- Tense (čas)
- Aspect (vid)
- Voice (slovesný rod)
- Person (osoba)
- Negative(ness) (zápor)