Contents

1 Introduction 9

2 Lemma and tag structure 11
  2.1 Lemma structure ............................................. 11
  2.1.1 Derivational Information ................................. 11
  2.1.2 Semantic Information ...................................... 12
  2.2 Tag Structure .................................................. 12
  2.2.1 Positional tags .............................................. 12
    2.2.1.1 1 – Part of speech .................................... 13
    2.2.1.2 2 – Detailed part of speech ......................... 14
    2.2.1.3 3 – Gender .............................................. 18
    2.2.1.4 4 – Number .............................................. 19
    2.2.1.5 5 – Case ................................................. 19
    2.2.1.6 6 – Possessor’s Gender ............................... 19
    2.2.1.7 7 – Possessor’s Number ................................ 19
    2.2.1.8 8 – Person .............................................. 20
    2.2.1.9 9 – Tense .............................................. 20
    2.2.1.10 10 – Degree of Comparison .......................... 20
    2.2.1.11 11 – Negation ......................................... 20
    2.2.1.12 12 – Voice ............................................ 21
    2.2.1.13 15 – Variant .......................................... 21
  2.2.2 Compact tags .............................................. 21
  2.2.3 Informal abbreviations ................................. 21

3 Names 23
  3.1 Personal names ............................................... 24
    3.1.1 von, van, etc. ........................................... 24
    3.1.2 Chinese names ............................................ 24
    3.1.3 Korean names ............................................. 25
    3.1.4 Foreignized Czech names ............................... 25
  3.2 Compound names (names consisting of other names) .......... 26
  3.3 Horses, Dj’s etc. ............................................. 27
  3.4 Sport clubs, etc. ............................................. 27
  3.5 Other ............................................................ 28
    3.5.1 Geographical names ...................................... 28
    3.5.2 Initials .................................................... 29
    3.5.3 Institutions, companies .................................. 29
    3.5.4 Sporting and other events .............................. 30
    3.5.5 Televisions ............................................... 31
    3.5.6 News, Magazines .......................................... 31
    3.5.7 Song names, etc. ......................................... 31
<table>
<thead>
<tr>
<th>Contents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Abbreviations</td>
<td>33</td>
</tr>
<tr>
<td>4.1 Gender</td>
<td>33</td>
</tr>
<tr>
<td>4.2 Normal abbreviations</td>
<td>33</td>
</tr>
<tr>
<td>4.3 Isolated letters</td>
<td>34</td>
</tr>
<tr>
<td>4.4 RM-systém, samopal SA-58</td>
<td>35</td>
</tr>
<tr>
<td>4.5 Units of measurements</td>
<td>35</td>
</tr>
<tr>
<td>4.6 Authors abbreviations</td>
<td>36</td>
</tr>
<tr>
<td>4.7 Academic titles</td>
<td>36</td>
</tr>
<tr>
<td>5 Colloquial Czech</td>
<td>37</td>
</tr>
<tr>
<td>5.1 Cos, jaks, kdys ...</td>
<td>37</td>
</tr>
<tr>
<td>5.2 Suffix -é in plural of neuter</td>
<td>37</td>
</tr>
<tr>
<td>6 Foreign words and phrases</td>
<td>39</td>
</tr>
<tr>
<td>6.1 Part of speech</td>
<td>39</td>
</tr>
<tr>
<td>6.1.1 English noun clusters</td>
<td>39</td>
</tr>
<tr>
<td>6.1.2 Examples</td>
<td>39</td>
</tr>
<tr>
<td>6.2 Articles</td>
<td>40</td>
</tr>
<tr>
<td>6.2.1 Citation use</td>
<td>40</td>
</tr>
<tr>
<td>6.2.2 Word use</td>
<td>40</td>
</tr>
<tr>
<td>6.2.3 Examples</td>
<td>40</td>
</tr>
<tr>
<td>6.3 Nouns</td>
<td>40</td>
</tr>
<tr>
<td>6.3.1 Citation use</td>
<td>40</td>
</tr>
<tr>
<td>6.3.1.1 English</td>
<td>40</td>
</tr>
<tr>
<td>6.3.2 Word use</td>
<td>41</td>
</tr>
<tr>
<td>6.3.2.1 English</td>
<td>41</td>
</tr>
<tr>
<td>6.4 Verbs</td>
<td>41</td>
</tr>
<tr>
<td>6.4.1 Citation use</td>
<td>41</td>
</tr>
<tr>
<td>6.4.1.1 English verbs</td>
<td>41</td>
</tr>
<tr>
<td>6.4.2 Word use</td>
<td>41</td>
</tr>
<tr>
<td>6.5 Slovak language</td>
<td>41</td>
</tr>
<tr>
<td>7 Errors</td>
<td>43</td>
</tr>
<tr>
<td>7.1 Characters</td>
<td>43</td>
</tr>
<tr>
<td>7.2 Separators, etc.</td>
<td>43</td>
</tr>
<tr>
<td>8 Hard to decide</td>
<td>45</td>
</tr>
<tr>
<td>8.1 až</td>
<td>45</td>
</tr>
<tr>
<td>8.2 jak</td>
<td>45</td>
</tr>
<tr>
<td>8.3 málo</td>
<td>47</td>
</tr>
<tr>
<td>8.4 moc</td>
<td>47</td>
</tr>
<tr>
<td>8.5 proto</td>
<td>48</td>
</tr>
<tr>
<td>8.6 svůj</td>
<td>48</td>
</tr>
<tr>
<td>8.7 tak</td>
<td>48</td>
</tr>
<tr>
<td>9 Sólokapři</td>
<td>51</td>
</tr>
<tr>
<td>9.1 Date and time</td>
<td>52</td>
</tr>
<tr>
<td>9.2 Numbers, numerals and quantifiers</td>
<td>52</td>
</tr>
<tr>
<td>9.3 Hyphenated composites</td>
<td>52</td>
</tr>
</tbody>
</table>
10 Insertion 55
  10.1 Possessive adjectives ........................................... 55
  10.2 Words ending with -ismus, -izmus .............................. 55
  10.3 Strange and unique things ........................................ 56
  10.4 Other ................................................................. 57

11 Errors in PDT 1.0 59
We are pleased to publish the first version of the manual for morphological annotation of Czech sentences. We believe that such guidelines can be of use to the users of Prague Dependency Treebank 1.0 (PDT 1.0), as well as for preparation of new data.

Let us recall the most important steps we passed in order to get about two million morphologically annotated words (PDT 1.0). At the very beginning, we put together a team of eight annotators – we did introduce them to a system of morphological tags we designed to describe Czech morphological properties; we also introduced them a morphological analyzer for processing isolated words we use (as a preprocessing step), and, last but not least, we did rely on their knowledge of Czech morphology they have acquired while studying at secondary school, i.e. we did not offer them any annotation guidelines.

One can assume that this strategy is too hazardous – how to deal with discrepancies the annotators produce to ensure the consistency of annotation? First, two annotators annotated each text file. Then, by a “blind” automatic procedure (no matter what word is processed – just comparing two strings) we detected words annotated differently. Consequently, the only one annotator (as a member of just two-member team) handled these cases and, also, checked the morphological annotations against the syntactic-analytical annotations. This way we replaced the absence of annotation guidelines by sequential elimination of discrepancies across both the morphological and syntactic-analytical levels of annotation.

Along the way we were writing this annotation manual. It is not intended as a comprehensive guide to the morphological annotation of Czech sentences (in contrast to the manual for syntactic-analytical annotations). The authors concentrate “only” on those cases which caused the most ambiguities and problems while annotating PDT 1.0. The ongoing effort is directed to the treating of not-yet-solved problematic cases in accord with the conventions of automatic morphological analyzer.

The morphological annotation of PDT 1.0 was carried out in the framework of experimental verification of the definition of formal representation of the analysis of Czech sentences (the project GAČR 405/96/0198, “Formal representation of language structures”). The material obtained in this way (data) is used in many domains of research in computational linguistics, above all as basic (training) data in projects of the automatic language analysis, the MŠMT research project MSM113000006, the “Laboratory for Language Data Processing” (the MŠMT project VS961510) and the Center for Computational Linguistics (the MŠMT project LN00A063). These data have been also used as verification material for various partial projects within the complex program GAČR 405/96/K214 (“Czech Language in Computer Age”). The “Center for Computational Linguistics” project financially supported work on these morphological annotation guidelines.

We are grateful to Petr Pajas – this document “as it is” would not appear without his XML and LaTex skills.

**Typographical conventions**

Vertical bar on the outer side of the page is used to highlight comments we make or suggestions we propose.

Gray is used to highlight something what should be checked.
Chapter 1

Introduction

Sometimes, the writer uses the word incorrectly – e.g. a name of a woman as a name of a man, surname as a first name, etc. It is necessary to annotate the real usage not the should-be usage.

Maybe it should be somehow marked, if we encounter it.

To get an idea what a foreign name, etc. mean it is useful to try to find using an internet portal, in an encyclopedia, on a map, etc. During annotation, we have found the following internet links useful:

**Portals**

http://www.seznam.cz – for Czech products, companies
http://search.seznam.cz/search.cgi?mod=f&hlp=y – for Czech companies
http://www.google.com
http://www.altavista.com (shop section for various searching products)

**Encyclopedias**

http://www.britannica.com
http://www.encyclopedia.com
http://www.encarta.msn.com

**Dictionaries**

http://slovnik.seznam.cz – various dictionaries

**Maps**

http://mapy.atlas.cz – Czechia
Chapter 2

Lemma and tag structure

2.1 Lemma structure

Lemma in PDT 1.0 has two parts. First part, the lemma proper, has to be a unique identifier of the lexical item. Usually it is the base form (e.g. infinitive for a verb) of the word, possibly followed by a number distinguishing different lemmas with the same base forms. Second part (optional) is not part of the identifier and contains additional information about the lemma, e.g. semantic or derivational information.

Note: There is a convention that if lemmas use numbers to distinguish lexical items with the same base form, they all have to use them - i.e. instead of sets of lemmas \{X, X-1, X-2\} or \{X, X-2, X-3\}, there should be a set \{X-1, X-2, X-3\}.

Note: The lemmas having different semantic suffixes should have different numbers.
In this manual we behave as the annotator. We try to mark such improper numbers by roman font (other part of the lemma is in italics). For example stop in akce Stop million will be marked as stop-1\_m and not stop-1\_m).

Table 2.1: Examples

<table>
<thead>
<tr>
<th>Whole lemma</th>
<th>Lemma proper</th>
<th>Second part</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemik</td>
<td>chemik</td>
<td></td>
</tr>
<tr>
<td>maso·(jídlo_apod.)</td>
<td>maso</td>
<td>_ (jídlo_apod.)</td>
</tr>
<tr>
<td>Bonn·G</td>
<td>Bonn</td>
<td>_ G</td>
</tr>
<tr>
<td>vazba-1_ (obviněněho)</td>
<td>vazba-1</td>
<td>_ (obviněněho)</td>
</tr>
<tr>
<td>vazba-2_ (spojení)</td>
<td>vazba-2</td>
<td>_ (spojení)</td>
</tr>
<tr>
<td>Martinů-1_Y_ (*4-1)</td>
<td>Martinů-1</td>
<td>_ Y_ (*4-1)</td>
</tr>
</tbody>
</table>

2.1.1 Derivational Information

The morphological component used in PDT 1.0, handles only inflection, not derivations – it means lemmas are rather shallow. However, sometimes the lemma contains information about lemmas it is derived. For example lemmas of possessive adjectives contain information about the noun they are derived from (otcův ← otec). The information is encoded in the following way – how many characters you have to remove from the end, and what string you have to add to get the deeper lemma. Only the proper lemmas are both input and output of this process.
Following examples illustrate this:

kardinálůvْ(*2) – remove two letters: kardinál
KarlůvْYْ(*3el) – remove 3 characters, add “el”: Karel
přijetí-2ْ(např.návrh)ْ(*5mout-2) – remove 5 characters, add “mout-2”: přijmout-2
Martinův-1ْYْ(*4-1) – remove 4 characters, add “-1”: Martin-1

Other examples:
SorosůvْSْ(*2)
chlapcůvْ(*3ec)
MáchůvْSْ(*2a)
Hlinkův-1ْSْ(*4a-1)
podáníْ(něco[někomu][někam])ْ(*3at)
prohlášeníْ(*4sit)
protíprávnostْ(*3ý)

2.1.2 Semantic Information
Some lemmas (esp. names) contain suffixes expressing semantic information about their use, etc.:
G – geographical name: Praha, Ústí nad Labem
Y – given (first) name, formerly used as default: Petr, John
S – surname (last name): Dvořák, Zelený, Agassi, Bush
E – name of a nationality: Čech, Kolumbijec
R – name of a product: Tatra (the car),
K – name of a company: Tatra (the company)
m – default – names of mines, stadiums, guerrilla bases, etc; also used for functional words in names.

2.2 Tag Structure

2.2.1 Positional tags
A positional tag is a string of 15 characters. Every position encodes one morphological category using one character (mostly upper case letters or numbers).

<table>
<thead>
<tr>
<th>Position</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>POS</td>
<td>Part of speech</td>
</tr>
<tr>
<td>2</td>
<td>SubPOS</td>
<td>Detailed part of speech</td>
</tr>
<tr>
<td>3</td>
<td>Gender</td>
<td>Gender</td>
</tr>
<tr>
<td>4</td>
<td>Number</td>
<td>Number</td>
</tr>
<tr>
<td>5</td>
<td>Case</td>
<td>Case</td>
</tr>
<tr>
<td>6</td>
<td>PossGender</td>
<td>Possessor’s gender</td>
</tr>
</tbody>
</table>

continued on next page
2.2 Tag Structure

<table>
<thead>
<tr>
<th>Position</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>PossNumber</td>
<td>Possessor’s number</td>
</tr>
<tr>
<td>8</td>
<td>Person</td>
<td>Person</td>
</tr>
<tr>
<td>9</td>
<td>Tense</td>
<td>Tense</td>
</tr>
<tr>
<td>10</td>
<td>Grade</td>
<td>Degree of comparison</td>
</tr>
<tr>
<td>11</td>
<td>Negation</td>
<td>Negation</td>
</tr>
<tr>
<td>12</td>
<td>Voice</td>
<td>Voice</td>
</tr>
<tr>
<td>13</td>
<td>Reserve1</td>
<td>Reserve</td>
</tr>
<tr>
<td>14</td>
<td>Reserve2</td>
<td>Reserve</td>
</tr>
<tr>
<td>15</td>
<td>Var</td>
<td>Variant, style</td>
</tr>
</tbody>
</table>

Some of the characters encode aggregation of more atomic values – for example: ‘X’ – means any value, ‘Y’ means masculine animate (‘M’) or inanimate (‘I’). Dash (‘-’) means no value (e.g. tense for nouns).

Not all combinations of tag values are possible. There is about 4K tags

Examples:

hraniční: AAIS4----1A---- standard adjective, masc. inanimate, singular, accusative, positive
potok: NNIS4-----A----- noun, masc. inanimate, singular, accusative, positive
karikaturistou: NNMS7-----A---- noun, masc. animate, singular, instrumental, positive
ODS: NNFXX-----A----8 noun, feminine, any number, any case, positive, abbreviation
podle: RR--2---------- preposition (non vocalized), requiring genitive
volen: VsYS---XX--AP--- verb, passive participle, masculine, singular, any person, any tense, positive, passive

2.2.1.1 1 – Part of speech

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adjective</td>
</tr>
<tr>
<td>C</td>
<td>Numeral</td>
</tr>
<tr>
<td>D</td>
<td>Adverb</td>
</tr>
<tr>
<td>I</td>
<td>Interjection</td>
</tr>
<tr>
<td>J</td>
<td>Conjunction</td>
</tr>
<tr>
<td>N</td>
<td>Noun</td>
</tr>
</tbody>
</table>
2.2.1.2 2 – Detailed part of speech

Further subcategorizes POS. The POS value is uniquely specified by SubPOS value.

Table 2.4: SUBPOS

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>Sentence boundary</td>
<td>C – numeral</td>
</tr>
<tr>
<td>,</td>
<td>Word krát (lit.: times)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>}</td>
<td>Conjunction subordinate (incl. aby, kdyby in all forms)</td>
<td>J – conjunction</td>
</tr>
<tr>
<td>:</td>
<td>Numeral, written using Roman numerals (XIV)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>:</td>
<td>Punctuation (except for the virtual sentence boundary word ###, which uses the 2.4 SUBPOS #)</td>
<td></td>
</tr>
<tr>
<td>=</td>
<td>Number written using digits</td>
<td>C – numeral</td>
</tr>
<tr>
<td>?</td>
<td>Numeral kolik (lit. how many/how much)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>@</td>
<td>Unrecognized word form</td>
<td>X – unknown</td>
</tr>
<tr>
<td>^</td>
<td>Conjunction (connecting main clauses, not subordinate)</td>
<td>J – conjunction</td>
</tr>
<tr>
<td>4</td>
<td>Relative/interrogative pronoun with adjectival declension of both types (soft and hard) (jaký, který, čí, ..., lit. what, which, whose, ...)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>5</td>
<td>The pronoun he in forms requested after any preposition (with prefix n-: něj, něho, ..., lit. him in various cases)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>6</td>
<td>Reflexive pronoun se in long forms (sebe, sobě, sebou, lit. myself / yourself / herself / himself in various cases; se is personless)</td>
<td>P – pronoun</td>
</tr>
</tbody>
</table>
2.2 Tag Structure

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Reflexive pronouns se (2.8 CASE = 4), si (2.8 CASE = 3), plus the same two forms with contracted -s: ses, sis (distinguished by 2.11 PERSON = 2; also number is singular only) This should be done somehow more consistently, virtually any word can have this contracted -s (cos, polívkus, ...)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>8</td>
<td>Possessive reflexive pronoun svoj (lit. my/your/her/his when the possessor is the subject of the sentence)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>9</td>
<td>Relative pronoun jenž, již, ... after a preposition (n–: něhož, niž, ..., lit. who)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>A</td>
<td>Adjective, general</td>
<td>A – adjective</td>
</tr>
<tr>
<td>B</td>
<td>Verb, present or future form</td>
<td>V – verb</td>
</tr>
<tr>
<td>C</td>
<td>Adjective, nominal (short, participial) form rád, schopen, ...</td>
<td>A – adjective</td>
</tr>
<tr>
<td>D</td>
<td>Pronoun, demonstrative (ten, onen, ..., lit. this, that, that ... over there, ...)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>E</td>
<td>Relative pronoun což (corresponding to English which in subordinate clauses referring to a part of the preceding text)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>F</td>
<td>Preposition, part of; never appears isolated, always in a phrase (nehledě (na), vzhledem (k), ..., lit. regardless, because of)</td>
<td>R – preposition</td>
</tr>
<tr>
<td>G</td>
<td>Adjective derived from present transgressive form of a verb Personal pronoun, clitical (short) form (mě, mi, ti, mu, ...); these forms are used in the second position in a clause (lit. me, you, her, him), even though some of them (mě) might be regularly used anywhere as well</td>
<td>A – adjective</td>
</tr>
<tr>
<td>H</td>
<td>Interjections</td>
<td>I – interjection</td>
</tr>
<tr>
<td>J</td>
<td>Relative pronoun jenž, již, ... not after a preposition (lit. who, whom)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>K</td>
<td>Relative/interrogative pronoun kdo (lit. who), incl. forms with affixes -ž and -s (affixes are distinguished by the category 2.16 VAR (for -ž) and 2.11 PERSON (for -s))</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>L</td>
<td>Pronoun, indefinite všechen, sám (lit. all, alone)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>M</td>
<td>Adjective derived from verbal past transgressive form</td>
<td>A – adjective</td>
</tr>
<tr>
<td>N</td>
<td>Noun (general)</td>
<td>N – noun</td>
</tr>
<tr>
<td>O</td>
<td>Pronoun svoj, nesvůj, tentam alone (lit. own self, not-in-mood, gone)</td>
<td>P – pronoun</td>
</tr>
</tbody>
</table>

continued on next page
<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Personal pronoun já, ty, on (lit. I, you, he) (incl. forms with the enclitic -s, e.g. tys, lit. you’re); gender position is used for third person to distinguish on/ona/ono (lit. he/she/it), and number for all three persons</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>Q</td>
<td>Pronoun relative/interrogative co, copak, cožpak (lit. what, isn’t-it-true-that)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>R</td>
<td>Preposition (general, without vocalization)</td>
<td>R – preposition</td>
</tr>
<tr>
<td>S</td>
<td>Pronoun possessive můj, tvůj, jeho (lit. my, your, his); gender position used for third person to distinguish jeho, její, jeho (lit. his, her, its), and number for all three pronouns</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>T</td>
<td>Particle</td>
<td>T – particle</td>
</tr>
<tr>
<td>U</td>
<td>Adjective possessive (with the masculine ending -ův as well as feminine -in)</td>
<td>A – adjective</td>
</tr>
<tr>
<td>V</td>
<td>Preposition (with vocalization -e or -u): (ve, pode, ku, ..., lit. in, under, to)</td>
<td>R – preposition</td>
</tr>
<tr>
<td>W</td>
<td>Pronoun negative (nic, nikdo, nijáký, žádný, ..., lit. nothing, nobody, not-worth-mentioning, no/none)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>X</td>
<td>(temporary) Word form recognized, but tag is missing in dictionary due to delays in (asynchronous) dictionary creation</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Pronoun relative/interrogative co as an enclitic (after a preposition) (oč, nač, zač, lit. about what, on/onto what, after/for what)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>Z</td>
<td>Pronoun indefinite (nějaký, některý, čikoli, cosi, ..., lit. some, anybody’s, something)</td>
<td>P – pronoun</td>
</tr>
<tr>
<td>a</td>
<td>Numeral, indefinite (mnoho, málo, tolik, několik, kdovíkolik, ..., lit. much/many, little/few, that much/many, some (number of), who-knows-how-much/many)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>b</td>
<td>Adverb (without a possibility to form negation and degrees of comparison, e.g. pozadu, naplocho, ..., lit. behind, flatly); i.e. both the 2.14 NEGATION as well as the 2.13 GRADE attributes in the same tag are marked by – (Not applicable)</td>
<td>D – adverb</td>
</tr>
<tr>
<td>c</td>
<td>Conditional (of the verb být (lit. to be) only) (by, bych, bys, bychom, byste, lit. would)</td>
<td>V – verb</td>
</tr>
<tr>
<td>d</td>
<td>Numeral, generic with adjectival declension (dvojí, desaterý, ..., lit. two-kinds/..., ten-...)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>e</td>
<td>Verb, transgressive present (endings -e/-č, -ic, -ice)</td>
<td>V – verb</td>
</tr>
</tbody>
</table>
2.2 Tag Structure

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>Verb, infinitive</td>
<td>V – verb</td>
</tr>
<tr>
<td>g</td>
<td>Adverb, forming negation (2.14 NEGATION set to A/N) and degrees of comparison 2.13 GRADE set to 1/2/3 (comparative/superlative), e.g. velký, zaříjmařý, ..., lit. big, interesting</td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>Numeral, generic; only jedny and nejedny (lit. one-kind/sort-of, not-only-one-kind/sort-of)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>i</td>
<td>Verb, imperative form</td>
<td>V – verb</td>
</tr>
<tr>
<td>j</td>
<td>Numeral, generic greater than or equal to 4 used as a syntactic noun (čtvero, desatero, ..., lit. four-kinds/sorts-of, ten-)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>k</td>
<td>Numeral, cardinal jeden, dva, tři, čtyři, půl, ... (lit. one, two, three, four); also sto and tisíc (lit. hundred, thousand) if noun declension is not used</td>
<td>C – numeral</td>
</tr>
<tr>
<td>m</td>
<td>Verb, past transgressive; also archaic present transgressive of perfective verbs (ex.: udělav, lit. (he-)having-done; arch. also udělaje (2.16 VAR = 4), lit. (he-)having-done)</td>
<td>V – verb</td>
</tr>
<tr>
<td>n</td>
<td>Numeral, cardinal greater than or equal to 5</td>
<td>C – numeral</td>
</tr>
<tr>
<td>o</td>
<td>Numeral, multiplicative indefinite (-krát, lit. (times): mnohokrát, tolikrát, ..., lit. many times, that many times)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>p</td>
<td>Verb, past participle, active (including forms with the enclitic –s, lit. ’re (are))</td>
<td>V – verb</td>
</tr>
<tr>
<td>q</td>
<td>Verb, past participle, active, with the enclitic -ř, lit. (perhaps) – could-you-imagine-that? or but-because- (both archaic)</td>
<td>V – verb</td>
</tr>
<tr>
<td>r</td>
<td>Numeral, ordinal (adjective declension without degrees of comparison)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>s</td>
<td>Verb, past participle, passive (including forms with the enclitic -s, lit. ’re (are))</td>
<td>V – verb</td>
</tr>
<tr>
<td>t</td>
<td>Verb, present or future tense, with the enclitic -ř, lit. (perhaps) – could-you-imagine-that? or but-because- (both archaic)</td>
<td>V – verb</td>
</tr>
<tr>
<td>u</td>
<td>Numeral, interrogative kolikrát, lit. how many times?</td>
<td>C – numeral</td>
</tr>
<tr>
<td>v</td>
<td>Numeral, multiplicative, definite (-krát, lit. times: pětkrát, ..., lit. five times)</td>
<td>C – numeral</td>
</tr>
</tbody>
</table>

continued on next page
CHAPTER 2. LEMMA AND TAG STRUCTURE

continued from previous page

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>Numeral, indefinite, adjectival declension (nejeden, tolikátý, ..., lit. not-only-one, so-many-times-repeated)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>y</td>
<td>Numeral, fraction ending at -ina; used as a noun (pětina, lit. one-fifth)</td>
<td>C – numeral</td>
</tr>
<tr>
<td>z</td>
<td>Numeral, interrogative kolikátý, lit. what (at-what-position-place-in-a-sequence)</td>
<td>C – numeral</td>
</tr>
</tbody>
</table>

Table 2.5: Obsolete values

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>!</td>
<td>Abbreviation used as an adverb</td>
</tr>
<tr>
<td>.</td>
<td>Abbreviation used as an adjective</td>
</tr>
<tr>
<td>~</td>
<td>Abbreviation used as a verb</td>
</tr>
<tr>
<td>;</td>
<td>Abbreviation used as a noun</td>
</tr>
<tr>
<td>3</td>
<td>Abbreviation used as a numeral</td>
</tr>
<tr>
<td>x</td>
<td>Abbreviation, part of speech unknown/indeterminable</td>
</tr>
</tbody>
</table>

2.2.1.3 3 – Gender

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Feminine</td>
</tr>
<tr>
<td>H</td>
<td>{F, N} – Feminine or Neuter</td>
</tr>
<tr>
<td>I</td>
<td>Masculine inanimate</td>
</tr>
<tr>
<td>M</td>
<td>Masculine animate</td>
</tr>
<tr>
<td>N</td>
<td>Neuter</td>
</tr>
<tr>
<td>Q</td>
<td>Feminine (with singular only) or Neuter (with plural only); used only with participles and nominal forms of adjectives</td>
</tr>
<tr>
<td>T</td>
<td>Masculine inanimate or Feminine (plural only); used only with participles and nominal forms of adjectives</td>
</tr>
<tr>
<td>X</td>
<td>Any</td>
</tr>
<tr>
<td>Y</td>
<td>{M, I} – Masculine (either animate or inanimate)</td>
</tr>
<tr>
<td>Z</td>
<td>{M, I, N} – Not feminine (i.e., Masculine animate/inanimate or Neuter); only for (some) pronoun forms and certain numerals</td>
</tr>
</tbody>
</table>
2.2 Tag Structure

2.2.1.4 4 – Number

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Dual, e.g. nohama</td>
</tr>
<tr>
<td>P</td>
<td>Plural, e.g. nohami</td>
</tr>
<tr>
<td>S</td>
<td>Singular, e.g. noha</td>
</tr>
<tr>
<td>W</td>
<td>Singular for feminine gender, plural with neuter; can only appear in participle or nominal adjective form with gender value Q</td>
</tr>
<tr>
<td>X</td>
<td>Any</td>
</tr>
</tbody>
</table>

2.2.1.5 5 – Case

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nominative, e.g. žena</td>
</tr>
<tr>
<td>2</td>
<td>Genitive, e.g. ženy</td>
</tr>
<tr>
<td>3</td>
<td>Dative, e.g. ženě</td>
</tr>
<tr>
<td>4</td>
<td>Accusative, e.g. ženu</td>
</tr>
<tr>
<td>5</td>
<td>Vocative, e.g. ženo</td>
</tr>
<tr>
<td>6</td>
<td>Locative, e.g. ženě</td>
</tr>
<tr>
<td>7</td>
<td>Instrumental, e.g. ženou</td>
</tr>
<tr>
<td>X</td>
<td>Any</td>
</tr>
</tbody>
</table>

2.2.1.6 6 – Possessor’s Gender

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Feminine, e.g. matčin, její</td>
</tr>
<tr>
<td>M</td>
<td>Masculine animate (adjectives only), e.g. otců</td>
</tr>
<tr>
<td>X</td>
<td>Any</td>
</tr>
<tr>
<td>Z</td>
<td>{M, I, N} – Not feminine, e.g. jeho</td>
</tr>
</tbody>
</table>

2.2.1.7 7 – Possessor’s Number

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Plural, e.g. nás</td>
</tr>
<tr>
<td>S</td>
<td>Singular, e.g. můj</td>
</tr>
</tbody>
</table>
2.2.1.8 8 – Person

Table 2.11: PERSON

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1st person, e.g. <em>pišu, pišeme</em></td>
</tr>
<tr>
<td>2</td>
<td>2nd person, e.g. <em>pišes, pišete</em></td>
</tr>
<tr>
<td>3</td>
<td>3rd person, e.g. <em>piše, pišou</em></td>
</tr>
<tr>
<td>X</td>
<td>Any person</td>
</tr>
</tbody>
</table>

2.2.1.9 9 – Tense

Table 2.12: TENSE

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Future</td>
</tr>
<tr>
<td>H</td>
<td>{R, P} – Past or Present</td>
</tr>
<tr>
<td>P</td>
<td>Present</td>
</tr>
<tr>
<td>R</td>
<td>Past</td>
</tr>
<tr>
<td>X</td>
<td>Any</td>
</tr>
</tbody>
</table>

2.2.1.10 10 – Degree of Comparison

Table 2.13: GRADE

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive, e.g. <em>velký</em></td>
</tr>
<tr>
<td>2</td>
<td>Comparative, e.g. <em>větší</em></td>
</tr>
<tr>
<td>3</td>
<td>Superlative, e.g. <em>největší</em></td>
</tr>
</tbody>
</table>

2.2.1.11 11 – Negation

Table 2.14: NEGATION

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Affirmative (not negated), e.g. <em>možný</em></td>
</tr>
<tr>
<td>N</td>
<td>Negated, e.g. <em>nemožný</em></td>
</tr>
</tbody>
</table>
2.2 Tag Structure

2.2.12 12 – Voice

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Active, e.g. <em>pišící</em></td>
</tr>
<tr>
<td>P</td>
<td>Passive, e.g. <em>psaný</em></td>
</tr>
</tbody>
</table>

2.2.13 15 – Variant

Table 2.16: VAR

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Basic variant, standard contemporary style; also used for standard forms</td>
</tr>
<tr>
<td></td>
<td>allowed for use in writing by the Czech Standard Orthography Rules</td>
</tr>
<tr>
<td></td>
<td>despite being marked there as colloquial</td>
</tr>
<tr>
<td>1</td>
<td>Variant, second most used (less frequent), still standard</td>
</tr>
<tr>
<td>2</td>
<td>Variant, rarely used, bookish, or archaic</td>
</tr>
<tr>
<td>3</td>
<td>Very archaic, also archaic + colloquial</td>
</tr>
<tr>
<td>4</td>
<td>Very archaic or bookish, but standard at the time</td>
</tr>
<tr>
<td>5</td>
<td>Colloquial, but (almost) tolerated even in public</td>
</tr>
<tr>
<td>6</td>
<td>Colloquial (standard in spoken Czech)</td>
</tr>
<tr>
<td>7</td>
<td>Colloquial (standard in spoken Czech), less frequent variant</td>
</tr>
<tr>
<td>8</td>
<td>Abbreviations</td>
</tr>
<tr>
<td>9</td>
<td>Special uses, e.g. personal pronouns after prepositions etc.</td>
</tr>
</tbody>
</table>

2.2.2 Compact tags

For most (but not all cases) just omit the dashes from positional tags. For more information, see

2.2.3 Informal abbreviations

In certain cases (including some places in this manual), the following tag abbreviations are used. Most of them are self-evident (dashes and rarely used fields dropped), as you can see in the following list:

Ngnc – noun; NFS1 = NNFS1-----A-----
Aagnc – adjective; AAXXX = AAXXX-----1A-----
Db – adverb; Db = Db---------------
Dg – adverb; Dg = Dg--------1A-----
Dgd – adverb; Dga2 = Dg--------2A-----
J* – conjunction; J^ = J^---------------

\(^2\text{http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/compact_tags.pdf}\)
CHAPTER 2. LEMMA AND TAG STRUCTURE

$J_-$, – conjunction; $J_\sim = J_-$, ----------------

$Rc, RRc$ – preposition, $RR7 = RR--7----------$

$RVc$ – vocalized preposition, $RV7 = RV--7----------$

$TT$ – particle; $TT = TT--------------$

$Ng-8, NNgXX-8$ – noun abbreviation; $NFXX-8 = NNFXX------A----8$

$AX-8, AAXXX-8$ – adjective abbreviation; $AAXXX-8 = AAXXX-----1A----8$

$Db-8$ – adverb abbreviation; $Db-8 = Db--------------8$

$Rc-8, RRc-8$ – preposition abbreviation; $RR7-8 = RR--7----------8$
Chapter 3

Names

Proper names (either directly or the lemmas they consist of) have suffixes marking the category of that name:

G – geographical name: Praha, Ústí nad Labem
Y – given (first) name, formerly used as default: Petr, John
S – surname (last name): Dvořák, Zelený, Agassi, Bush
E – name of a nationality: Čech, Kolumbijec
R – name of a product: Tatra (the car),
K – name of a company: Tatra (the company)
m – default – names of mines, stadiums, guerilla bases, etc; also used for functional words in names.

The lemma should start with upper case if the word is always in upper-case in names (Tatra is always in uppercase, but banka not).

Keeping this categorization in the same level as lemmas is quite unsustainable and very unsuitable.

1. In theory every word can occur in any category. For example, new in New York (G), New Jersey Devils (sport club – K), New Jersey Devils cards (product – R), etc. Because it would explode the lexicon, usually common words (besides new and alike, all functional words) have only two lemmas: one for common words and one for all names (using default category m). But such approach is highly unsystematic and works only for a small corpus.

2. But even then, the system is not used consistently – some functional words even do not have the above mentioned two versions (normal and m). For example, nad in Ústí nad Labem, should be nad-2_m, but there is no such a lemma. Similarly a in a.s. should be annotated differently when part of a name of some company and when not, but it is not.

3. Moreover lemmas having different categories are formally not connected, e.g. if you see Martin-3_K, you do not know if it is derived from Martin-1_Y (Martin + Martin, s.r.o) or Martin-2_G (DS Martin, a.s.).

Proposed solution:

- The G, K, etc. categories should be independent of the morphology, and should be assigned to phrases on a different level. This would also require some enhancement to the annotation tool DA.
- Only the words that always (i.e. >90% or so) belong to some category would marked with that category (e.g. new has no special suffix, England has G).
- Personal name would be annotated as separate lemmas Petr Pánvička – Pánvička_S not (pánvička)_s.
• Names containing name of a person, where the original link is not perceived (usually geographical names that do not contain possessive construction) have separate entries. N.B. that current guidelines require all the following words to be annotated with lemmas containing G (incl. ostrov, and úžina!)

Columbus (town in Ohio) – Columbus-2;G not (Columbus-1;S);
Martin (town in Slovakia)– Martin-2;G not (Martin-1;S);
Beringova úžina – (Beringuš-2;S´ užina), not Beringuš-2;G úžina-2;G
Ostrov Sergeje Kirova – (ostrov Sergej-2;Y Kirov-1;S);
Kirov (town in Russia) – Kirov-2;G not (Kirov-1;S);

3.1 Personal names

Some names are sometimes declined, sometimes not (Bill – o Bill Clintonovi, o Billu Clintonovi, o Billovi). The tag for nondeclined form is NgXXA.

3.1.1 von, van, etc.

For names (e.g Ludwig van Beethoven) the van, etc. phrase is perceived as a surname – annotate it that way. For other it is still perceived as geographical name (e.g Kryštof Harant z Polžic a Bezdužic). Of course the borderline is fuzzy.

Examples:

Ludwig van Beethoven – Ludwig-2;Y van-2; tˇ-(v_hol.;jménech) Beethoven-2;S
Vincent van Gogh – Vincent-2;Y van-2; tˇ-(v_hol.;jménech) Gogh-2;S
Kryštof Harant z Polžic a Bezdužic – Kryštof-2;Y Harant-2;S z-1 Polžice-2;G a-1 Bezdužice-2;G
Brigida z Háje – Brigida-2;Y z-1 Háj-2;G

3.1.2 Chinese names

Usage

The surname precedes the given name. In most cases, the whole name is used (not just the family name). The thing is complicated by the fact, that many Chinese living abroad often change the order of their name or use their given name as a surname, etc. The discussion below can help you to determine, which part of a name is the given name and which part is the surname. If you are in doubt annotate them all as given names (Y).

That was the original recommendation, but probably annotating them as S would be better, because they are often used that way (You can say Clinton for Bill Clinton, but you cannot say Po for Po Li).

Surnames

There are relatively few surnames in China (200 most common surnames account for >96% of all surnames). Most of them consist of one syllable (Wang, Li, Chen, etc.) Only few surnames consist of two syllables (Ou-yang, Mo-qi, Si-ma, Pu-yang). Married women do not get their husband’s surname.
Given names
Mostly two syllables, often connected with a dash (however sometimes separated by a space). Some can be widely used, some can be unique. Often it is impossible to say whether it is a name of a male or a female. The second syllable is usually used in informal addressing. The first syllable can be shared by all siblings. In traditional China a person had several given names during his/her life.

Most common Chinese surnames (in Pinyin):
Cai, Ceng-Zeng, Chen, Chen-Shen, Deng, Gao, Guo, He, Hu, Huang, Li, Liang, Lin, Lü, Ma, She, Sun, Tang, Wang, Wu, Xie, Xu, Yang, Ye, Zhang, Zhao, Zheng, Zhu

Links
http://www.wlu.edu/hhill/names.html – Chinese names explained

3.1.3 Korean names
Korean names behave similarly as Chinese names. Surname precedes given name. Given name of most Koreans consists of two parts, in Latin alphabet often connected with a dash. Most common Korean surnames are (45% of the population): Kim, Lee (often spelled as Rhee, Yi or Li), Park.

Examples:
Yang Sung-jin – S: Yang, Y: Sung, jin
Yang Sungjin – S: Yang, Y: Sungjin
Kim Il-Sung (former dictator of North Korea) – S: Kim, Y: Ir, Sen
Kim Ir-Sen (= Kim Il-Sung) - S: Kim, Y: Il, Sung
He Wung – S: He, Y: Wung

3.1.4 Foreignized Czech names
Sometimes you can encounter names that are Czech in their origin, but are somehow altered to fit other languages (diacritics is omitted, female and male surnames are the same – e.g. Judy Sedivy).

Use the following guidelines to decide the lemma and tag for such a name:

- a name that does not distinguish female and male variant, should have just one lemma and three different tags (gender M, F, X)
  
  Peter Janda – Janda S + NNMXX------A----- or NNMS1------A-----
  Jane Janda – Janda S + NNFX------A-----
  Jane a Peter Janda – Janda S + NNXXXX------A-----

- a name that has the same spelling as in Czech, should use the Czech lemma Jane Janda
  – Janda S + NNFX------A-----

- a name with altered spelling has its own lemma (with ,t suffix) Judy Sedivy – Sedivy, t
  + NNFX------A-----

1If {M,F} gender is introduced, the tag NN{M,F}XX------A----- should be used.
CHAPTER 3. NAMES

3.2 Compound names (names consisting of other names)

All lemmas of autosemantic words in compound names must have the category determined by the whole name (e.g. K, R). The lemmas of functional words contain default type category (m).

The problem is that a name of one type can occur as part of a name of a different type:

New England – G
New England Association of Chemistry Teachers – K
New England Association of Chemistry Teachers Journal – R

England is G noun in the first, K adjective in the second and R adj. in the third name.

If the lemma of the category you need does not exist and you have to insert a new one, do not care about numbering of lemmas, somebody else will do it (it would impossible to ensure that the numbers were unique across all annotators). That means, if there is other lemma having just different category (e.g. there is England>G available, but you need England>R), just change the category label.

Using the above-proposed separation of morphology and name categorization, the New England example would be annotated quite easily (only England is marked by a category (G) by the morphological analyzer, the rest is done by some other kind of tool):

(new England>G)c, (((new England>G)c association of chemistry teacher)k, (((new England>G)c association of chemistry teacher)k journal)r

If the annotator did not recognize the components of the name (e.g. it is in Burmese), (s)he would annotate just the highest level.

The categorization is sometimes quite tricky – you do not know, whether to consider a phrase a name or a name plus normal word:

Nobelova nadace – Nobelúv>K nadace>K 3
Nobelův stůl (e.g. in a museum) – Nobelúv>S stůl
Nobelova cena – hard to say (m vs. normal), decided: Nobelúv>S cena.

Examples:

Brownův pohyb – Brownův>S
Cena J. Debrau – Debrau>S cena
Mérieuxův ústav – Mérieuxúv>K ústav (Should be ústav>K but is not)
Divadlo J. Grossmana – divadlo>K J-4:B>K Grossman>K
příloha Kolumbus (in Lidové noviny) – Kolumbus>m
v Dobrovského ulici nejezdí ... – Dobrovský>G
v Dobrovského nejezdí ... – Dobrovský>G
poliklinika Dobrovského (unofficial, it is located in D. Street) – Dobrovský>G

Using the separation of morphology and name categorization, this is quite easy:

Nobelova nadace – (Nobelúv>S nadace)k
Nobelův stůl (e.g. in a museum) – Nobelúv>S stůl
Nobelova cena – easy to say: (Nobelúv>S cena)m.

2See the beginning of 3.
3The lemmas have different numbers (e.g. Nobelúv-1>S, Nobelúv-2>K).
3.3 Horses, DJ’s etc.

Horses have all kind of names (e.g. Vinná rěva, Deprivace, He Shall Reign, La Paloma Monitor, Frýdlant, Gold End, Lučina, Green Peace, Areál, First, Bounty), and quite often you do not know if it is female or male (sometimes even female like names belong to a male horse). One clue is, that in an Oak (a horse contest type), all horses are young mares – females.

In PDT 1.0 the names of horses where mostly not annotated correctly – simply any available name was selected (Otherwise, a new lemma with category Y would have to be inserted in each case: e.g. Deprivace would be Deprivace,Y, annotated as deprivace, He Shall Reign annotated as normal English phrase: he_to, shall_to, reign_to).

In our opinion, if the Y category were independent of the lemma, the horse name should be annotated correctly.

Similar problem is with the names of musical groups and DJ’s. For famous groups and DJ’s enter separate lemmas, for others use normal available lemmas.

3.4 Sport clubs, etc.

Name of the town in the club name: if only the town is noted, it is annotated as a geographic name (G), if the whole name of the club is noted, it is annotated as an institution (K). It is analogous to countries. (Česko vs. Německo are annotated as G)

Examples:

Cheb vs. Plzeň – Cheb,G Plzeň,G
SKP Union Cheb vs. Plzeň – SKP,B,K Union,K4 Cheb,K Plzeň,G

Of course, it can be a problem to know it with foreign clubs. If you do not know, annotate it as an institution (K).

Examples:

Chelsea – part of London, UK
Chelsea – Chelsea,G
Chelsea FC – Chelsea,K FC,L,B,K,w(…)
Ferencvaros – part of Budapest, Hungary
Ferencvaros – Ferencvaros,G

4In PDT 1.0, the lemma is Union, but it should Union,K
CHAPTER 3. NAMES

Ferencvaros TC – Ferencvaros₂₆G₆K₆B₂K₆
Sparta – Sparta₂₆G₆K₆
Sparta Praha – Sparta₂₆G₆K₆Praha₆K₆
Viktorie Žižkov – Viktorie₂₃G₃K₁(jméno_sport.klubu)Žižkov₁G₃K₁
Udinese – Udinese₅G₅K₅₊NNNXX-----₈A-----

It is the adjective of Udine (town in NE Italy), the official name of the football club is Udinese Calcio (calcio = football). However in Czech, the name is perceived as a noun and as the name of that club, therefore it is probably better to use it in that way:

To determine, whether something is a name of a town or a club, you can try to find that name on a map (e.g. http://www.expedia.com/pub/agent.dll?qscr=mmfn) and also find the club (e.g. http://www.soccerage.com).

Using the above-proposed ⁵ separation of morphology and name categorization, this looks much more consistent:

Cheb vs. Plzeň – (Cheb₂G₆K₆)₁(Penze₂G₆K₆)₁
SKP Union Cheb vs. Plzeň – (SKP₂B₆K₆UnionG₆K₆Cheb₂G₆K₆)₁(Penze₂G₆K₆)₁
Ferencvaros – (Ferencvaros₂G₆K₆)₁
Ferencvaros TC – (Ferencvaros₂G₆TC-₆B₆K₆)₁
Chelsea – (Chelsea₂G₆K₆)₁
Chelsea FC – (Chelsea₂G₆FC-₁B₆K₆w₆(...)₁)
Viktorie Žižkov – (Viktorie₂₂G₂K₂(jméno_sport.klubu)Žižkov₂G₂K₂)
Sparta – Sparta₂₂G₂K₂
Sparta Praha – (Sparta₂₂G₂PrahaG₂K₂)₁
Udinese – (Udinese₂₅G₅t₁)₁
Udinese Calcio – (Udinese₂₅G₅t calcio₂G₅t₁)₁

The name of the sport club often contains some abbreviation. Some are common and present in the analyzer’s lexicon (e.g. FC, AC) some are quite unusual (e.g. EV, ERC, EC, ERC, EG, VS, AS). If they are not present in the lexicon, entering them, suffixing the lemma by _B₆K₆w and using NNNXX-----₈A----- as tag,

3.5 Other

Insisting on inclusion of name categories (K, R, etc.), implies explosion of number of lemmas. We follow each examples section by analogous examples using the above- proposed separation of morphology and name categorization (see 3.2).

3.5.1 Geographical names

Streets

We suppose that the word ulice, etc. is always present, even if elided on the surface.

Examples:

Dlouhá – dlouhýG₆₊AAFS₁-----₁₈A-----
Dlouhá ulice – dlouhýG₆₊AAFS₁-----₁₈A-----ulice₊NNFS₁-----₈A-----
Palackého, Dobrovského, etc. – PalackýG₆₊DobrovskýG₆₊NNMS₂-----₈A-----

⁵See the beginning of the 3.
Examples:

\[
\begin{align*}
&\text{Dlouhá} - (\text{dlouhý})_G + AAFS1----1A---- \\
&\text{Dlouhá ulice} - (\text{dlouhý ulice})_G; or (\text{dlouhý})_G; ulice + AAFS1----1A---- NNFS1----A--- \\
&\text{Palackého, Dobrovského, etc.} - (\text{Palacký}_S)_G; (\text{Dobrovský}_S)_G + NNMS2------A----
\end{align*}
\]

Town names
Words in one-word names consisting that were originally adjectives are annotated as nouns.

Examples:

\[
\begin{align*}
&\text{Hluboká} - \text{Hluboká}_G + NFS1 \\
&\text{Dobrá Voda} - (\text{dobrá}_G + AFS1 Voda)_G; (\text{Součást, názvu, Odolena, Voda}) + NFS1 \\
&\text{Ohrada u Hluboké} - \text{Ohrada}_G + NFS1 u_\text{m} + RR2 \text{Hluboká}_G + NFS2
\end{align*}
\]

3.5.2 Initials

A separate character for aggregate gender \{M, F\} would be good (for initials following a letter in newspaper, an initial before a foreign last name, foreign names, etc.).

3.5.3 Institutions, companies

This category contains for example companies, foundations, shops, clubs, sport clubs, restaurants, etc. All autosemantic words in names of restaurants have lemmas with K. The exceptions are functional words that are annotated as default type (m).

Examples:

\[
\begin{align*}
&\text{Porcela Plus: Plus-3 + TT} \\
&\text{Bar Viola} - \text{bar-2}_K, \text{Viola-2}_K \\
&\text{U Medvídka} - u_\text{m}, \text{medvídek-2}_K \\
&\text{La cambusa} - \text{Le-1}_m, t, (\text{Franc., člen jako, souč., jmen., a, názvů})^7, \text{cambusa}_K, t \\
&\text{Restaurant HaPi} - \text{restaurant-2}_K, \text{HaPi}_K
\end{align*}
\]

\footnote{Frequent names of towns and names when POS changes, have separate entries. Therefore not (hluboká)_G.}

\footnote{In the current morphological lexicon, the m is missing.}
CHAPTER 3. NAMES

restaurace Jin Jiang – restaurace-1, jin-2K, jiang-2K,t
Franouzská restaurace v Obecním domě – francouzský-2K, restaurace-2K, v-2m obecní-2K
dům-2K
Hospůdka U vylitýho mrože – hospůdka-2K u-2m vylitý-2K mrož-2K

Examples:

Bar Viola – (bar, Viola-1)K or (bar, viola)-1K (select anyone, if you do not know the orig.)
U Medvidku – (u-1, medvidek)-1K
La cambusa – (le-1, la cambusa)-1K
Restaurant HaPi – (restaurant, HaPi)-1K
Čínská restaurace Jin Jiang – (čínský, restaurace, jin, jiang)-1K
restaurace Jin Jiang – (restaurace, jin, jiang)-1K
Francouzská restaurace v Obecním domě – (francouzský restaurace, v-1 (obecní dům))-1K
Hospůdka U vylitýho mrože – (hospůdka, u-1, vylitý, mrož)-1K

3.5.4 Sporting and other events

All events should receive special lemmas with m. However, if it is registered as a company
and used in that meaning, then it should be K. If not certain use m.

Examples: 8

Paris Indoor – (Paris-2m,t Indoor-1m,t + NNNXX------A------)
US Open – (US-3B,t Open-1m Open-1m 1a------8 Open-1m,t Open-1m + AAXXX------1A------9
akce Stop milion – stop-1m milion-1000000,m

Pohár mistrů – pohár-1m mistr-1m
Mistrovství světa – mistrovství-1m svět-1m

Examples:

Paris Indoor – (Paris-2m,G,t Indoor-1m,t)m
US Open – (US-2B,l Open-1m,(americký) Open-1m,l)m
akce Stop milion – (stop-1 milión-1000000)m
Pohár mistrů – (pohár mistr)-m
Mistrovství světa – (mistrovství svět)-m

8Many of these entries are not in the lexicon, therefore the actual numbers can be different once it is there.
See note in 2.1, e.g. mistrovství: mistrovství-1, mistrovství-2,m, mistrovství-3,R, etc.
7We think, it is perceived as noun, probably inanimate, in Czech.
3.5 Other

3.5.5 Televisions

Generally televisions are annotated as institutions (K). Only, if a company runs several channels, then the channels are annotated as products (R); but it is currently used only with Czech(oslovak) public television (ČT1, ČT2 and F1).

Examples:

ČT – ČT_B_K
ČT1 – ČT1_B_R
Nova – Nova_K
NBC – NBC-4_B_K
CNN – CNN-1_B_K_y_b_t

3.5.6 News, Magazines

All autosemantic word in names of news or magazines have lemmas with R. Currently, some of the newspapers are in the lexicon as institutions (e.g. Sme), this is not correct. Foreign names are often used as in plural, even if in the original there are in singular.

Examples:

Sme – Sme_R_{noviny} + NNXX
Zeitung – Zeitung-1_R_{soucí_názvu_něm_novin} + NFPX or NISX

3.5.7 Song names, etc.

Names of songs, TV programs etc. are annotated as normal words. The only reason is practical – it would cause explosion of the lexicon. If the categories and morphology are separated (see beginning of 3), these items can be annotated as R or m.
Chapter 4

Abbreviations

For discussion about inserting abbreviation not present in the morphological lexicon, see 10

4.1 Gender

Abbreviations can be used with different genders (e.g. ODS – feminine (strana) or neuter). Any abbreviation can have neuter gender. If the gender cannot be disambiguated by the context, use the gender used elsewhere in article. If the author mixes genders or there are no disambiguating contexts, use the gender inherent gender of the abbreviation. In Czech, is usually easy to determine – it is the gender of the head of unabbreviated equivalent (e.g. ODS – strana → f). With foreign abbreviations it is much more problematic, different people use different genders (e.g. because of different translation). If you are not certain which of the gender is most widely used, use the default neutrum.

Examples:

UK – F (univerzita)
FBI (Federal Bureau of Investigation) – I (úřad), N (default or byro), F (probably á la CIA), MP (pl., referring to the members of the FBI)
CIA (Central Intelligence Agency) – F (agentura)

4.2 Normal abbreviations

Normal abbreviations have sometimes as a lemma the abbreviation (and sometimes the original unabbreviated word. Usually the former method is used for abbreviation that are more common then the unabbreviated word (and for abbreviation of multi word expressions). But it is not always true.

For discussion about determining the gender of an abbreviation, see 4.1

Examples:

např.: například B + Db----------8
P.S.: post-2 B t (lat. po např. P.S.) + RR--X----------8
scriptum: B t L (př. P.S.) + NNNX-----A---8
n.L.: nad-1 B 1 + RR--7----------8, Labe B G + NNNS7-----A---8

1Should be nad-2 B m, but is not.
4.3 Isolated letters

**Note:** The following is still not official.

Isolated letters (e.g. *A-konto*) are handled as abbreviations. The only exception is if they are not in the name (*zápas skupiny B*). Many of the annotations suggested below are still not offered by the morphological analyzer. Moreover, sometimes the morphological analyzer is constrained to offer appropriate lemma and tag only if the letter is followed by a dot. Should be repaired.

You have to select (or insert) the lemma according to the semantic category:

- **K-0**: first (and most middle) names
- **K-4**: names of institutions
- **K-5**: geographical names
- **K-6**: names of products
- **K-7**: other names (sporting events, etc)
- **K-9**: last (and some middle) names
- **k-8**: other abbreviations (not names)
- **k-3**: other letters (not abbreviations, not in names)

Frequent abbreviations have their own lemmas, for example *V* – *V-1’volt* or *k*: *ABC k.s.*

Tag selection (or insertion):

- noun: gender is known: NNgXX—–A—8 (g ∈ {MF, IN})
- noun: gender is unknown: NNXXX----A----8
- adjective: AAXXX----1A----8 or AAgXX----1A----8
- others: X0---------------1 (variant of X0------------- for one letter words)

**Examples:**

A: *A-mužstvo* – a-3ˇ (označení pomocí písmene) + AAXXX----1A----8

d: odst. 1 písm. d) – d-3ˇ (označení pomocí písmene) + NNNXX----A----8

A: *A konto* (or *A-konto*) – A-6ˇ B-6ˇ R + AAXXX----1A----8

a: *ABC a.s.* – akciový B-8ˇ (jen akciová společnost) + AAXXX----1A----8

s: *na s. 128 – strana 4ˇ B-8ˇ (v knize, rukopise,...) + NNFXX----A----8

1. It is hard to decide, whether an isolated letter is an abbreviation or a label using a letter (e.g. a-3). For example, *B* in *B-konto* can be from *bežný*, but *A* in *A-konto* probably means better than *B*. Maybe not, maybe yes, who knows. What is important, we mostly annotate texts written by people that do not know. Therefore it would be reasonable to merge these two possibilities together. Maybe annotate all single letters as abbreviation, the possible exception could be labels of paragraphs and cases (odst. 1 písm. d) or za a).
2. Letters in similar configuration as nouns in noun cluster should be treated as nouns – they should be annotated as nouns. See also 6.1.1.

3. The category of a name (K, G, R, etc) and lemma selection should be orthogonal – see also 3.2.

Examples:

A: A-muˇ zstvo – (a-8(ost. zkratka nebo oznaˇ cen´ı) + NNNX------A----8
d: odst. 1 písm. d) – d-3(př. odst. a, za a) + NNNX------A----8
A: A konto (or A-konto) – (a-8(ost. zkratka nebo oznaˇ cen´ı)...)+NNXXX------A----8
a: ABC a.s. – (... akciovy_B.(jen akciovaˇspoleˇcnost))x + AAXXX---1A----8
s: na s. 128 – strana-4_B.(v_knize, rukopise,...) + NNFXX------A----8

4.4 RM-systém, samopal SA-58

An abbreviation preceding a noun is an adjective, an abbreviation following a noun is a noun. We would suggest to annotate them all as nouns (see 6.1.1). Does it mean that HIV in HIV virus and virus HIV have different POS.

Examples:

RM-systém – RM:B:K + AAXXX------1A----8 NNXXX------A----8
samopal SA-58 – SA-2:B:R + NNXXX------A----8
virus AH 3 B - AH-1:B + NNXXX------A----8
virus HIV- HIV:B:L:U_(lidsky_virus_zpúsobující AIDS) + NNXXX------A----8

4.5 Units of measurements

Units called after some males person (V – volt, A – ampér, etc.), have inanimate gender. However, units using degrees (°C, °F) have masculine animate gender, because the word stupně is always present (even if omitted in the written text). Absolute temperature uses as the unit called Kelvin (K) not degree of Kelvin. Therefore the unit has inanimate masculine gender. However, if the author uses it erroneously as degree, the tag as to be masculine animate.

Examples:

C: Ráno byly 3 °C. – Celsius:B – NNMX------A----8°
C: Ráno byly 3 C. (read as Ráno byly tři stupně Celsia) – Celsius:B – NNMX------A----8
K: Teplota 5000 K. – Celsius:B – NNMX------A----8°
K: Teplota 5000 °K. – Celsius:B – NNMX------A----8°

If the C character is preceded by some character trying to look as the degree symbol ° (eg. ·C, o C, O C), then you should mark it as an error – as lemma insert the degree symbol ° and as tag X@------------1. It should be converted into a punctuation mark.

On Czech keyboards usually Shift+<key-on-the-left-from-1>, followed by Space. Or on any keyboard Alt+0176.
4.6 Authors abbreviations

The author’s name abbreviations used in newspapers (e.g. Ber, mas, jst, ...) have lemma as the form +99_B_S and tag NNXXX------A----8. There is X for gender because usually we do not know it. If the {M,F} gender is introduced, it should be used here. These abbreviations are not present in the lexicon, therefore you have to insert them.

Examples:

ač: PRAHA (ČTK, ač) Problém Gabčíkova ... – ač-99_B_S – NNXXX------A----8
gap: DUKOVANY (gap) Na základě posudku ... – gap-99_B_SNNXXX------A----8

4.7 Academic titles

Titles distinguish genders – there has to be one lemma for men, and one lemma for women (JUDr-1: doktor práv) vs. JUDr-2: (doktorka práv)); to keep it consistent the masculine has number 1, the feminine has number 2. We think, the titles should have the same form for women and men. Just the tag should be different, with possibility to have X if the gender is not known (e.g. a letter subscribed as Dr. A. B.)
Chapter 5

Colloquial Czech

If an official alternative to the colloquial form exist, then the colloquial form has the same tag except a different variant (‘5’, ‘6’, ‘7’, ev. ‘3’ – see 2.2.1.13).

Examples:

\textit{které: stavení, které} – P4NP4----------5
\textit{Novákovic: Novákovic pes} – NovákůvěS_{(2)}-AUXXM----------6 \footnote{In PDT 1.0, this is sometimes obsoletely annotated as AUMS1M----------6 or NNXX---A---6}
\textit{takovejšma: takovyhle} – AAFP7------1A---6
\textit{hovadinama: hovadina} – NNFP7--------A---6
\textit{naši: pro naší atletiku (officially short: naši)} – můj_{(přívlast.)} – PSFS4-P1------6

5.1 Cos, jaks, kdys ...

We tagged these words as if they were without -s and added -9 at the end.

\begin{itemize}
  \item In our opinion it would be better to divide such an expression in two words (e.g. cos → co + být, analogous to abych → aby + být) and tag them like two normal words, just with some variant recognize it.
\end{itemize}

5.2 Suffix -é in plural of neuter

Should not be treated as misspelling, but annotated as (colloquial) variant of official -á forms (variant ‘5’).

Examples:

\textit{které: stavení, které} – P4NP4----------5
Chapter 6

Foreign words and phrases

General rule

1. For a longer phrase (or citations) in a foreign language, use morphology of that language (but distinguish genders M and I??) (Hence citation use).

2. For a single word or shorter phrase use Czech morphology. (Hence word use) The borderline is fuzzy, of course.

6.1 Part of speech

Many foreign words used in Czech sentence can have different part of speech than in their original language. Usually the hint is how it behaves in different context, if it is declined as a noun, if agrees with its head, etc.

All foreign lemmas have _t suffix.

It would be good to somehow distinguish foreign words in word or citation use.

6.1.1 English noun clusters

All nouns in attributive use are annotated as adjectives.

That’s quite problematic:

1. Virtually all English nouns can be used as attributes of other nouns
2. It is imported to Czech: *Staropramen Extraliga, Český Telecom Cup, etc.

We think, it should be annotated as two nouns.

6.1.2 Examples

V kostele XY zpívala Musica Bohemica.

Bohemica annotated as a noun; in Latin it is an adjective.

Reason: When the phrase is declined, Bohemica is declined as a noun (žena): pozvali Musicu Bohemicu, *pozvali Musicu Bohemicou

Annotation: Musica_t_K + NFS1A, BohemicaL_t_L_K NFS1A

To je trochu ad hoc.

hoc is annotated as a noun; in Latin it is an adverb.

Annotation: ad_t RRX, hoc_t NXXXA

In the following, the section headers refer to the categories of the foreign language.
CHAPTER 6. FOREIGN WORDS AND PHRASES

6.2 Articles

English *an* should be a form of *a*.

Articles merged with a preposition (fra *du*, ita *della*, deu *im*, *aufs*, zur) are treated as prepositions (??Split into two words??)

Arabic short words (##)/articles, ?prepositions) are treated as articles.

6.2.1 Citation use

Same as single words

Should distinguish gender, number, and/or case Therefore: TTgnc or AAgnc ??

6.2.2 Word use

Tag: TT----------------
Lemma: Usually the same as the form

Originally, we wanted to treat articles as adjectives. Forms having different gender, number and/or case, would have the same lemma (*der* for forms *der*, *die*, *das*, *des*, *dem*, *den*). The problem is that Czech does not respect the original categories (*nebezpečný* La Manche – *la* in French F, in Czech the phrase is I; *Los Angeles* – in Spanish pl, in Czech sg.)

6.2.3 Examples

*l* l-5, t (př, l’Arc, stažený_tvar.fr, členu) + TT
l’ L-10, (př, L’Aqua, stažený_tvar.fr, členu) + TT (should be m)
l la la-2_t + TT
il il-t, (it, len) + TT
as as_t + TT (Arabic)
al al_t + TT (Arabic)
el el_t + TT (Arabic)
della della_t + RRX (sometimes incorrectly annotated as AXXX)
am am_t + RRX

6.3 Nouns

6.3.1 Citation use

6.3.1.1 English

To keep it simple, number of English nouns is annotated in the same way in citation use as in word use. That means X is used instead S for nouns in singular. The difference is in X is used instead S for nouns in singular. The difference is in cases – in citation use, the case is always X, but word use it can sometimes be declined.

If the name categories and lemmas were independent, everything annotated as L-10 would become l-5 (see 3.2)
6.3.2 Word use

6.3.2.1 English

The nouns in singular in English have number annotated as X (English singular. is often used in Czech as plural). For nouns that are usually declined mark the case even if in base form, for nouns that are nondeclined mark it as X.

Examples:

flow: oba dva cash flow (oficiální i skutečný) ... – flow\textsubscript{t} – NNIXX------A-----
statement: v cash flow statementu ... – statement\textsubscript{t} – NNIS6------A-----
statement: Náš cash flow statement ... – statement\textsubscript{t} – NNIS1------A-----
flow: Náš cash flow ... – flow\textsubscript{t} – NNIXX------A-----
girl: Beatles: Girl – girl\textsubscript{t} – NNFXX------A-----
girls: A teď zahrajeme písničku Girls. – girl\textsubscript{t} – NNFXX------A-----

6.4 Verbs

6.4.1 Citation use

6.4.1.1 English verbs

The following tags are applied:

- Present non3sg: go VB-X---XP-AA---
- Present 3sg: goes VB-S---3P-AA---
- Imperative: go Vi-X---X--A-----
- Infinitive: go Vf--------A-----
- Past tense: went Vp-X---XR-AA---
- Passive participle gone Vs-X---XX-AP---

If it is hard to determine the base form usage, annotate it as infinitive. If it is hard to decide between past tense and passive participle, use past tense. In PDT 1.0, most of the verbs using base form were annotated using the default – infinitive.

Examples:

be: to be or not to be – be\textsubscript{t} (angl. být, v názvech apod.) – Vf--------A-----
do: Do it right now! – do\textsubscript{t} – Vi-X---X--A-----

6.4.2 Word use

Usually the tags and lemmas are the same as in citation use.

6.5 Slovak language

If a Slovak word has the same form as corresponding Czech one (e.g. prepositions), you should annotate it as if it were Czech. Otherwise it has to be annotated as any other foreign language.
Chapter 7

Errors

The text can contain errors. It is reasonable to correct some of them, preserving the original form. However, only low-level errors – spelling and morphology should be corrected (We do not want to correct Engels’ text into Heidegger’s). Never correct a colloquial form by an official one (e.g. zelené města → zelená města, bez noh* bez nohou), even if the analyzer does not know the form 1.

The errors have to be just marked, do not edit the file. Try to insert lemma and tag as if the form is correct, and use the DA support for marking errors – it inserts the text “(Chyba)” at the end of lemma or tag. If the lemma is correct, insert it after the tag, otherwise insert it after the lemma, if you do not know just insert it somewhere. If you want to add some comment, write it before the closing parenthesis, preceded by a dash (e.g. (Chyba-nad c by méla být čárka, ne hůček)). This convention makes it easy to find the errors automatically.

7.1 Characters

Sometimes, foreign characters had been be screwed (e.g. Françoise), and therefore the morphological analyzer did not recognize the whole word. Mark it as a lemma error (do not edit the file), it has to be corrected and run thru the analyzer once more. There is a problem with letters not contained in Latin 2, they should be replaced by corresponding characters without diacritics. In the future, Unicode (2 bytes or UTF) should be considered.

7.2 Separators, etc.

Sometimes, the text contains o or I as bullets or separators. They should be marked for deletion (Press L (delete) in the lemma or tag list).

---

1You have to insert a new lemma and/or tag – see 10 for more details.
Chapter 8

Hard to decide

8.1 až

až-1 + J,
2 až 3 (but not od 2 až do 3 – see až-3)
nabízí přibližení až přijetí
až-2 + J,
tak .. až: Nabízí se tak okatě, až je to hanba.
.. začnou pohybovat, až nakonec určí, že ..
Bylo mi 24, a byl jsem plný touhy se pomstít. Až jsem se ocitl před člověkem, který
dostal zabrať víc než já.

If omitted, the sentence stays grammatical. It is often possible replaceable by teprve.

Dostanete až 250 mil zdarma.
kam až: Kam až půjdeš?
Až on me přesvědčil, že tomu tak bude.

Modifies functional word (should be probably TT)
až + conj: Je geolog a až pak filozof
až + prep: z Brna až do Prahy (Cf. až-1)

8.2 jak

jak-1,„L_”(živočich) + NNMnC-----A-----
Obvious.
jak-2 + J,

1. Meaning že (cannot be replaced by jakpak)

Jak řekl M. Zeman, bude třeba ..
Jak ukazuje vývoj posledních let, je to ..
Jak známo, ...
Skutečnost, jak už to býva, byla trochu jiná. However, rarely it can be Db – depending on
the interpretation

Viděl, jak upadla.
Meaning Viděl, že upadla. – J,
CHAPTER 8. HARD TO DECIDE

Meaning
Viděl, jakým způsobem upadla. – Db
Kamera zabírá poslance, jak otvírají krabici

2. Time, meaning když, až, jakmile
Přijdu, (hned) jak budu hotově.
Hned jak budu moct, zavolám.

3. In comparison, meaning než, jako:
Byl větší jak oněk
rychlý jak vrtrěk

4. Condition (coll.), having the meaning jestliže, když
Jak budeš zlobit, nepůjdeš nikamě

Japonskému turistovi upadla lžička, jak chtěl zmáčknout spoušť foťáku.
Poslední šanci, jak se probojovat do .., bude..
Stát to měl spravovat zvláštním ministerstvem (jak tomu např. v Rakousku)

jak-2 + J
In the phrase jak ... tak ..., having the meaning of i...i . However cf. jak-3 2.
Byli tam jak odborníci, tak amatéři.

jak-3 + Db
Pronominal adverb

1. Interrogative – manner or extend (expr. jak pak).
Jak se jmenuješ?
Jak je to možné?
Sometimes expressing large extend (often in exclamations).
Jak ten čas leží
Jak (pak) by neš. Japa by ne.
Líbí se ti to? – A jak!

2. Relative – marks subordinative adverbial clause (mostly manner expressing comparison, often with tak – however cf. jak-2 + J+)
Jak řekli, tak udělali
jak dlouho, jak je možné (tak .., jak ..)
Jak si kdo ustele, tak si lehne

3. Relative (coll.) – meaning co, který
ten člověk, jak jsem ti o něm říkal

4. Indefinite
budě jak budě (the verb is repeated)
jak kdo, jak kdě, jak kdy,

?? Jak se kůže sama obnovuje, postupně vylučuje ..
?? Jak jsem chodil o berlich, tak jsem si zničil i druhé koleno.
8.3 málo

Similar to moc.

\[\text{málo-1} \text{ (málo + méně) + \text{málo peněz}) + \text{Ca-c-}}\]

It has to be modified (in the shallow syntax) by a noun in genitive. Has only two forms:

málo and mála (only in genitive).

Máme málo zájemců.

bez mála peněz

před málo lety

Je jen o málo důslednější. – but Je málo důsledný. is málo-3 (Dg)

Udělal to jako jeden z mála odborníků, ...

Udělal to jako jeden z mála. – ?? not modified by anything

Udělal to jako jeden z mála, co přišli.

\[\text{málo-2} \text{ (při mýlém co měl)} + \text{NNNc-} \text{A-}\]

vystačit s málem

vařit z mála

Děkuji. – Za málo. 

\[\text{málo-3} \text{ (málo + před jm., při byl málo důsledný)} + \text{Dg-} \text{DA-}\]

Málo mluví, hodně dělá.

Je málo důsledný.

Ve srovnání s loňskou sezónou je to velmi málo. – you can say méně.

Zdržím se jen málo.

8.4 moc

Similar to málo.

\[\text{moc-1} \text{ (nad někým; politická, vojenská; plná,...)}\]

Obvious.

převzít moc

moc proletariátu

udělám, co je v mé moci

měm moc

\[\text{moc-2} \text{ (mnoho něčeho (se subst. v gen.) + Ca-x-}}\]

Cannot be replaced by velmi. Can mean příliš, but is more colloquial. It has to be modified (in the shallow syntax) by a noun in genitive.

Má moc peněz.

Všeho moc škodi.

\[\text{moc-3} \text{ (velmi, ve spojení s adj., při moc hezká)} + \text{Db}\]

Can be replaced by velmi (except ellipses). Modifies an adjective, adverb or verb.

Je moc hezká.

Vím to moc dobře.

Moc se snažil.

Ve srovnání s loňském je to moc. – ellipse.
8.5 **proto**

proto-1 `(proto:,a,proto:,ale,proto,...) + J`

Coordinative conjunction expressing consequence (implication). Structure: reason → consequence. Replaceable by *tedy*. Usually *a proto* or *a ... proto*

Nesplnil úkol, *(a) proto nedostal odměnu.*

Každé proč má své proto.

Německo se začalo dusit, a rozhodlo se proto omezit ...

proto-2 `(dal.mu.co.proto,tak,proto!) + Db`


Udělal to proto, aby/že mu pomohl.

co proto: dát někomu co proto; dostat co proto

no proto: Říkal, že tam přece jen půjde – *No proto!* (Sometimes classified as a modal particle)

8.6 **svůj**

svůj-1 `(přívlast.) + P8gnc---------v`

Obvious.

svůj-2 `(být svůj) + AOgn----------v`

Problem with tags, analyzer probably needs update.

Vzít za své.

Víme své. Víme svoje.

8.7 **tak**

In general:

- replaceable by *a proto* ⇒ *J*
- replaceable by *tím způsobem, stejně, zrovna* ⇒ *Db*

**tak-2 + J**

Coordinative conjunction. If one of the clause is subordinative then *tak* has the meaning of an adverb: (Cf. *Bál se, tak si pískal*. – *J* vs. *Kdyby se bál, tak si pískal – *Db*)

1. A consequence — meaning *(a) proto, tedy*

*Bál se, *(a) tak si pískal,“*

Neudělali..., příspěvek tak budou muset vrátit.

Byly zakázány, a tak přitahovaly

Zmizí bariéry, a tak bude možné využívat ..

Zpozdila se, a tak musela běžet.

Jsou profici, tak ať se podle toho zařídí/

Počítá se s tím, že některé se sloučí, i tak bude třeba ..

2. A conjunction — in *jak – tak*

*tak-3 + Db*
1. Refering to something known, to other sentence, etc.

   tak – jak: Bylo to tak, jak jsem myslel.
   jak – tak: Jak řekli, tak udělali.
   Přesně tak.
   tak zvaný
   Ať je to tak nebo tak ...
   jen tak: Udělal to jen tak.
   tak tak: Stihl to (jen) tak tak.
   to: Stalo se tak při ..
   Tak se tehdy žilo

   Sub-Clause, tak Main-clause:
   Když – tak: Když jsem počítal já, tak mi vyšlo velké číslo.
   Pokud – tak: Pokud to není diskriminace, tak nevidím důvod ..
   Dokud se člověk raduje, tak je život pěkný.
   Kdyby – tak: Kdyby/Pokud by se bál, tak by si pískal.
   (Cf. Bál se, tak si pískal. – J^)

2. Expressing amount (usually large) of a property, etc.

   Kam tak rychle?
   tak jako: Je tak velký jako já.
   Zmizel z povědomí tak jako jeho pomník;
   Nabízí se tak okatě, až je to hanba.
   To je ale tak daleko ..
   tak vysoká; tak oslaben, že ..
   Buďte tak laskav.
   ani tak o ..., jako o ...: Nejde ani tak o mzdu, jako o ...
   přibližně: Dostane se na burzu asi tak třetí den od ..
   hned tak: Hned tak nepřijde. (koneckoců)
   odmítá to, stejně tak jako ...
   .. a zrovna tak hýzdit;
   tak jako tak
Chapter 9
Sólokapři

Hradec Králové
Králová-3G-2S -(Dvůr Králové) + NNF S2-----A----- It is hradec that belonged to králová

strana
na jedné straně ..., na druhé straně ...: druhý-1 (jiný), strana-1-2(v_prostoru) nerespektované ze strany Israele: strana-3-2(v_soudu, ..

stát
stane se ministrem: stát-2-2(něco_se_přihodilo)

s=to
být sto něco udělat lemma = sto-3-2(býtsto), znacka TT---------------

tudíž
always J’ ##Why is there the other possibility

vážit
vážit cestu – vážit na váze? nebo ctít někoho

vedení
Everything except elektrické vedení type, is considered as form of vedení-1

Examples:
pod vedením kamarádů – vedení-1-2(*7ěst-1)
vedení podniku – vedení-1-2(*7ěst-1)
čínské vedení – vedení-1-2(*7ěst-1)
elektrické vedení – vedení
CHAPTER 9. SÓLOKAPŘI

9.1 Date and time

\[ v + \]

A day – accusative (4): \( v \) sobotu, \( v \) neděli
A month – locative (6): \( v \) lednu, \( v \) září
An hour – accusative (4): \( v \) 4 hodiny, \( v \) 6 hodin
Ve dne – locative (6) – \( \text{NNIS6-----A---9} \) -special kind of locative that occurs only in this context (\( v \) noci is also in locative):
Month in a date – genitive (2): \( 25. \) září, \( 2. \) října

9.2 Numbers, numerals and quantifiers

An adjective modifying a quantified expression agrees in case with the noun not the numeral.

Examples:
\( za \) (gen) těch (gen) mizerných (acc) deset (gen) korun

\( \text{Deset} \) (nom) nejlepších (gen) sportovců (gen) ukázalo

\( 1x \)

Lemma: as form. Tag \( C\text{v-}--------- \)

Example
\( 1x - 1x + C\text{v-}--------- \)

\( 4x5 \)

Should be split into three parts. E.g. \( 4x5 \to 4, x, 5 \)

9.3 Hyphenated composites

If the hyphenated word ends with -o, and by a replacement of that -o by an adjective ending we obtain an adjective (normal or possessive), the lemma for the word is that adjective (e.g. česko-německý – česko \( \to \) český, Karlo-Ferdinandova – Karlo \( \to \) Karlův). Some word cannot be viewed as derived from adjectives, but rather from nouns (e.g. rap- jazzová – rap \( \to \) rap vs. rapovo-jazzová – rapovo \( \to \) rapový). However, the lemma of that noun cannot be used as a lemma for the hyphenated form, and a new lemma (having different number) has to be introduced.

That is extremely inconvenient (padlý na hlavu) – virtually any noun can be used in such a context, therefore for every noun, there have to be two lemmas – one for normal usage and one for hyphenated usage. We strongly suggest to allow any noun to have a hyphenated form in several variants – at least the bare base form and form ending in -o (variant ‘1’).

Examples:

\( \text{srbsko-černohorská} – \text{srbský} – \text{A2-}---------\text{A----} \)

\( \text{Univerzita Karlo-Ferdinandova} – \text{Karlov \_} \text{K\_*(3el)} – \text{A2-}---------\text{A----} \)

\( ^1\)Better: \( \text{Karlov \_Y\_*(3el)} – \text{A2-}---------\text{A----} \). See 3.2

52
9.3 Hyphenated composites

Univerzita Karel-Ferdinandova – Karel-2; K – A2 --------- A-----
rap-jazzová: rap-2 – A2--------- A-----
| Better: rap – A2--------- A-----
rapo-jazzová: rap-2 – A2--------- A-----
| Better: rap – A2--------- A----1
rapovo-jazzová: rapový – A2--------- A-----
Chapter 10

Insertion

If the possibilities offered by morphological analyzer are not suitable, you have to insert new lemma and/or tag. If you insert a new lemma, you have to ensure, that the lemma (lemma proper) you insert is not already used. That usually means adding unique numbers to distinguish lexical items having the same base form.

10.1 Possessive adjectives

Lemmas of possessive adjectives show how the get the noun they are derived from (see also 2.1.1). For example:

- *kardinálův* (*2) – remove two letters: *kardinál*
- *Karlův* → *Y* (*3el) – remove 3 characters, add "el": *Karel*
- *Martinův-1* → *Y* (*4-I*) – remove 4 characters, add "-1": *Martin-1*

Examples:

- *premiérův* (*2)
- *Sorosův* → *S* (*2)
- *chlapcův* (*3ec)
- *Švehlův* → *S* (*2a)
- *Máchův* → *S* (*2a)
- *Hlinkův-1* → *S* (*4a-I*)
- *Benderův-1* → *S* (*4-I*)

10.2 Words ending with -ismus, -izmus

The base form should use *-izmus* ending, the form using *-ismus* is treated as variant '1'. Currently still some entries do not follow this convention.

Examples:

- *mechanizmus: mechanizmus* – NNIS1------A------
- *mechanismus: mechanizmus* – NNIS1------A-----1
- *exhibicionismus: exhibicionismus* – NNIS1------A-----

---

1 The examples show the desired state, in the current version of morphological analyzer they are regarded as separate lexical items (they have different lemmas)
CHAPTER 10. INSERTION

exhibicionismus: exhibicionismus – NNIS1------A----1
nacionalizmus: nacionalizmu – NNIS1------A----
nacionalismus: nacionalizmus – NNIS1------A----1

10.3 Strange and unique things

Transcription of pronunciation
Lemma: as the form, tag: NNXXX
Examples:
  vyslovujeme “zpjev” – zpjev + NNXXX

Isolated morphemes
Lemma: as the form, tag: NNXXX
Examples:
  ...ve slovech končících na -ství pišeme...: ství + NNXXX

Geometry
We can meet an article about a geometric theme sometimes. It means, that there occur a lot of triangles ABC, abscissas (lines) PQ, RS, AB and so on in that article. It is necessary to create a new lemma ending 98 for every mentioned figure.

Chess codes
Lemma: The code + -1,w. Tag NNNXX------A----8 (neuter because pole is neuter)
Example
  Jh8 – Jh8-1,w + NNNXX------A----8

Crippled forms
Lemma: the same as the form + _t
Tag: normal if possible, otherwise NNXXX / AAXXX according to the POS
Examples:
  Waklaf Hafel – Waklaf_t + NNMS1, Hafel_t + NNMS1
  Gaptschikowo – Gaptschikowo_t + NNNS1
  v Gaptschikowo – Gaptschikowo_t + NNNXX
10.4 Other

This section contains especially examples of previously inserted lemmas/tag. Some of them are already in the dictionary, however they mainly serve as an inspiration, when inserting similar things.

**ad hoc**

*ad-xít + RRX, hoc-xít + N*

**pele-mele**

For example as a heading in a newspaper *pele + TT, mele + TT*

**zprostředkovací vs. zprostředkovat**

The morphology is rather shallow. It means, for example, that lemma for *zprostředkovací* is *zprostředkovací* (precisely *zprostředkovací*(“2t)), and *zprostředkovat* as it was in the past.
Chapter 11

Errors in PDT 1.0

- \textit{HaDivadlo\_Y} \rightarrow \textit{HaDivadlo\_K}
- Theaters – most of the theaters do not have K category (search for \textit{Divadlo} case sensitive): \textit{Divadlo v Celetně}, \textit{Divadlo Husa}, etc.
- S/NWS/1993/mf930701:105-p4s1 – los should be los-2
- \textit{které} should be 5 not corrected as an error. (look for \textit{spell}\_které)
- `<title id="S/NWS/1992/lnd92251:095-p1s1"> U5 should be with m (US-3\_B\_m\_t) not K.
- Novákovic, Perotovic, .. should be AUMS1M--------6, but it is either AUgncM——–6 or NNXX-----A----6 (e.g. S/NWS/1992/lnd92254:051-p5s8)
- S/NWS/1992/lnd92258:077-p91s1

<table>
<thead>
<tr>
<th>cash</th>
<th>flow</th>
<th>statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is</td>
<td>cash_t</td>
<td>flow_t</td>
</tr>
<tr>
<td></td>
<td>AAXXX----1A----</td>
<td>NNFXX------A----</td>
</tr>
<tr>
<td>Should</td>
<td>cash-2_t</td>
<td>flow-2_t</td>
</tr>
<tr>
<td>be</td>
<td>A2--------A----</td>
<td>AAXXX----1A----</td>
</tr>
<tr>
<td>Should</td>
<td>cash_t</td>
<td>flow_t</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>