MANUAL FOR MORPHOLOGICAL ANNOTATION

Jirka Hana and Anna Feldman

version 2008-12-07

Table of Contents

1 Basic Rules

1.1 Descriptive, not prescriptive tagging

Probably the most important rule of annotation is to annotate the words as the writer used them and not as they should have used them. If some writer is unfortunate enough to think that Victor Hugo was a woman and uses it that way, it must be annotated with feminine gender, including any agreeing adjectives or verbs.

The only exception are clear typos. Annotate it with the correct tag and mark it as a typo. For example, if Victor Hugo is used as a male in several paragraphs, and then in one sentence it a single character of the verb makes it a feminine verb, we can safely assume that this is a typo and not the language of the author. However, this has to be used with great care and if you are unsure whether something is a typo or not, assume it is not. Colloquial language should never be corrected!

1.2 Annotating words not phrases

Annotating compound phrases as unit is avoided nearly at all cost. While linguistically, it might make sense, it is hard to say where to put the boundary. So we do not analyze such expressions as one unit, but annotate each word separately. One day there will be another layer above the morphological layer which would consider such phrases to be a single unit (together with company names, etc).

Examples:

křížem krážem (Czech, crisscross) – two words
tem bolee (Russian, let alone) – tot+PDNS7, bolee+Dg-------2A----
as soon as possible – annotated as four words

Currently, for technical reasons, this applies even to words separated by a punctuation. We assume that in the future, some of these cases will be treated as a single unit. However, at present, do *not* mark them as segmentation errors.

Examples:

O'Brien – annotated as *O + ' + Brien*
Yahoo! – annotated as *Yahoo + !*
John's – annotated as John *+ ' + s*
you're – annotated as *you + ' + are*
English-German – annotated as *English + - + German*
*čto-nibud' (*Russian *something)* – annotated as *čto + - + nibud'*

The only exceptions are certain foreign phrases – see Section 6.

1.3 POS changes

Sometimes, a word of one POS becomes a word of another POS. Say, in Czech, many adjectives became nouns as well; for example *nemocný* (both *sick* and *sick person*); similarly, in Russian, *moroženoje (frozen* and *ice cream), učenyj (learned* and *scholar).*

Annotating these words with the new POS should be used conservatively. Only clear and well established words should be annotated with the new POS. Otherwise it is better to analyze such words in their original POS. For example, in many languages, many strees have the form adjective + *Street* and it is common to use the adjective only when referring to the street. Thus, in Czech, *Dlouhá ulice* (lit: Long street) is usually called simply *Dlouhá* (lit: Long). This should be analyzed as an ellipsis and the word *Dlouhá* should be simply tagged as an adjective. If in doubt, tag with the original POS (and possibly add a comment).

2 Lemma and tag structure

2.1 Lemma structure

Lemma is 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 (this distinction is optional).

Some lemmas are accompanied by additional information, usually as e.g. a comment or explanation. Such comment is introduced by _^ and, strictly speaking, is not part of the lemma.

Czech Examples

Whole lemma	Lemma proper	Second part
beautiful	*beautiful*	
bank-1_^(institution)	*bank-1*	*_^(institution)*
vazba-2_^(river_bank)	*bank-2*	*_^(river_bank)*

2.2 Tag structure

We use so-called positional tags. Such a tag is a string of n characters, where n is dependent on a language. Every positions encodes one morphological category using one character (mostly upper case letters or numbers). This system is based on the Czech positional tagset.

For example, in Czech positional tags have 15 positions:

Position	Name	Description
1	POS	Part of speech
2	SubPOS	Detailed part of speech
3	Gender	Gender
4	Number	Number
5	Case	Case
6	PossGender	Possessor’s gender
7	PossNumber	Possessor’s number
8	Person	Person

9	Tense	Tense
10	Grade	Degree of comparison
11	Negation	Negation
12	Voice	Voice
13	Reserve1	Not used
14	Reserve2	Not used
15	Var	Variant, style

While the number and order of positions is dependent on a language, for any given language every tag has the same length and contains all positions in the same order.

All our tagsets, regardles of the language, share certain features:
- the first position is part-of-speech (POS)
- the second position detailed POS (SubPOS)
- SubPOS uniquely determines POS
- N/A values are encoded as '-' (tense or case for an interjection)
- 'X' means any value (e.g. used for gender of Russian adjectives in plural). We do not use any other aggregate values

We tried to design the individual tagsets as similar as possible to each other and to the original Czech tagset (for example, in all languages, 'N' is used for nouns, 'S' for singular , etc.).

Czech examples:

hraniční (border$_{adj}$): *AAIS4----1A----*
standard adjective, masc. inanimate, singular, accusative, positive (not negated)
potok (stream): *NNIS4-----A----*
noun, masc. inanimate, singular, accusative, positive
ODS (Civic Democratic Party): *NNFXX-----A---8*
noun, feminine, any number, any case, positive, abbreviation
podle (according to): *RR--2----------*
preposition (non vocalized), requiring genitive
see http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html

Russian examples:

Член партии по возможности старался не говорить ...
Член NNMS1-----A----
noun, masculine, singular, nominative, positive (not negated)
партии NNFS2-----A----
noun, feminine, singular, genitive, positive
по RR—6----------
preposition (not-vocalized) requiring locative case
возможности NNFS6-----A----
noun, feminine, singular, locative, positive
старался VpMS----R-AA---
past participle, masculine, singular, (past), positive, active voice
не TT----------

particle
говорить Vf--------A----
infinitive, positive
Note, however, that the Russian tagset is under development so some details might change.

2.2.1Informal abbreviations

Often we use the following tag abbreviations. Most of them are self-evident (dashes and rarely used fields are omitted), as you can see in the following list:

*NN*gnc – noun; *NNFS1 = NNFS1-----A----*
*Aa*gnc – adjective; *AAXXX = AAXXX----1A----*
Db – adverb; *Db = Db----------*
Dg – adverb; *Dg = Dg-------1A----*
*Dg*d – adverb; *Dg2 = Dg-------2A----*
*J^* – conjunction; *J^ = J^----------*
J, – conjunction; *J, = J,----------*
*Rc, RR*c – preposition, *RR7 = RR--7----------*
*RV*c – vocalized preposition, *RV7 = RV--7----------*
TT – particle; *TT = TT----------*

NNgXX-8 – noun abreviation; NFXX-8 = NNFXX-----A---8
AAXXX-8 – adjective abreviation; AAXXX-8 = AAXXX----1A---8
Db-8 – adverb abreviation; Db-8 = Db----------8
Rc-8, RRc-8 – preposition abreviation; RR7-8 = RR—7---------8

To keep the text simple, we use NNXXX and AAXXX for nouns and adjectives that do not change their form and if the language distinguishes grammatical gender, it is not specified for the noun.

3 Proper Names

Lemmas of proper names that are written with an initial capital in an usual text should be capitalized as well (lemma of *Berlin* is *Berlin*, not *berlin*). This applies even if the particular token is written with all lower case letter (e.g. for typographic reasons), thus both *Joseph* and *joseph* have the same lemma *Joseph*.

All personal names are annotated as nouns, even if they have a form of another POS.

Czech Examples:

Josef Nový – Nový + NNMS1-----A---- (*Nový* is a surname, *nový (new)* is an adjective)
Josef Sázel - Sázel + NNMS1-----A---- (*Sázel* is a surname, *sázel (planted*$_{masc.sg}$*)* is a verb)

3.1 von, van, etc.

Foreign prepositions like *van* or *von* in names (e.g. *Ludwig van Beethoven*) are annotated as regular prepositions (no case distinction). The noun following the preposition (e.g. *Beethoven*) is annotated as any other surname if it is used that way in current Czech. If it is still perceived and used as a place name, it should be annotated that way.

Czech Examples:

Ludwig van Beethoven – Ludwig+NNMS1 van+RR--X Beethoven+NNMS1
as a last name - you can say *Beethoven wrote*
Kryštof Harant z Polžic a Bezdružic – Kryštof +NNMS1 Harant+NNMS1 z-1+RR--2 Polžice NNFS2 a-1+J^ Bezdružice NNFS2
as regular place names - you cannot says *Polžic a Bezdružic wrote* or *Polžic wrote*

3.2 Adjectives and Nouns in Geographical names

Streets

We suppose that the word *street*/*ulice*/.., etc. is always present, even if elided on the surface. Therefore adjectival names are annotated as adjectives not as nouns.

Czech Examples

Dlouhá (lit: Long$_{fem.sg}$) – Dlouhý + AAFS1----1A----
Dlouhá ulice (lit: Long$_{fem.sg}$ street$_{fem.sg}$) – Dlouhý + AAFS1----1A---- ulice + NNFS1-----A----
Palackého (*Palacký*$_{masc.sg.gen}$ *(St)*) *– Palacký + NNMS2-----A----*

Towns

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

Czech Examples

Hluboká (lit: Deep$_{fem.sg}$) – Hluboká + NFS1
Dobrá Voda (lit: Deep$_{fem.sg}$ Water$_{fem.sg}$) – Dobrá + AFS1 Voda + NFS1
Ohrada u Hluboké (lit: Hedge$_{fem.sg}$ near Deep$_{fem.sg.gen}$) – Ohrada + NFS1 u + RR2 Hluboká + NFS2

3.3 Unusual names - horses, DJ's etc.

Names that Horses, some Djs have all kind of names (e.g. *He Shall Reign, La Paloma Monitor, Gold End, Green Peace, First, Bounty*). If it is possible to analyze them as individual words, do it. Otherwise, analyze them as a single unit (assign the tag of the whole unit to the first word of the name, and tag the rest as X0----------).

3.4 Other – TODO

Czech Examples

Porcela Plus: *Porcela + NNFS1 Plus + TT* (*Porcela* is the head, it declines)
Paris Indoor – Paris+AAXXX----1A---- Indoor+NNNXX-----A----

US Open – US+AAXXX----1A---8 Open+NNISX----1A----

4 Abbreviations and Symbols

By default, the lemma of an abbreviation is the abbreviation itself and the tag is a tag of non-declined noun, with 8 as variant (*NNXXX-----A---8*). If it is clear that the abbreviation stands for a different POS or has a particular gender or number use the appropriate tag (say *AAXXX----1A---8* if it is an adjective). Frequent abbreviations may have their own lemmas and tags offered by the analyzer. If they apply, use them, but do not enter such specific analyses by yourself.

Czech examples:
trojúhelník ABC (triangle ABC) – *ABC + NNXXX-----A---8*
přímka PQ (line PQ) – *PQ + NNXXX-----A---8*
samopal SA-58 (machine gun SA-58) – *SA + NNXXX-----A---8*
RM-systém – RM + NNXXX-----A---8
Dr Josef Wagner - Dr + NNXXX-----A---8 (note we make no gender distiction)
virus AH 3 B – AH + NNXXX-----A---8
FBI zjistilo, že ... (FBI found$_{sing.neut}$ that ...) – FBI + *NNNSX-----A---8*
(gender and number by aggreement w/ verb)
FBI zjistila, že ... (FBI found$_{sing.fem}$ that ...) – FBI + *NNFSX-----A---8* (ditto)
Detektivové FBI zjistili, že ... (Investigators from FBI found that ...*) – FBI + NNXXX-----A---8* (default tag)
gap: BERLIN (gap) Na základě posudku ... – gap + NNXXX-----A---8
this is an abbreviation identifying the author of the article
Jh8 (chess code) – *Jh8 + NNXXX-----A---8*

např. (e.g.) – *například + Db----------8* (a frequent abbr, with a dedicated lemma)

4.1 Isolated letters

Isolated letters whether abbreviations (*B.Clinton)* or symbols (*zápas skupiny B, odstavec b*) are handled as abbreviations. The letter itself is the lemma (possibly followed by a number). The only exception is when there is a special lemma for a frequent abbreviation offered by the analyzer (*rok for r.*). Do not insert dedicated lemmas.

Czech Examples

A: *A-mužstvo* (A-team) – *A + NNNXX-----A---8*
d: *odst. 1 písm. d)* (paragraph 1 d) – *d + NNNXX-----A---8*
A: *A konto – A + NNNXX-----A---8*
A: *16 A – Amper + NNIXX-----A---8*
a: *ABC a.s.* (ABC Inc) – *akciový + AAXXX----1A---8*
r: r. 1998 (year 1998) – *rok + NNIXX-----A---8*
s: *na s. 128* (on page 128) – *strana + NNFXX-----A---8*

TODO :Should labels labels of paragraphs and cases (*odst. 1 písm. d)* or *za a.*) be kept separate.?
TODO: How about *O* in *O'Brien*? Maybe all letters without 8? Or two variants, and in case of doubt use the non 8.

5 Colloquial and Archaic Language

If an official alternative to the colloquial form exist, then the the colloquial form has the same tag except a different variant (‘5’, '6' or '7'). Archaic forms have variant '2' or '3'. Consult the manual for a given tagset.

Czech Examples

které: stavení, které – P4NP4---------5
Novákovic: Novákovic pes – Novákův – AUXXXM--------6
takovejhlema: *takovýhle – AAFP7----1A---6*
hovadinama: *hovadina – NNFP7-----A---6*
naší: *pro naší atletiku* (officially short: *naši*) – *můj – PSFS4-P1------6*

6 Foreign words and phrases

In the following, we use the term *Czech* to mean the primary language of the annotated document and *foreign language* as the language of the foreign phrase.

TODO: Would be nice to mark all foreign phrases somehow (again what is foreign and what isn't this is fuzzy).

6.1 Two ways

There are two ways for annotating foreign words and phrases:

1. Longer phrases or citations that act as a single unit should analyzed as a single unit – the first word is tagged as such a word and the rest is given tag X0----------. Hence we will refer to these expression as *citation-use*.

   Czech examples:

   - *Dostihy vyhrála She will reign.* (*She will reign won the race*)
     Horse name: *She*+NNNXX *will*+X0 *and*+X0 *reign*+X0
   - *Nakonec zahráli I saw her standing there*. (*At the end, they played I saw her standing there.*)
     Song name tagged as: I+NNNXX; the rest gets X0.

2. Single words and phrases that are perceived as having internal structure (for example, the individual words inflect) and are incorporated into the sentence are tagged individually. This usually includes names of people, companies, cities
   etc. consisting of single nouns, article plus noun etc. Hence we will refer to this foreign expressions as *word-use*.

   Czech examples:
   - *kniha o Bill Clintonovi* (*book about B. C.*) – two nouns *Bill* [NNMXX] *Clinton* [NNMS3]
   - *Pozvali Musicu Bohemicu. (They invited Musica Bohemica)* – two nouns: *Musica*+NNFS2 *Bohemica*+NNFS2

The borderline is fuzzy, of course.

6.2 General rules

Foreign words and phrases are tagged according to their usage in Czech (by the particular speaker/writer), not according to their original native morphology.

Czech examples:
- *La Manche* – masculine noun because one says *bouřlivý*$_{masc}$ *La Manche* (stormy LM)
  In French, it is a feminine noun.
- *Udinese* – noun (name of a football club, NNNXX-----A----).
  In Italian, it is an adjective of *Udine* (town in NE Italy), the official name of the football club is *Udinese Calcio* (*calcio = football*).

Another general rule is – do not spend with it more time than necessary. As long as it is clear that a particular phrase is foreign, it is not that crucial whether it is annotated as a singular noun or a noun not distinguishing gender.

If a foreign phrase is introduced by a Czech noun phrase, annotate the foreign phrase as nondeclinable neuter noun (NNNXX-----A----).

Czech example:
- *Zazpívali písničku Uwanuma (They sang a song named Uwanuma) – Uwanuma* is annotated as NNNXX-----A----, because it is introduced by a Czech noun *písničku (song).*
- *Zazpívali písničku I love to hate you. (They sang a song named I love to hate you)* – the whole phrase *I love to hate you* is annotated as NNNXX-----A----, that means *I* is annotated as NNNXX-----A----; the rest gets X0, see §6.1.

6.3 Nondeclined and declined words

Some words or names from foreign languages may be used in both declined and nondeclined forms.

Czech Example:
- *kniha o Bill [NNXXX-] Clintonovi[NNMS3-] (a book about Bill Clinton)*
- *kniha o Billu [NNMS3-] Clintonovi [NNMS3-] (a book about Bill Clinton)*

Forms that are equal to the base form, and could thus be considered both declined and nondeclined, should be tagged as declined (unless the word is never declined).

Czech Example:

- *Bill [NNMS1] Clinton [NNMS1] přijel do Prahy. (Bill Clinton arrived to Prague)*

6.4 Czech Examples

It is hard to provide general language independent rules on how to annotate foreign expressions, because each language treats foreign expressions differently. Instead, we provide examples of various phenomena in Czech and how they should be annotated. This can then be used as a guideline for other languages.

English NP

Determiners and attributes (adjectives, modifying nouns, etc) are all annotated as non-declinable adjectives (AAXXX), because they are perceived that way in Czech. (Cases when the modifiers are complex NPs are probably best analyzed as single units).

Examples:

- *Náš cash flow se stabilizoval (Our cash flow stabilized) – cash+AAXXX, flow+NNISX- or NNIS1*
- *v cash flow statementu jsme uvedli (in the cash flow statement they wrote) cash+AAXXX, flow+AAXXX, statement+NNIS6*
- *Náš cash flow statement … (our cash flow statement) – as above or as a single unit NNIS1*
- *oba dva cash flow (oficiální i skutečný) (both cash flows (official and real)) … – flow + NNIXX* (gender based on agreement with *oba dva*)

However, modifying NPs in names of sport events should be annotated as NPs and not as AAXXX. The reason is that such construction started to be used even with non-foreign words where they are clearly perceived as a sequence of NPs, for example *Staropramen Extraliga* or *Český Telecom Cup.* Thus *Motorola* in *Motorola Cup* should be annotated as a feminine noun.

Other examples

V kostele XY zpívala Musica Bohemica. (In the Church XY, there was singing Musica Bohemica)

> *Musica Bohemica* are annotated as 2 nouns; in Latin it is a noun + adjective.
> **Reason:** When the phrase is declined, *Bohemica* is declined as a noun (*žena*):
> *pozvali Musicu Bohemicu, *pozvali Musicu Bohemicou*
> **Annotation:** *Musica_,t + NFS1A, Bohemica_,t NFS1A*

To je trochu ad hoc. (That's a little bit ad hoc)

> *hoc* is annotated as a noun; in Latin it is an adverb.
> **Annotation:** *ad_,t RRX, hoc_,t NXXXA*

Al-Kajdá zmizela z ... (Al-Qaeda disappeared from ...) – Al + AAXXX, Kajdá – NNFS1[1]
Přeplaval celý La Manche (He swam over English Chanel) – La+AAXXX, Manche+NNIS1
Přijeli jsme do La Defence. (We arrived to La Defance) – La+AAXXX, Defence+NNFXX- (or NFXXX)
Jak napsal Corriere della Serra, ... (As Corriera della Serra wrote ...) - single unit: NNIS1- based on the form of the verb
z Los Angeles (*from Los Angeles)* – AAXXX + NNNS2 (in Spanish plural)
Nakonec zaspívali Girl od Beatles. (At the end, they san Girl from Beatles) – Girl + NNFXX
Zahrajeme písničku Girls. (*We will play the song Girls) – Girls + NNFXX*
A zase jsme u toho To be or not to be. (And we are back to ..) – as a unit NNX

7 Errors

The text can contain errors. As stated in one of the basic rules, it is important to annotate the words as the writer used them and not as they should have used them

Only clear typos or errors resulting from text conversions, etc. should be corrected. Never, correct author's language and never correct colloquial language. If you are unsure whether to correct something or not, do not correct it.

The errors have to be just marked, do not edit the file. Insert lemma and tag as if the form were correct, but add a comment "*Error"* or "*Error* - <description>". This convention makes it easy to find the errors automatically.

7.1 Characters

Sometimes, foreign characters had been be messed up (e.g. *Jifi* instead of *Jiří) Fran?oise* instead of *Françoise* or *Ji i* or *Jiφi* instead of *Jiří*), and therefore the morphological analyzer did not assign them the correct analysis.

If it was analyzed as a single word, simply mark it as an error and insert the correct analysis. If they were not analyzed as a single word (Fran?oise would be analyzed as Fran + ? + oise). Mark all of them as errors and insert the correct analysis to the first one of them. However, only technical errors should be corrected like this, if it was the author who put the „improper" character there, it should be annotated that way and not corrected (for example, The Economist has a policy never to use diacritics exept for French or German). Thus *Jiri* or *Francoise* should be annotated as *Jiri* or *Francoise* not as an error.

7.2 Separators, etc.

Sometimes, the text contains *o* or *I* as bullets or separators. They should be marked for deletion – choose one of the the first lemma/tag pairs, edit them (press F4 or right-click on it an choose Edit from the menu) and enter X into the comment (if you want to add additional comment enter it as "*X* - <comment>".

[1] The current tokenizer treats it as 3 tokens, treating it as a single noun would be probably better.

8 Specific cases

TODO

- Transcription of pronunciation
- Isolated morphemes
- Crippled forms

9 Specific languages

9.1 Specific languages – Czech

- *sto a pětiset in "**sto-, pětiset**- a tisícikoruny" Lemma: as the cardinal numeral (sto, pětset), tag A2--------A----.*

9.1.1Hyphenated composites

If the hyphenated word ends with *-o,* and replacing that -o by an adjective ending yields an adjective (normal or possesive), the lemma for the word is that adjective (e.g. *česko-německý – česko → český, Karlo-Ferdinanova – Karlo → Karlův*). Some word cannot be viewed as derived from adjectives, but rather from nouns (e.g. *rap-jazzová – rap → rap* vs. *rapovo-jazzová – rapovo → rapový*).

Examples

srbsko-černohorská – srbský – A2--------A----
Univerzita Karlo-Ferdinandova – Karlův – A2--------A----
Univerzita Karel-Ferdinandova – Karel – A2--------A----
rap-jazzová: *rap – A2--------A----*
rapo-jazzová: *rap – A2--------A----* (could have a variant 1)
rapovo-jazzová: *rapový – A2--------A----*

9.2 Specific languages – Russian

9.3 Negation

The values A/N in the negation slot of a tag refer to the presence (N) or absence of a negative prefix не- (ne) for open class words:

Examples:

ненацеленной – нацеленный + AAFS7----1N----
нацеленной – нацеленный + AAFS7----1A----
невнятные – невнятные + AAXP1----1N----
незнание – знание + NNNS1-----N----

незнакомых – знакомый + NNXP2-----N----

Note that for pronouns the slot has always N/A value. Whether they have negative meaning or not is specified by their lemma.

Words that not negated *synchronically* (they may still have negative semantics, but the initial не is not a morphological prefix anymore.)

Examples:

ненависть – ненависть + NNFS4-----A---- (there is no *нависть*)
непереносимый – непереносимый + AAMS4----1A---- (*переносимый* has a very different meaning[2]).

9.3.1*Adjectives*

If you are unsure which SubPOS to use:
- if deciding between A (regular adj), M (long participle) or U (possessive adj) use A
- if deciding between C (short adj) and c (short participle) use C.

You may want to add a comment.

Indeclinable adjectives (e.g. *khaki*) are annotated as AAXXX----1A----. They usually follow the noun

9.3.2 *Numerals*

Nol' and numerals above 999 (*tysjača, milion* etc) are annotated as regular nouns.

dvojka, devjatka - nouns
pjatok, desjatok - nouns

dvuxletnij - adjective
pjatiletka - noun

9.3.3 *Prepositions*

Some prepositions consists of several words. In such a case, only the final word receives the regular RR--X or RV--X tag, the preceding parts of prepositions receive RF tag. The set of compound prepositions should be small and most of them are already offered as such by the analyzer. Be conservative in inserting new ones, if possible, annotated them as a regular combination of prepositions with other words. If in doubt, do not use the RF tag.

Examples:

в интересах – в+RR--2 *интерес*+NNMP6
несмотря на – несмотря+RF *на*+RR—4

[2] One may imagine that the lemma could be *непереносимый* with the tag being AAMS4----1N---- and there would be simply no form of the lemma with a tag AAMS4----1A----. However, at least for now, assume that such words are not negated.