A manual for resources creators

Important: Feel free to add your comments when you see a problem, don't understand something, some issue is not addressed, something can be explained better, etc.

But please: Make sure it is obvious it was you who added the text. THe best thing is to start your comment with your name and highlight the whole text with some color, for example:

Joe: I really don't understand this part.

Lemmas.

Roughly: a lemma (base form) is the form of a word that the word is listed in a dictionary under.

For example:

Lemma	Language	Lemma of	But not of	Unclear
house	eng	house, houses	household	house (verb)
this	eng	this, these	that, the	
a	eng	a, an	the, one	
good	eng	good, better, best	goodwill	well
slow	eng	slow, slower, slowest	slowness	slowly
blanco	spa	blanco, blanca, blancos, blancas		
go	eng	go, goes, going, went, gone		

Generally, forms of a single lemmas should have the same POS and even SubPOS. There

can be exceptions, but only if one knows what they are doing (e.g. one might argue that English nouns and verbs should have the same lemma, in our case, this is not really a problem, because the lemmas are the same.).

TODO: Should possessives have the personal pronoun as their lemma?
TODO: Decide on pronouns in general

It is better to lump together forms that maybe should be separate, than to split forms according to a wrong boundary. For example, it is much better to assign a single lemma to adjectives and the corresponding adverbs, than to assign one lemma to *slow* and *slowly* and another to *slower, slowest, slower*$_{adv}$, *slowest*$_{adv}$.

Personal Pronouns

TODO: finalize
For historical reasons, forms having the same person and number have the same lemma

For example: *I* and *me* have the same lemma (probably *I*), *she* and *her* too (*she*), but *she*, *he, you* have each a distinct lemma.

Tagset design

Introduction

Default shape of the tagset

Unless there is a reason to do it otherwise use the following tagset slots, values, etc.

TODO: Mark all those that are unlikely to change.

Default Positions

p	POS	Part of Speech
s	SubPOS	Detailed Part of Speech
g	Gender	
y	Animacy	
n	Number	
c	Case	
f	PossGender	Possessor’s Gender
m	PossNumber	Possessor’s Number
e	Person	
t	Tense	
b	Aspect	
d	Grade	Degree of comparison

a	Negation	
v	Voice	
y?	Definiteness	TODO: decide which position
i	Var	Variant, Style, Register, Special Usage

Notes

1. **All tagsets have**
    - POS - first position
    - SubPOS - second position
    - Var - the last slot, use '8' value for abbreviations and '-' if variant distinction is not needed
2. **aa**
3. **Abbreviations**
    - lowercase a-z letters
    - **unique within a particular tagset**
    - avoid clash with the above abbreviation
    - if possible avoid clash with abbreviations used in other tagsets
4. a

POS Values

A	Adjective
C	Numeral
D	Adverb
I	Interjection
J	Conjunction
N	Noun
P	Pronoun
V	Verb
R	Preposition
T	Particle
X	Unknown, not determined, unclassifiable; foreign.
Z	Punctuation

Do not change this set of values unless really needed.
TODO: Participles are under A or V or both? Should this be language dependent?
TODO: Determiner - special POS?

Other values

For other slots, take RUssian as the basis.

TODO: Reflexivity - subpos vs special slot

Wildcards

- Use sparingly

- Use for words that do not distinguish certain category even though other words of the same SubPOS do. For example, in a language with noun cases some borrowed nouns might not distinguish case, use 'X' instead of considering the word ambiguous.
- There are no partial wildcards (e.g. masculine and feminine, but not neuter gender). In such case consider the word ambiguous.
- In general, the SubPOS distinguishes which positions must be specified and which are N/A (have '-' value). However, for closed-class words, a N/A value instead of 'X' or introducing a new SubPOS. This is done to increase readability. For example, in Czech, personal pronouns distinguish gender only in 3rd person. Instead of either using different SubPOS for non-3rd persons and 3rd person, or using 'X' for gender for non-3rd person, N/A value is used.
- It is not always clear when to consider a word ambiguous, when to use the wildcard and when to use N/A value.

Adding new positions

When a language distinguishes some category not captured by the default tagset, you may need to add it. But not automatically. Sometimes using more complex value set for another position (usually SubPos) may be a better solution. TODO talk about compromise between tagset length and straightforwardness.

No new positions for closed classed words

- No slots are introduced for distinctions made only by closed classed words (they are already captured by their lemmas).

  For example, the proximity distinction of *this* and *that* are captured by their lemma so there is no reason to encode it by the tagset as well (this would be true even if the same distinction was be made by other pronouns/adverbs/determiners, e.g. *there*, *then*, etc.

  However, when a particular distinction is made by both open and closed class words, specify the position for CC words, too. For example, since person and number is introduced for other classes, we use it for pronouns too. Thus, *I* would be sg and *we* plural, even though their lemmas capture such a distinction as well.

  Reasonable number of subPos values can be introduced for classes of CC words - see below.

(Sub)POS

- The set of basic POS values should be followed unless there is a strong reason to do otherwise. For example:
  - participles should not be a main POS, even when a particular grammar book classifies them as such, instead they should be classified as verbs, adjectives, and/or adverbs.
  - articles/determiners - TODO decide what to do. Maybe a different approach in lgs with articles and without? (In Romance, determiners have POS "D", articles are DA, demonstrative determiners DD, demonstrative pronouns PD; in Czech/Russian determiners are withing adjectives and pronouns (eg. possessive pronouns

    - numerals - numerals are considered a separate category, even though one can argue it should not be. Thus *one* is a cardinal numeral, not a noun, and *first* is an ordinal numeral not an adjective. But see a note on Numeral x Noun below.
- Reasonable number of subPos values can be introduced for classes of CC words, especially for traditional categories and categories distinguished by other tagsets (e.g. personal/possessive/interrogative/... pronouns)

Genitive vs. possesive

Some languages lost most of their case systems (English, French). However, for historical reasons, possessive forms are sometimes still referred to as genitive. If such genitive forms have adjective-like properties (agree in gender/number) and treating them as forms with different SubPos (e.g. possessive pronouns/adjectives) instead of forms with a different case simplifies the situation do it.

Possessor's gender/number.

For some languages it makes sense to distinguish two gender slots (similarly numbers):

1. Agreement gender (adjectives, verbs, possessive prons, ..) and lexical gender of nouns
2. Possessor's gender (possessive pronouns, possessive adjectives)

Note: It might be argued that the lexical gender should share the slot with the possessor and not with the agreement gender. However, for various reasons (including historical) it is done this way.

TODO: what if the language has both (all three) type of genders but no word distinguishes both of them at the same time. Would we want to save slots by using the same slot. For example if Russian/Czech did not have possessive adjectives.

Example - Russian:

- agreement/lex gender: slot 3, values: M/F/N/X,
  distinguished for nouns, adjectives, pronouns, numerals, verbs
- possessor gender: slot 7, value: M/F/N/X
  distinguished for possesive adjectives, possesive pronouns

ženščina 'woman' fem - NNM**F**S1-----A----
mužnin 'man's' masc & fem -AU**M**XS2**F**----A----
    the possessed thing is feminine, the possessors gender is not masculine
ona 'she' fem PP**F**-S1**-**--3I------
moja 'my' fem PS**F**XS1-S1I------
    the possessed thing is feminine, the possessors gender is not captured ('-')
ejo 'her' fem PS**X**XXX**M**S3I------
    the possessed thing is of any gender, the possessor is feminine
    one might argue the agreement gender should be N/A instead of X

Definiteness

Introduced only when encoded morphologically by open-class words (not just determiners), i.e. it is marked by affixes (or clitics spelled together with another word - but see section on clitics). Tagsets of languages that encode it by orthographically separate words do not capture it (see also section *No new positions for closed classed words*).

For example, both English and Bulgarian distinguish definiteness, but an English tagset would not capture it because articles are separate words, while a Bulgarian tagset probably would, because definite articles are written together with their host (not necessarily the definite noun).

Clitics

- spelled together with their hosts - capture them by a dedicated slot.

    I would not put that info into the subpos position. Enclitics are attached to the word, so they are somewhat peripheral. In many languages they are even spelled as separate words (although usually still pronounced together with their host). The SubPos position should capture a finer distinction of the POS and both of these should classify the main word, not the clitic. Linguistically, the cleanest solution would probably be to use two tags - one for the host and one for the enclitic. However, that would be rather unusual from the NLP perspective and could cause problems for some tools. Since the inventory of clitics is rather limited, I think that putting it into a separate position is reasonable. If there was a large set of clitics with their own cases, numbers, etc. we would have to come up with something different (e.g. separate clitics from their host, e.g. you're -> you + 're)

Slot: Var
Var - the last slot, use '8' value for abbreviations and '-' if variant distinction is not needed

Other

Numerals

A Cardinal numerals x Nouns

In many languages, numerals above 99 (100, 1000, 10^6, ...) behave (sometimes optionally) like nouns (they decline like other nouns, have a lexical gender, etc.). Usually, it is not easy to decide whether they are pure nouns or something in between a numeral and noun, unless one knows the language well.

On the other hand, this particular distinction is not that terribly important. If they have some properties of nouns (e.g. they have a gender, while other *cardinal* numeral do not), classify them as nouns. In most cases, 10^6 and above are nouns.

For example, in Czech, *million*, *billion*, etc are nouns: they have gender and they decline the same way as other nouns. The numerals 100 and 1000 are somewhere between.

One can either decline 1000 as a noun or it can be indeclinable (as many other numerals). Therefore, 1000 is annotated as a noun in some contexts and as a numeral in others.