Morphologically annotated corpus by Filosoft

- description of the data format

The character encoding is UTF8.

The corpus is tagged for sentences, clauses, and morphology. Punctuation marks are separate tokens. The morphological readings have been automatically disambiguated, but several tokens (most notably participles) still have more than one analysis.

Every token (and its analysis) is on a separate line; for every input word form, the structure of the word (e.g. stem, derivational suffix, inflectional affix), the word class and inflectional categories (e.g. number and case) are given.

The beginning of a sentence is signalled by <s> on a separate line; the ending by </s> on a separate line. Before every sentence, there is a description of its source. The beginning of this description is signalled by <ignoreeri> on a separate line; the ending by </ignoreeri> on a separate line.

Clause boundaries are marked by html entities which are glued to the end of analysis of tokens (punctuation marks or conjunctions). &kiilualgus; signals the beginning of an inserted clause, and &kiilulopp; signals its end. &kindel_piir; signals a breaking point between clauses.

Example:
laserikiirte laseri_kiir+te //_S_ pl g, //

The input token is laserikiirte (an inflectional form of “laser beams”). The analysis follows, separated by 4 spaces and shows that the lemma “lasreikiir” is a compound word (the border is marked with underscore), its inflectional ending is –te, it is a noun (_S_), plural (pl) genitive case (g)

If the word is a derived one or a compound, then:
1. The stem is separated from the previous component by “_”.
2. The inflectional affix is separated from the previous component by “+”.
3. The derivational suffix is separated from the previous component by “=”.

Only the rightmost component is lemmatised. There may be up to 5 stems in a compound word. The "+" and "=" are not used in a very principled way, so for practical purposes they could be deleted.

In foreign proper names consisting of more than one word, like New York, only the last word inflects, e.g. singular inessiv New Yorgis ‘in New York’. Such names are treated in an ad hoc manner as compounds; the blank is retained as a separator:

New Yorgis New York+s //_H_ sg in, //

Output may consist of more than one analysis, because a word form may be ambiguous. Analyses are separated by 4 spaces.

Example for purustatud (destroyed):
purustatud purusta+tud //_V_ tud, // purusta=tu+d //_S_ pl n, // purusta=tud+0 //_A_ // purusta=tud+0 //_A_ sg n, // purusta=tud+d //_A_ pl n, //

A zero-ending is marked as "0". If a word cannot have an ending (e.g. a conjunction), its ending is still depicted as "0", but the grammatical categories are left empty. The postfix ("gi" or "ki") is appended to the inflectional ending.

The base form stem is the same as the dictionary headword, except for verbs. If a word is a verb, then its lemma (the dictionary headword) is created by adding an ending "ma" to the stem, with the exception of the following 6 verb stems: "ei", "ära", "är", "kuulukse", "tunnukse", "näikse". (The task of generating the verb lemmas is put on the shoulders of the user, because in Estonian lexicographical practice these 6 exceptional verb stems are often also treated in some exceptional and un-documented way.)

Grammatical categories

(about Estonian morphology, see also http://www.cl.ut.ee/korpused/morfliides/seletus.php?lang=en, http://www.filosoft.ee/index_en.html, http://www.eki.ee/teemad/morfoloogia/)

In Estonian, a clitic "gi" or "ki" (meaning "even, too") may be attached to almost every word as a post-fix after an inflectional ending. It does not change the grammatical behaviour of the word, thus it is not reflected in the inflectional categories.

Part of speech abbreviations:

A = Adjective (positive)
C = Adjective (comparative)
D = Adverb
G = Genitive attribute, i.e. indeclinable adjective
H = Proper noun
I = Interjection
J = Conjunction
K = Adposition (pre- or postposition)
N = Numeral (cardinal)
O = Numeral (ordinal)
P = Pronoun
S = Common noun
U = Adjective (superlative)
V = Verb
X = Verb particle
Y = Abbreviation or acronym
Z = Punctuation

Inflections

Declinable words (adjectives, numerals, pronouns, nouns):

? = case undefined (e.g. a shortened word form)
sg n = singular nominative
sg g = singular genitive
sg p = singular partitive
sg ill = singular illative
sg in = singular inessive
sg el = singular elative

sg all = singular allative
sg ad = singular adessive
sg abl = singular ablative
sg tr = singular translative
sg ter = singular terminative
sg es = singular essive
sg ab = singular abessive
sg kom = singular komitative
adt = additive (grammatically, it is the same as singular illative)
pl n = plural nominative
pl g = plural genitive
pl p = plural partitive
pl ill = plural illative
pl in = plural inessive
pl el = plural elative
pl all = plural allative
pl ad = plural adessive
pl abl = plural ablative
pl tr = plural translative
pl ter = plural terminative
pl es = plural essive
pl ab = plural abessive
pl kom = plural komitative

Verbs:

The verb inflection of Estonian is really a very complicated question; all the grammar books and dictionaries that describe the verb morphology of Estonian, give slightly different descriptions.

When one thinks about the word grammar as a system of categories, one has categories like tense, mood, number, person, time etc. However, when one looks at the word grammar from how you can cut the wordforms into formatives (roots, inflectional endings), one notices that the same formative can represent different sets of grammar categories (e.g. verb ending 'sid' always indicates either "second person, singular" or "third person, plural") or that different formatives may represent the same set of grammar categories (e.g. verb endings 'ks', 'ksid').

In case of Estonian verb inflection, the differences between the system of grammatical categories and the system of inflectional formatives are so big that instead of giving the set of grammatical categories for every word form, we follow the tradition that gives (for a wordform) the formative as a shorthand tag for a set of grammatical categories. This way we hide some of the regular ambiguity of the formatives. If necessary, one may expand the portmanteau tags to explicitly list all the possible combinations of grammatical categories. It is just a question of representation.

b = indic present third singular active afirmative
d = indic present second singular active afirmative
da = infinit
des = gerund
ge = imper present second plural active afirmative
ge = imper present second plural active negative
gem = imper present first plural active afirmative
gem = imper present first plural active negative
gu = imper present third plural active afirmative
gu = imper present third plural active negative
gu = imper present third singular active afirmative
gu = imper present third singular active negative

ks = condit present active negative
ks = condit present first plural active afirmative
ks = condit present first singular active afirmative
ks = condit present second plural active afirmative
ks = condit present second singular active afirmative
ks = condit present third plural active afirmative
ks = condit present third singular active afirmative
ksid = condit present second singular active afirmative
ksid = condit present third plural active afirmative
ksime = condit present first plural active afirmative
ksin = condit present first singular active afirmative
ksite = condit present second plural active afirmative
ma = supine active illative
maks = supine active translative
mas = supine active inessive
mast = supine active elative
mata = supine active abessive
me = indic present first plural active afirmative
n = indic present first singular active afirmative
neg ge = imper present second plural active negative
neg gem = imper present first plural active negative
neg gu = imper present passive negative
neg gu = imper present third plural active negative
neg gu = imper present third singular active negative
neg ks = condit present active negative
neg nud = indic imperfect active negative
neg nuks = condit past active negative
neg o = imper present second singular active negative
neg o = indic present active negative
neg vat = quotat present active negative
neg = negative
nud = indic imperfect active negative
nud = partic past active
nuks = condit past active negative
nuks = condit past first plural active afirmative
nuks = condit past first singular active afirmative
nuks = condit past second plural active afirmative
nuks = condit past second singular active afirmative
nuks = condit past third plural active afirmative
nuks = condit past third singular active afirmative
nuksid = condit past second singular active afirmative
nuksid = condit past third plural active afirmative
nuksime = condit past first plural active afirmative
nuksin = condit past first singular active afirmative
nuksite = condit past second plural active afirmative
nuvat = quotat past active afirmative
nuvat = quotat past active negative
o = imper present second singular active afirmative
o = imper present second singular active negative
o = indic present active negative
s = indic imperfect third singular active afirmative
sid = indic imperfect second singular active afirmative
sid = indic imperfect third plural active afirmative
sime = indic imperfect first plural active afirmative
sin = indic imperfect first singular active afirmative
site = indic imperfect second plural active afirmative
ta = indic present passive negative
tagu = imper present passive afirmative
tagu = imper present passive negative
taks = condit present passive afirmative
taks = condit present passive negative

takse = indic present passive afirmative
tama = supine passive
tav = partic present passive
tavat = quotat present passive afirmative
tavat = quotat present passive negative
te = indic present second plural active afirmative
ti = indic imperfect passive afirmative
tud = indic imperfect passive negative
tud = partic past passive
tuks = condit past passive afirmative
tuks = condit past passive negative
tuvat = quotat past passive afirmative
tuvat = quotat past passive negative
v = partic present active
vad = indic present third plural active afirmative
vat = quotat present active afirmative
vat = quotat present active negative