

General System for Normal and Phonetic Inflection

*Ștefan Diaconescu** sdiaconescu@softwin.ro

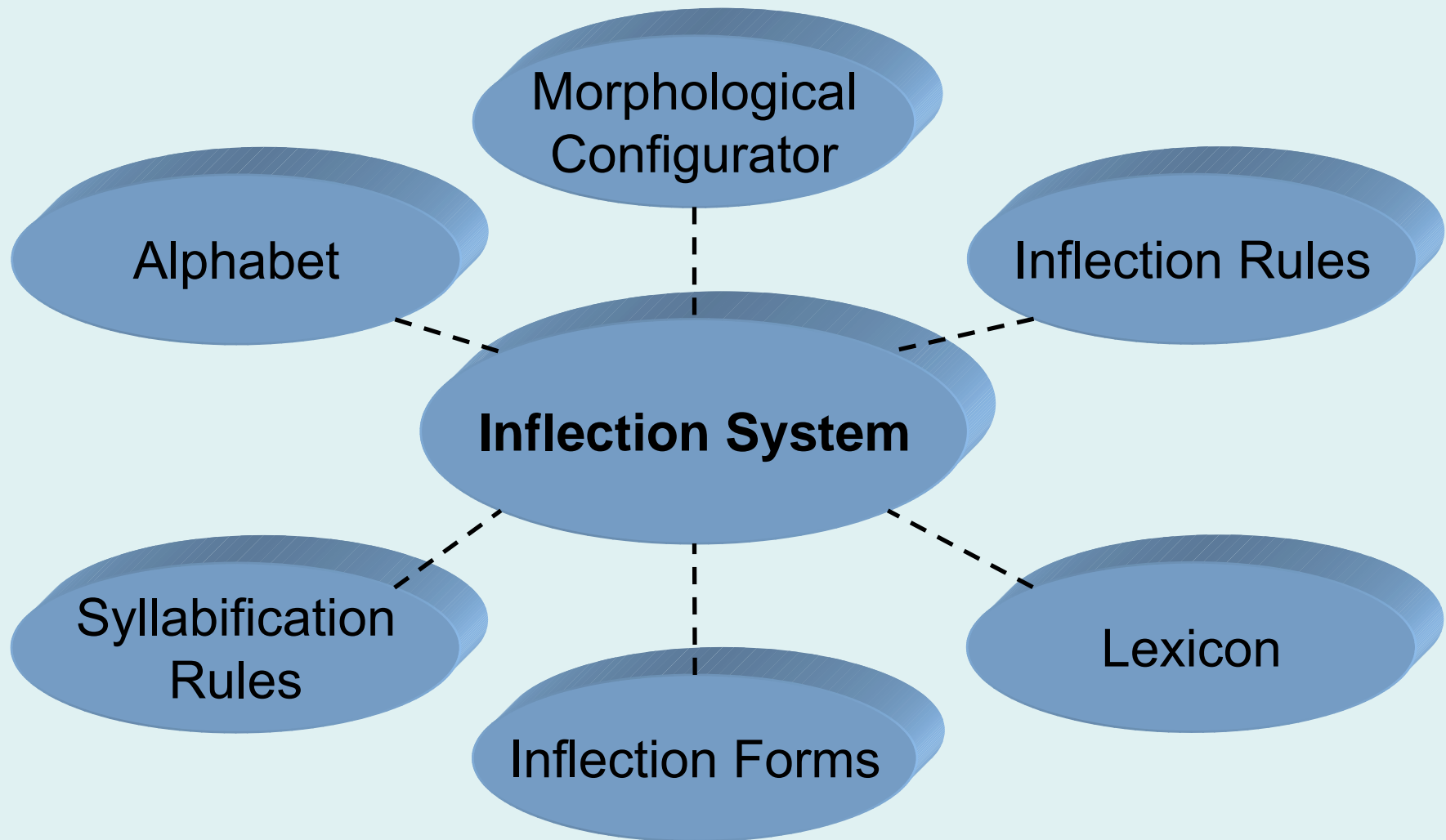
*Cristi Ingineru** ingineru@softwin.ro

*Felicia Codirlaşu** fcodirlasu@softwin.ro

*Monica Rizea** mrizea@softwin.ro

*Oana Bulibașa** obulibasa@softwin.ro

*SOFTWIN

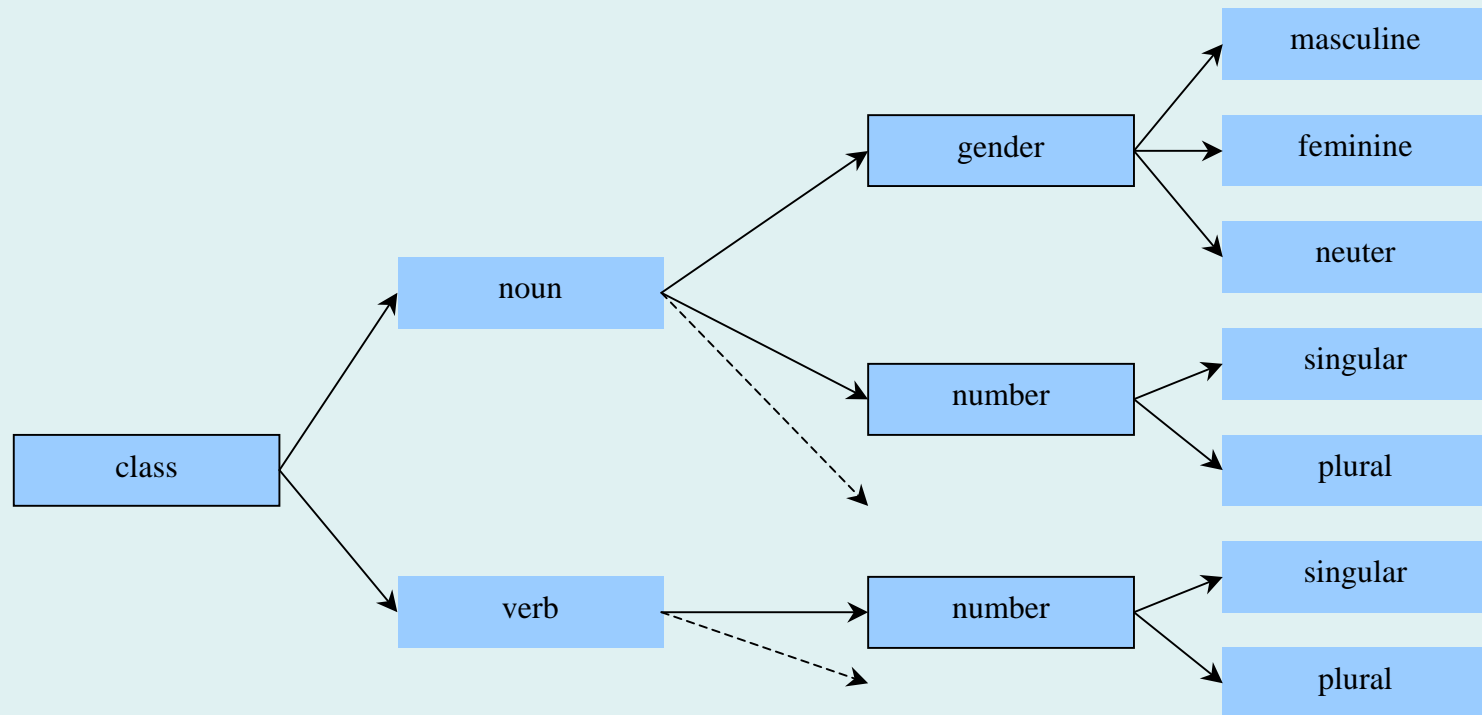


Grammar Abstract Language – GRAALAN

- Meta language that allows the description of a serial of linguistic aspects regarding one natural language or the correspondences between two natural languages
- It can describe any natural, insulated or non insulated inflected (synthetic or analytic) language
- Flexible
- Very easy to use by the linguists

The Morphological Configurator

- Describes the morphologic structure of a language
- Organized as an Attribute Value Tree (AVT)
 - attribute nodes: morphological categories
 - value nodes: morphological categories values
- Other types of information attached to each node:
 - the abbreviation
 - the category is inflected or not
 - belongs to a lemma or not
 - belongs to a supplement or not



Tree

```

[clasa                / name = Clasa, abbreviation = Cls, inflection = no /
  = substantiv       / name = Substantiv, abbreviation = Subst, lemma = yes,
                    / lexicon = input /
  [tip substantiv    / name = TipSubstantiv, abbreviation = TipSubst, inflection = no /
    = comun          / name = Comun, abbreviation = Com, lemma = yes,
                    / lexicon = input /
    , propriu        / name = Propriu, abbreviation = Pr, lemma = yes,
                    / lexicon = input /
  ]
[animatie            / name = Animatie, abbreviation = Animat, inflection = no /
  = animat           / name = Animat, abbreviation = Anim, lemma = yes,
                    / lexicon = input /
  , inanimat         / name = Inanimat, abbreviation = Inanim, lemma = yes,
                    / lexicon = input /
]

```

.....

The Alphabet

- Defines all symbols used in a language
 - phonetic alphabet
 - normal alphabet
 - special symbols
 - stress markers

- Defines the structures of a language's alphabet and the relations between those types of symbols
 - groups
 - classes

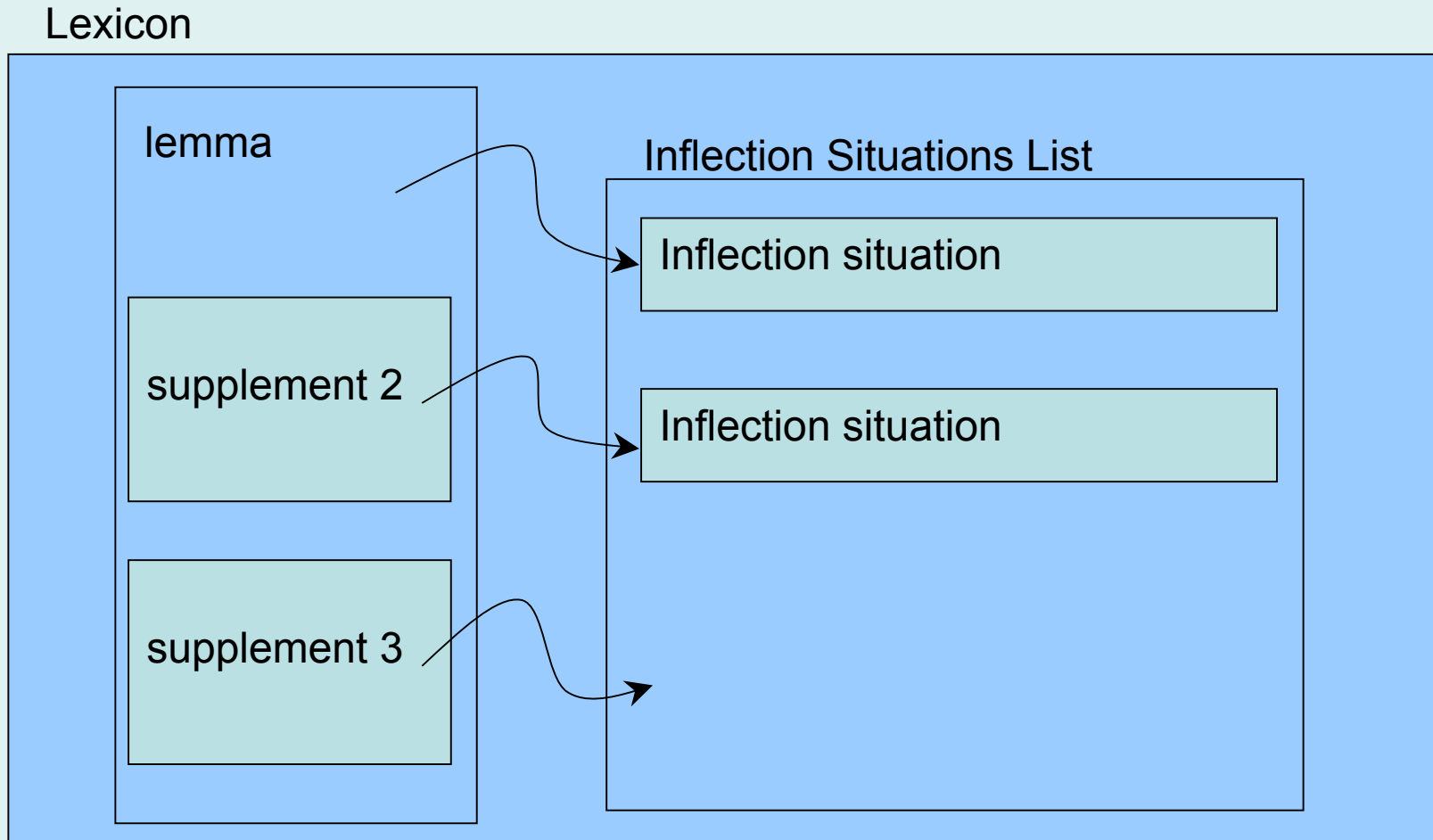
Examples for each Romanian alphabet entity type:

ə	character code = "ə" type = internal label = mid_central_unrounded stressed = yes order key = 1.1
A	character code = "A" type = internal label = A order key = 2.1
-	character code = "‐" type = <i>internal</i> label = hyphen special function = <i>connector</i> order key = 3.39
ı	stress code = "ˈ" type = primary label = primary_stress order key = 4.0
iou	group code = (("il"/"oO"/"uU") [("&semivowel_i;&mid_back_rounded;&semivowel_u;"])] label = triphthong_iou
A, Ă, Â, B, C, D, E...	class label = capital_letter elements = ("A", "Ă", "Â", "B", "C", "D", "E", "F", "G", "H", "I", "Î", "J", "K", "L", "M", "N", "O", "P", "Q", "R", "S", "Ş", "T", "Ţ", "U", "V", "W", "X", "Z")

The Lexicon

- Defines words, expressions and lexical / morphological / syntactic structures
- Types of entries:
 - words (lemma, supplement, wordform)
 - morphemes (roots, prefixes, suffixes, etc.)
 - inflection situation
 - multi word expression
 - morphological analytical structures
 - syntactic structures

Lemma is the only type of entry from lexicon involved in the inflection process, with the following associated information.



Entry00017711: Entry word lemma

Text "cântec"

Phonetic "k'ɨntek"

Syllabification

Euphonic "cân/tec"

Phonetic "k'ɨ/ntek"

Gloss "Şir armonios de sunete emise cu vocea sau cu un instrument"

Morphology

Inflection situation SubstTipComunInaniNeutrNomSg

Inflection rule Flex_SubstNeutru

Supplement

Text "cântece"

Phonetic "k'ɨntetʃe"

Number 2

Syllabification

Euphonic "cân/te/ce"

Phonetic "k'ɨ/nte/tʃe"

Morphology

Inflection situation SubstTipComunInaniNeutrNomPl

Markers x

end of entry

The Syllabification Rules

- Set of rules for the separation of a word into syllables, whether spoken or written.
- Syllabification types:
 - euphonic syllabification
 - phonetic syllabification
 - morphologic syllabification

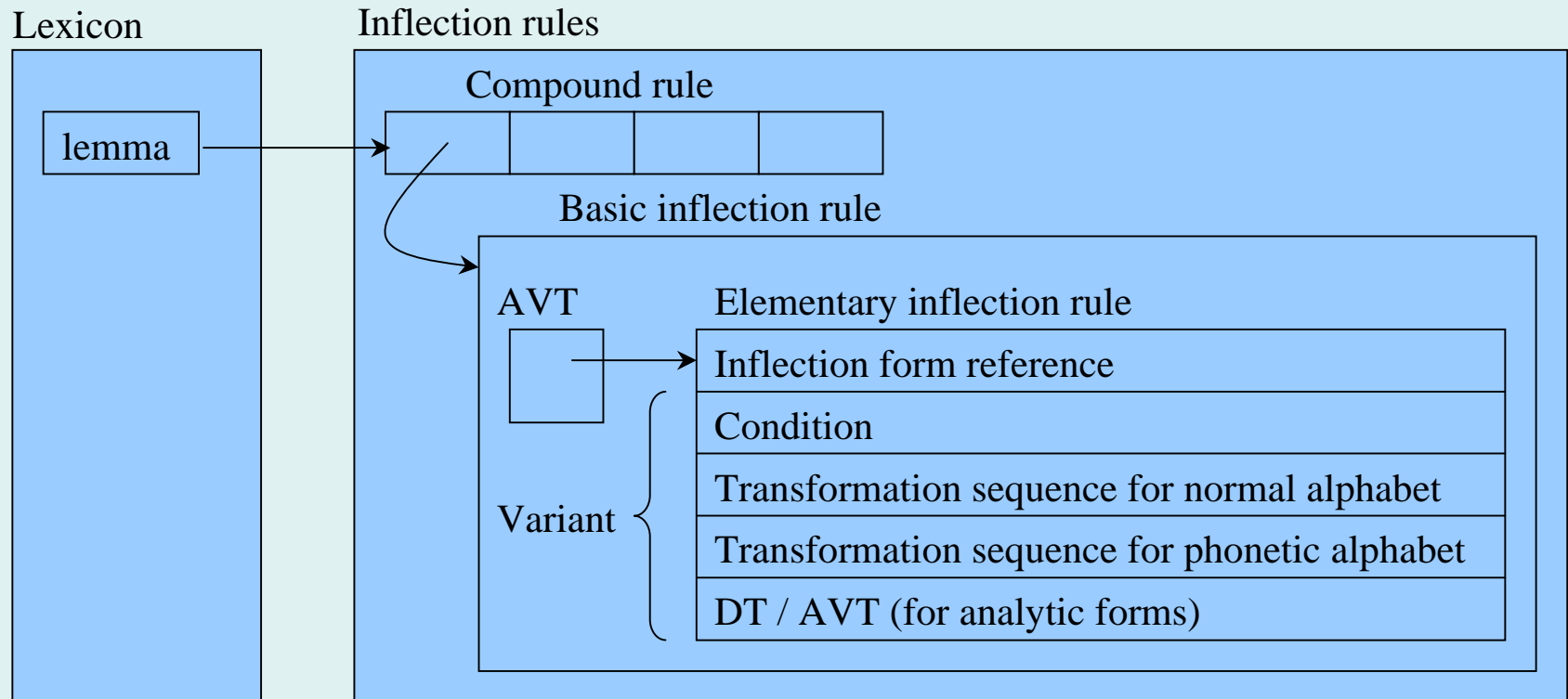
Euphonic	Rule "&vowel;" - "&semivowel;" + "&vowel;"
Phonetic	Rule "&phon_vowel; " + "&phon_semivowel;" - "&phon_semivowel;" + "&phon_semivowel;" + "&phon_vowel;"

Note: &vowel; or &semivowel; are labels referencing entities from the alphabet section, like alphabetic and phonetic characters or classes.

The Inflection Rules

- Define the actions that can be used to generate the inflection forms

- Types of inflection rules:
 - basic rules formed by an AVT and elementary transformation rules attached to the leafs of this AVT
 - compound rules: a list of basic rules



Basic Rule Subst_masc1:

[clasa = substantiv]

[tip substantiv = comun]

[animatie = animat, inanimat]

[gen = masculin]

[numar = singular

[caz = nominativ [articulare = nearticulat (EtL1: **alphabetic - phonetic -**)

, hotarat (EtS11:

/* last letter is a consonant - băiat, elev */

if(&consonant;) alphabetic insert "ul"**phonetic insert**

"&close_back_rounded;&alveolar_lateral_approximant;"

/* last letter is "i" - ochi */

if("i") alphabetic insert "ul"**phonetic insert**

"&close_back_rounded;&alveolar_lateral_approximant;"

...]

, genitiv ... , dativ ... , acuzativ ...]

, plural ...

]]

The Inflection Forms

- The result of the inflecting process are the inflection forms
- Complex structure consisted in:
 - text (alphabetic and phonetic)
 - syllabification (euphonic, phonetic and morphologic)
 - structure (tri, central word, auxiliary words)
- Each auxiliary word has a complete description:
 - text (alphabetic and phonetic)
 - lemma label
 - inflection situation
 - relation name
 - belongs or not (has a syntactic value or not)

ETF_Entry00018335_1:

Entry Text "un cent"

Phonetic "'un tʃ'ent"

Reference Entry00018335

[clasa = substantiv]

[tip substantiv = comun]

[animatie = inanimat]

[gen = masculin]

[numar = singular]

[caz = nominativ]

[articulare = nehotarat]

Syllabification

Euphonic "un cent"

Phonetic "'un tʃ'ent"

Tri 1 left

Central word

Text "cent"

Phonetic "tʃ'ent"

...

...

[clasa = substantiv]

[tip substantiv = comun]

[animatie = inanimat]

[gen = masculin]

[caz = nominativ]

[numar = singular]

[articulare = nearticulat]

Auxiliary words

Text "un"

Phonetic "'un"

Reference Art01

[clasa = articol]

[tip articol = nehotarat]

[caz = nominativ]

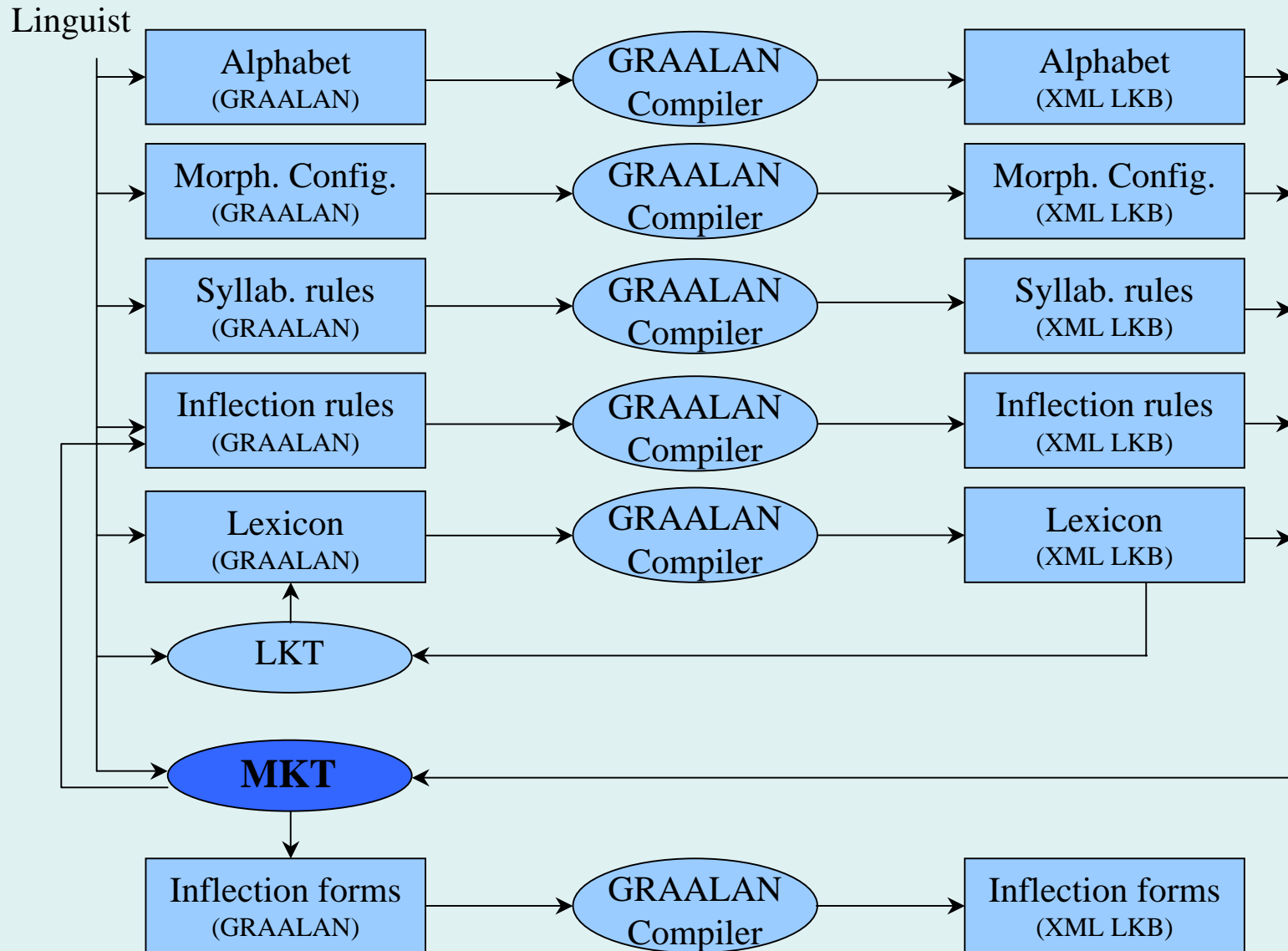
[gen = masculin]

[numar = singular]

Belongs = yes

@acord-art@

end of entry



Alphabet XML LKB	←	Alphabet DTD
Syllabification XML LKB	←	Syllabification DTD
Morphology XML LKB	←	Morphology DTD
Inflection rules XML LKB	←	Inflection rules DTD
Inflection forms XML LKB	←	Inflection forms DTD
Lexicon XML LKB	←	Lexicon DTD

The Morphological Knowledge Tool

- The primary scope of MKT: creating the inflection forms
- Friendly interface in order to help the linguists to verify and to correct the inflection forms
- All the corrections made by the linguist (exceptions) are saved in a controlled manner

I. The word tree preparation

- a. Copying Morphological Configurator
- b. Loading lemma's inflection situation
- c. Intersection
- d. Expanding

II. The Inflecting

- a. Loading inflection rule
- b. Generating inflection form nodes
- c. Generating inflection forms
- d. Loading inflection forms
- e. Syllabification

Actions:

- a. Replace
- b. Delete
- c. Insert

Type:

- a. char
- b. word

Where:

- a. Alphabetic / Phonetic text
- b. Syllabification

- Morphological analyzer
- Grammar checker
- Inflection
- Indexing/Searching
- Lemmatizer
- Speller
- Hyphenating
- Lexical dictionaries
- Human assisted machine translation
- Computer assisted machine translation
- Automatic machine translation
- etc.

The Morphological Configurator:

Name	Total attributes	Inflected attributes	Total values
Class	819	111	2,122
Class = Noun	9	3	27
Class = Article	13	2	38
Class = Adjective	12	0	30
Class = Pronoun	112	12	324
Class = Numeral	154	43	447
Class = Verb	478	35	1,142
Class = Adverb	33	8	89
Class = Preposition	1	1	4
Class = Conjunction	3	3	10
Class = Interjection	2	2	5

The Morphological Configurator:

Name	EC Number	Maximum EC length	Attributes names	Values names
Class	19,462	36	67	211
Class = Noun	312	12	6	18
Class = Article	82	8	4	15
Class = Adjective	420	14	9	21
Class = Pronoun	1,118	16	13	45
Class = Numeral	1,124	16	20	63
Class = Verb	16,320	34	24	73
Class = Adverb	68	14	12	32
Class = Preposition	3	2	1	4
Class = Conjunction	7	4	3	10
Class = Interjection	3	4	2	5

The Alphabet:

Sign types	Signs number	Examples
normal alphabet	66	A, a, B, b, ...
phonetic alphabet	36	, , k, ...
special characters	64	~, ?, -, ...
stress markers	2	primary_stress, secondary_stress, ...
groups	360	a_group, tch_group, sh_group, ...
classes	17	capital_letter, small_letter, vowel, semivowel, consonant, diphthong, ...

The Lexicon:

- 76,000 lemmas
 - 66,000 mono-word lemma
 - 10,000 multi-word lemmas
- 115,000 senses
- 103,000 supplements
- 12,700 multiword expressions.

The Syllabification Rules:

- 723 normal syllabification rules
- 723 phonetic syllabification rules

The Inflection Rules:

- Inflection situations (EC) that have inflection rules: 19,202
- Inflection situations that do not have inflected rules: 260
- Variants: 28,317
- Multiword variants: 19,935
- "Defective" situations: 2,936
- Mono-word variants: 8,382
- Multiword variants with 2 words: 7,785
- Multiword variants with 3 words: 6,554
- Multiword variants with 4 words: 3,196
- Multiword variants with 5 words: 1,908
- Multiword variants with 6 words: 492.

The Inflection Forms:

- 17,928,056 inflection situations
 - 2,075,978 mono-words
 - 15,852,078 multi-words
- 9,946,686 inflection forms
 - 1,019,783 mono-words
 - 8,926,903 multi-words

- The system has a great generalization degree
- Inflection forms are obtained relatively fast and complete
- It is very easy to correct the exceptions
- Generated inflected forms have a complex structure
- The system has the possibility to generate reports and various statistics

- Q&A