

TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT
EESTI KEELE OSAKOND

Ekstraktmorfoloogia meetodiga tuletatud keeletehnoloogia vadja noomeni vormisõnastiku näitel

Magistritöö kaitsmine
2019.06.18

Kristian Kankainen

Juhendajad dotsent Heinike Heinsoo ja PhD Külli Prillop

Magistritöö eesmärk

Luua noomenitega vormisõnastik, mis on:

- keeleõppes kasutatav
- automaatselt teisendatav keeletehnoloogiaks
- täiendatav ilma programmeerimisoskusteta

Magistritöö eesmärk

Luuu noomenitega vormisõnastik, mis on:

- ✓ keeleõppes kasutatav
- automaatselt teisendatav keeletehnoloogiaks
- täiendatav ilma programmeerimisoskusteta

Magistritöö eesmärk

Luuu noomenitega vormisõnastik, mis on:

- ✓ keeleõppes kasutatav
- ✓ automaatselt teisendatav keeletehnoloogiaks
- täiendatav ilma programmeerimisoskusteta

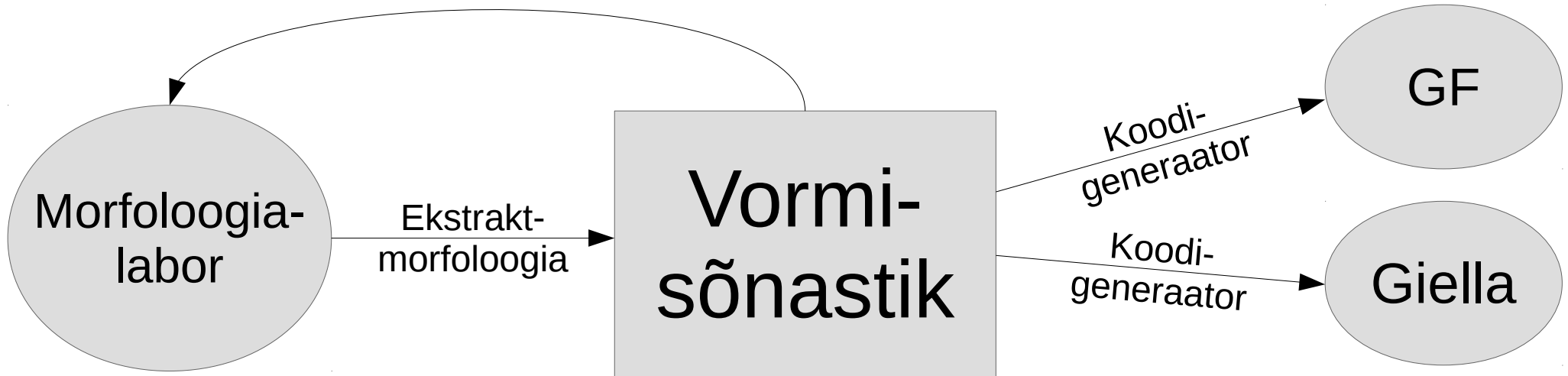
Magistritöö eesmärk

Luuu noomenitega vormisõnastik, mis on:

- ✓ keeleõppes kasutatav
- ✓ automaatselt teisendatav keeletehnoloogiaks
- ✓ täiendatav ilma programmeerimisoskusteta

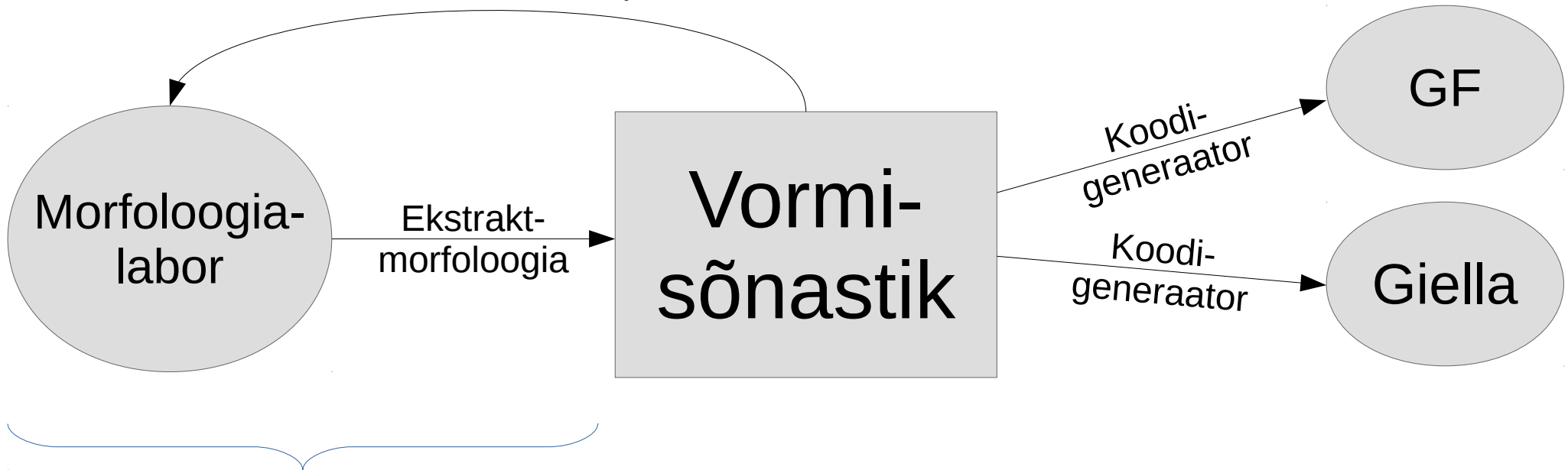
Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa



Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa

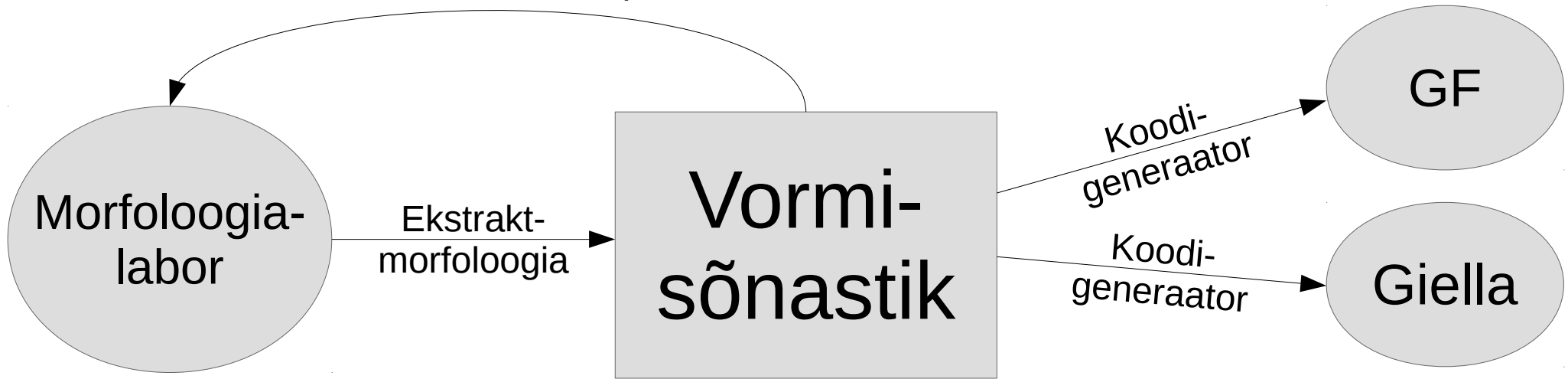


Morfoloogialabor on kasutajaliides

- Språkbanken'i arendatud rakendus (Karp)
- rakendab ekstraktmorfoloogiat
- sõna lisamine vormisõnastikku:
 1. sisesta sõnavorm
 2. vali õige tüüpsõna (sünteesib muutvormid)
 3. tüüpsõna puudumisel sisesta kõik muutvormid

Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa



Morfoloogialabor on kasutajaliides

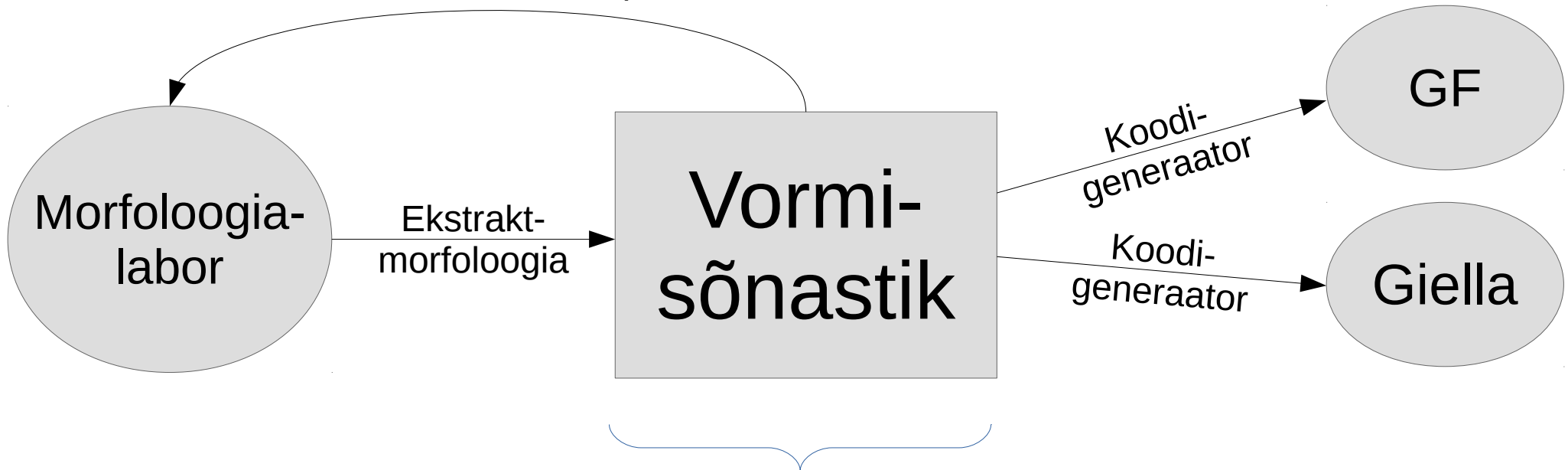
- Språkbanken'i arendatud rakendus (Karp)
- rakendab ekstraktmorfoloogiat
- sõna lisamine vormisõnastikku:
 1. sisesta sõnavorm
 2. vali õige tüüpsõna (sünteesib muutvormid)
 3. tüüpsõna puudumisel sisesta kõik muutvormid

Ekstraktmorfoloogia meetod

- eraldab muutvormide tabelist tüüpsõnamalli

Magistritöö ülevaade

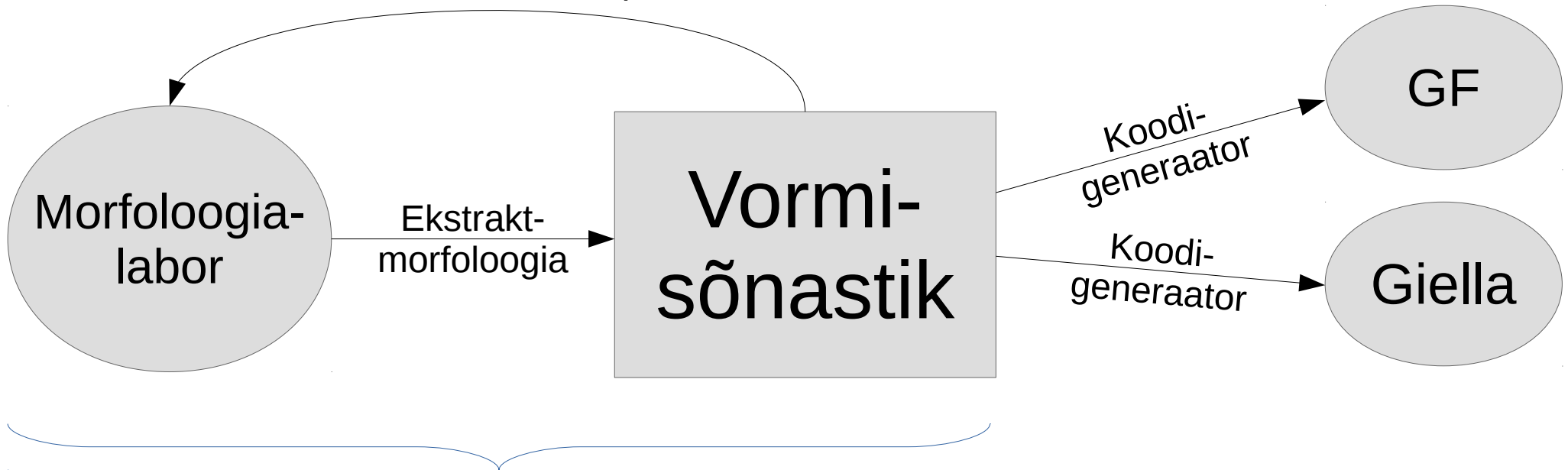
Tüüpsõnade ühtlustamine
käändkondade kaupa



- kokku 881 noomenit
 - ½ Heinsoo Sõnakopittõja sõnastik, ½ Tsvetkov
 - 24 käänet, levinud õppematerjalide järgi
 - 6 põhikäänet Tsvetkovi sõnaraamatust
 - Tsvetkovi Jõgõperä murre on Vaipoolistatud
- kokku 231 tüüpsõna(malli)
- kodeerib kõik informatsiooni rahvusvahelisesse standardi Lexical Markup Framework (ISO 24613:2007)

Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa



Ühtlustamise protsess:

- jagada tüüpsõnad Ariste käändkondade järgi
- sarnaste tüüpsõnade seast valida üks
- muuta teised lekseemid valitud tüüpsõna järgi
- korrata

Alguses oli üle 500 tüüpsõna, lõpuks 231.

Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa



Muuda sõnavormi – muudad programmikoodi

Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa

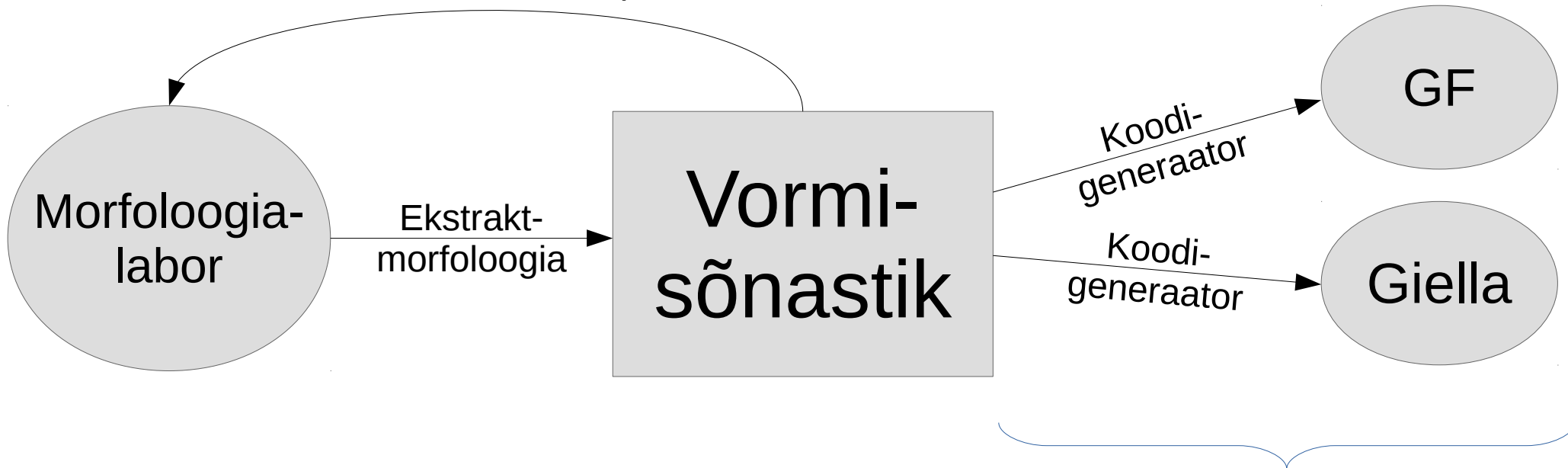


Muuda sõnavormi – muudad programmikoodi

Arvutimorfoloogia ei jää toppama arvutilingvisti taha

Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa



- generaator teisendab leksikoni ja tüüpsõnamallid programmikoodi
- vormisõnastik defineerib *mis*
- programmikood defineerib *kuidas*
- morfoloogia integreeritud:
 - prog. keelde Grammatical Framework
 - Giella keeletehnoloogia taristu
- generaatoreid võib lisada

Magistritöö ülevaade

vadja-vene
vestmik

Tüüpsõnade ühtlustamine
käändkondade kaupa

Morfoloogia-
labor

Ekstrakt-
morfoloogia

Vormi-
sõnastik

Koodi-
generaator

GF

Koodi-
generaator

Giella

- generaator teisendab leksikoni ja tüüpsõnamallid programmkoodi
- vormisõnastik defineerib *mis*
- programmkood defineerib *kuidas*
- morfoloogia integreeritud:
 - prog. keelde Grammatical Framework
 - Giella keeletehnoloogia taristu
- generaatoreid võib lisada

Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa



- generaator teisendab leksikoni ja tüüpsõnamallid programmikoodi
- vormisõnastik defineerib *mis*
- programmikood defineerib *kuidas*
- morfoloogia integreeritud:
 - prog. keelde Grammatical Framework
 - Giella keeletehnoloogia taristu
- generaatoreid võib lisada

Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa

Morfoloogia-
labor

Ekstrakt-
morfoloogia

Vormi-
sõnastik

Koodi-
generaator

GF

analüsaator ja
süntesaator

Giella

õigekirja-
kontrollija

vadja-vene
vestmik

- generaator teisendab leksikoni ja tüüpsõnamallid programmkoodi
- vormisõnastik defineerib *mis*
- programmkood defineerib *kuidas*
- morfoloogia integreeritud:
 - prog. keelde Grammatical Framework
 - Giella keeletehnoloogia taristu
- generaatoreid võib lisada

Magistritöö ülevaade

Tüüpsõnade ühtlustamine
käändkondade kaupa

Morfoloogia-
labor

Ekstrakt-
morfoloogia

Vormi-
sõnastik

Koodi-
generaator

GF

analüsaator ja
süntesaator

Giella

õigekirja-
kontrollija

generaator
Koodi-

magistritöö
tabelid

- generaator teisendab leksikoni ja tüüpsõnamallid programmkoodi
- vormisõnastik defineerib *mis*
- programmkood defineerib *kuidas*
- morfoloogia integreeritud:
 - prog. keelde Grammatical Framework
 - Giella keeletehnoloogia taristu
- generaatoreid võib lisada

vadja-vene
vestmik

Magistritöö eesmärk

Luuu noomenitega vormisõnastik, mis on:

- ✓ keeleõppes kasutatav
- ✓ automaatselt teisendatav keeletehnoloogiaks
- ✓ täiendatav ilma programmeerimisoskusteta

Magistritöö eesmärk

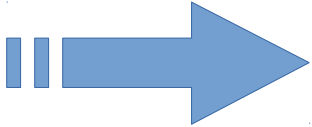
Luuu noomenitega vormisõnastik, mis on:

- ✓ keeleõppes kasutatav
- ✓ automaatselt teisendatav keeletehnoloogiaks
- ✓ täiendatav ilma programmeerimisoskusteta

Suurõd passibõd!

Ekstraktmorfoloogia meetod

- Ahlberg, Malin, Markus Forsberg ja Mans Hulden (2014)
- Eraldab (ekstraheerib) käändetabelist tüüpsõnamalli
- Tabeli korduvad osad: muutujad e tehniline tüvi
- Tüüpsõnamallid on inimloetavad

Muutvorm	Morfoloogia		Tehn. tüvi	Muutv. mall	Morfoloogia
<i>katto</i>	SG NOM		<u>kat</u> t <u>o</u>	x_1 t x_2	SG NOM
<i>katod</i>	SG GEN		<u>kat</u> <u>o</u> d	x_1 x_2 d	SG GEN

$\text{katto}(x_1, x_2) \longrightarrow \text{katto}(\text{čiut}, o)$

Muutv. mall	Tehn. tüvi	Morfoloogia
x_1 t x_2	<u>čiut</u> t <u>o</u>	SG NOM
x_1 x_2 d	<u>čiut</u> <u>o</u> d	SG GEN

kattoa
VS
tüttöä