Creating Linguistic Resources with the Grammatical Framework

Aarne Ranta

LREC Tutorial, Malta, 17 May 2010
1 The GF System and Simple Multilingual Grammars 4
  1.1 GF = Grammatical Framework 4
  1.2 Installing and using the GF system 6
  1.3 Working with context-free grammars in GF 7
  1.4 Abstract and concrete syntax 9
  1.5 Making a grammar multilingual 11
  1.6 Visualizing trees and word alignment 16
  1.7 From abstract trees to parse trees 16
  1.8 Scaling up the grammar 18
  1.9 Exercises 19

2 Morphological Paradigms and Lexicon Building 19
  2.1 Morphology 19
  2.2 Lexical categories 20
  2.3 Example: resource module for English verb inflection 21
  2.4 Smart paradigms 24
  2.5 Phases of morphology implementation 27
  2.6 Building a lexicon 28
  2.7 Nonconcatenative morphology: Arabic 29
  2.8 Exercises 34

3 Basics of a Linguistic Syntax Implementation 34
  3.1 Syntax in the resource grammar 35
  3.2 The key categories and functions 35
  3.3 Predication: building clauses 36
  3.4 Complementation: building verb phrases 38
  3.5 Determination: building noun phrases 39
  3.6 Modification: adding adjectives to nouns 40
  3.7 Lexical insertion 41
  3.8 A miniature resource grammar for Italian 42
3.9 Italian verb phrases ........................................ 45
3.10 Italian noun phrases ........................................ 46
3.11 Predication, at last .......................................... 48
3.12 To do ......................................................... 49
3.13 Ergativity in Hindi/Urdu ................................. 49
3.14 Exercises .................................................. 51

4 Using the Resource Grammar Library in Applications 51
4.1 Software libraries .......................................... 51
4.2 Using the library: natural language output .............. 52
4.3 Two kinds of grammarians ............................... 54
4.4 Meaning-preserving translation ........................... 54
4.5 Example: abstract syntax for a ”Face” community ... 56
4.6 Functors and interfaces .................................... 58
4.7 Modules of a domain grammar: ”Face” community ... 60
4.8 Free variation ............................................... 61
4.9 Overview of the resource grammar API ............... 62
4.10 Exercises .................................................. 66

5 Developing a GF Resource Grammar 66
5.1 The principal module structure ......................... 67
5.2 Effort statistics, completed languages .................. 72
5.3 How to start building a language, e.g. Marathi ......... 72
5.4 Assignment: a good start ................................ 75
Notice. This tutorial is an updated version of the one used at the GF Summer School 2009 (grammaticalframework.org/summerschool.html). It was first presented on an on-line course in April 2009. The summer school in August 2009 had 30 participants from 20 countries. 15 new languages were started. Since the summer school, the library has grown from 12 to 16 languages.

The goal of this tutorial is to introduce a fast way to resource grammar writing, by explaining the practical use of GF and the linguistic concepts in the resource grammar library.

For more details, we recommend

- the tutorial on the GF homepage grammaticalframework.org
- GF Book by A. Ranta, forthcoming at CSLI Publications

We cannot stress enough the importance of your own work on the code examples and exercises using the GF system!

1 The GF System and Simple Multilingual Grammars

Contents

What GF is
Installing the GF system
A grammar for John loves Mary in English, French, Latin, Dutch, Hebrew
Testing grammars and building applications
The scope of the Resource Grammar Library
Exercises

1.1 GF = Grammatical Framework

GF is a grammar formalism: a notation for writing grammars
GF is a functional programming language with types and modules
GF programs are called grammars
A grammar is a declarative program that defines

- parsing
- generation
- translation

**Multilingual grammars**

Many languages related by a common abstract syntax

\[
\begin{align*}
\text{Mary loves John} & \quad \text{Maria Ioannem amat} \\
\text{Fred Mary (Compl Love John)} & \\
\text{Marie aime Jean} & \quad \text{موس آھبت اتژ جان}
\end{align*}
\]

**The GF program**

**Interpreter** for testing grammars (the GF shell)

**Compiler** for converting grammars to useful formats

- PGF, Portable Grammar Format
- speech recognition grammars (Nuance, HTK, ...)
- JavaScript

**The GF Resource Grammar Library**

Morphology and basic syntax

Common API for different languages

Currently (April 2010) 17 languages: Bulgarian, Catalan, Danish, Dutch, English, Finnish, French, German, Interlingua, Italian, Norwegian, Polish, Romanian, Russian, Spanish, Swedish, Urdu.

Under construction for 16 languages: Arabic, Esperanto, Farsi, Greek (Ancient), Hebrew, Icelandic, Japanese, Latin, Latvian, Maltese, Mongol, Portuguese, Swahili, Thai, Tswana, Turkish.
GF run-time system

PGF grammars can be embedded in Haskell, Java, and Prolog programs. They can be used in web servers.

- fridge magnet demo: tournesol.cs.chalmers.se:41296/fridge
- translator demo: tournesol.cs.chalmers.se:41296/translate

1.2 Installing and using the GF system

Go to the GF home page, and follow shortcuts to either:

- Download: download and install binaries
- Developers: download sources, compile, and install

The Developers method is recommended for resource grammar developers:

- latest updates and bug fixes
- version control system

Starting the GF shell

The command $gf starts the GF shell:

$ gf

This is GF version 3.1.6.
License: see help -license.
Using the GF shell: help

Command h = help

> help

gives a list of commands with short descriptions.

> help parse

gives detailed help on the command parse.

Commands have both short (1 or 2 letters) and long names.

1.3 Working with context-free grammars in GF

These are the simplest grammars usable in GF. Example:

    Pred.  S ::= NP VP ;
    Compl. VP ::= V2 NP ;
    John.  NP ::= "John" ;
    Mary.  NP ::= "Mary" ;
    Love.  V2 ::= "loves" ;

The first item in each rule is a **syntactic function**, used for building trees: Pred = predication, Compl = complementation.

The second item is a **category**: S = Sentence, NP = Noun Phrase, VP = Verb Phrase, V2 = 2-place Verb.

Importing and parsing

Copy or write the above grammar in file zero.cf.

To use a grammar in GF: import = i

    > i zero.cf
To **parse** a string to a tree: `parse = p`

```plaintext
  > p "John loves Mary"
  Pred John (Compl Love Mary)
```

Parsing is, by default, in category $S$. This can be overridden.

**Random generation, linearization, and pipes**

Generate a random tree: `generate_random = gr`

```plaintext
  > gr
  Pred Mary (Compl Love Mary)
```

To **linearize** a tree to a string: `linearize = l`

```plaintext
  > l Pred Mary (Compl Love Mary)
  Mary loves Mary
```

To **pipe** a command to another one: `|`

```plaintext
  > gr | l
  Mary loves Mary
```

**Graphical view of abstract trees**

![Graphical Tree Diagram](image)

In Mac:

```plaintext
  > p "John loves Mary" | visualize_tree -view=open
```
In Ubuntu Linux:

```bash
> p "John loves Mary" | visualize_tree -view=oeg
```

You need the Graphviz program to see the view.

**Graphical view of parse trees**

![Parse Tree Diagram]

```bash
> p "John loves Mary" | visualize_parse -view=open
```

### 1.4 Abstract and concrete syntax

A context-free rule

```
Pred. S ::= NP VP
```

defines two things:

- **abstract syntax**: build a tree of form `Pred np vp`
- **concrete syntax**: this tree linearizes to a string of form `np vp`

The main idea of GF: separate these two things.
Separating abstract and concrete syntax

A context-free rule is converted to two judgements in GF:

- `fun`, declaring a syntactic function
- `lin`, giving its linearization rule

\[
\text{Pred. } S ::= \text{NP VP} \implies \text{fun Pred : NP -> VP -> S}
\]
\[
\text{lin Pred np vp = np ++ vp}
\]

Functions and concatenation

**Function type**: \(A \rightarrow B \rightarrow C\), read ”function from \(A\) and \(B\) to \(C\)”

**Function application**: \(f \ a\ b\), read ”\(f\) applied to arguments \(a\) and \(b\)”

**Concatenation**: \(x \ ++\ y\), read ”\(x\) followed by \(y\)”

Cf. functional programming in Haskell.

Notice: in GF, ++ is between token lists and therefore ”creates a space”.

From context-free to GF grammars

The grammar is divided to two modules

- an abstract module, judgement forms `cat` and `fun`
- a concrete module, judgement forms `lincat` and `lin`

<table>
<thead>
<tr>
<th>Judgement</th>
<th>reading</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>catC</code></td>
<td>(C) is a category</td>
</tr>
<tr>
<td><code>fun f : T</code></td>
<td>(f) is a function of type (T)</td>
</tr>
<tr>
<td><code>lincat C = L</code></td>
<td>(C) has linearization type (L)</td>
</tr>
<tr>
<td><code>lin f xs = t</code></td>
<td>(f) (xs) has linearization (t)</td>
</tr>
</tbody>
</table>

Abstract syntax, example

abstract Zero = {
    cat
    S ; NP ; VP ; V2 ;
    fun
Pred : NP -> VP -> S ;
Compl : V2 -> NP -> VP ;
John, Mary : NP ;
Love : V2 ;
} } 

Concrete syntax, English

concrete ZeroEng of Zero = {
lincat
  S, NP, VP, V2 = Str ;
lin
  Pred np vp = np ++ vp ;
  Compl v2 np = v2 ++ np ;
  John = "John" ;
  Mary = "Mary" ;
  Love = "loves" ;
} } 

Notice: Str (token list, "string") as the only linearization type.

1.5 Making a grammar multilingual

One abstract + many concretes

The same system of trees can be given

- different words
- different word orders
- different linearization types

Concrete syntax, French

concrete ZeroFre of Zero = {
lincat
  S, NP, VP, V2 = Str ;
lin
  Pred np vp = np ++ vp ;
  Compl v2 np = v2 ++ np ;
  John = "Jean" ;
Just use different words

Translation and multilingual generation

Import many grammars with the same abstract syntax

> i ZeroEng.gf ZeroFre.gf
Languages: ZeroEng ZeroFre

Translation: pipe linearization to parsing

> p -lang=ZeroEng "John loves Mary" | l -lang=ZeroFre
Jean aime Marie

Multilingual generation: linearize into all languages

> gr | l
Pred Mary (Compl Love Mary)
Mary loves Mary
Marie aime Marie

Multilingual treebanks

Treebank: show both trees and their linearizations

> gr | l -treebank
Zero: Pred Mary (Compl Love Mary)
ZeroEng: Mary loves Mary
ZeroFre: Marie aime Marie

Concrete syntax, Latin

concrete ZeroLat of Zero = {
lincat
  S, VP, V2 = Str ;
Different word order (SOV), different linearization type, parameters.

**Parameters in linearization**

Latin has *cases*: nominative for subject, accusative for object.

- *Ioannes Mariam amat* "John-Nom loves Mary-Acc"
- *Maria Ioannem amat* "Mary-Nom loves John-Acc"

**Parameter type** for case (just 2 of Latin’s 6 cases):

```
param Case = Nom | Acc
```

**Table types and tables**

The linearization type of NP is a **table type**: from Case to Str,

```
lincat NP = Case => Str
```

The linearization of John is an **inflection table**,

```
lin John = table {Nom => "Ioannes" ; Acc => "Ioannem"}
```

When using an NP, **select** (!) the appropriate case from the table,

```
Pred  np vp = np ! Nom ++ vp
Compl v2 np = np ! Acc ++ v2
```
Concrete syntax, Dutch

concrete ZeroDut of Zero = {
  lincat
    S, NP, VP = Str ;
    V2 = {v : Str ; p : Str} ;
  lin
    Pred np vp = np ++ vp ;
    Compl v2 np = v2.v ++ np ++ v2.p ;
    John = "Jan" ;
    Mary = "Marie" ;
    Love = {v = "heeft" ; p = "lief"} ;
}

The verb heeft lief is a discontinuous constituent.

Record types and records

The linearization type of $V2$ is a record type with two fields

    lincat V2 = {v : Str ; p : Str} 

The linearization of Love is a record

    lin Love = {v = "hat" ; p = "lieb"} 

The values of fields are picked by projection (.)

    lin Compl v2 np = v2.v ++ np ++ v2.p
Concrete syntax, Hebrew

```plaintext
concrete ZeroHeb of Zero = {
  flags coding=utf8;
  lincat
  S = Str;
  NP = {s : Str; g : Gender};
  VP, V2 = Gender => Str;
  lin
  Pred np vp = np.s ++ vp ! np.g;
  Compl v2 np = table {g => v2 ! g ++ "וְהָא" ++ np.s};
  John = {s = "יְהוָה"; g = Masc};
  Mary = {s = "רְאוֹב"; g = Fem};
  Love = table {Masc => "רְשָׁם"; Fem => "רְאוֹבָו"};
  param
  Gender = Masc | Fem;
}
```

The verb agrees to the gender of the subject.

**Variable and inherent features, agreement**

NP has gender as its **inherent feature** - a field in the record

```plaintext
  lincat NP = {s : Str; g : Gender}
  lin Mary = {s = "mry"; g = Fem}
```

VP has gender as its **variable feature** - an argument of a table

```plaintext
  lincat VP = Gender => Str
```

In predication, the VP receives the gender of the NP

```plaintext
  lin Pred np vp = np.s ++ vp ! np.g
```

**Feature design**

Deciding on variable and inherent features is central in GF programming.
Good hint: dictionaries give forms of variable features and values of inherent ones.
Example: French nouns

- *cheval* pl. *chevaux* masc. noun

From this we infer that French nouns have variable number and inherent gender

\[
\text{lincat } N = \{ s : \text{Number} \to \text{Str} ; g : \text{Gender} \}\]

### 1.6 Visualizing trees and word alignment

![Diagram of trees and word alignment]

### 1.7 From abstract trees to parse trees

Link every word with its **smallest spanning subtree**
Replace every **constructor function** with its **value category**

**Generating word alignment**

In L1 and L2: link every word with its smallest spanning subtree
Delete the intervening tree, combining links directly from L1 to L2
Notice: in general, this gives **phrase alignment**

*Notice:* links can be crossing, phrases can be discontinuous

**Word alignment via trees**

```
> parse "John loves Mary" | aw -view=open
```

**A more involved word alignment**
Building applications

Compile the grammar to PGF:

```bash
$ gf -make ZeroEng.gf ZeroFre.gf ZeroLat.gf ZeroGer.gf ZeroHeb.gf
```

The resulting file `Zero.pgf` can be e.g. included in fridge magnets:

![Example fridge magnet image]

1.8 Scaling up the grammar

Zero.gf is a tiny fragment of the Resource Grammar

The current Resource Grammar has 80 categories, 200 syntactic functions, and a minimal lexicon of 500 words.

Even S, NP, VP, V2 will need richer linearization types.

More to do on sentences

The category S has to take care of

- tenses: *John has loved Mary*
- negation: *John doesn’t love Mary*
- word order (German): *wenn Johann Maria lieb hat, hat Maria Johann lieb*

Moreover: questions, imperatives, relative clauses

More to do on noun phrases

NP also involves

- pronouns: *I, you, she, we*
- determiners: *the man, every place*

Moreover: common nouns, adjectives
1.9 Exercises

1. Install gf on your computer.
2. Learn and try out the commands align_words, empty, generate_random, generate_trees, help, import, linearize, parse, put_string, quit, read_file, translation_quiz, unicode_table, visualize_parse, visualize_tree, write_file.
3. Write a concrete syntax of Zero for yet another language (e.g. your summer school project language).
4. Extend the Zero grammar with ten new noun phrases and verbs.
5. Add to the Zero grammar a category A of adjectives and a function ComplA : A \rightarrow VP, which forms verb phrases like is old.

2 Morphological Paradigms and Lexicon Building

Contents

Morphology, inflection, paradigm - example: English verbs
Regular patterns and smart paradigms
Overloaded operations
Inherent features in the lexicon
Building and bootstrapping a lexicon
Nonconcatenative morphology: Arabic

2.1 Morphology

Inflectional morphology: define the different forms of words

- English verb sing has the forms sing, sings, sang, sung, singing

Derivational morphology: tell how new words are formed from old words

- English verb sing produces the noun singer

We could do both in GF, but concentrate now on inflectional morphology.
Good start for a resource grammar

Complete inflection system: 1-6 weeks
Comprehensive lexicon: days or weeks
Morphological analysis: up to 200,000 words per second
Export to SQL, XFST, ...

What is a word?

In abstract syntax: an object of a basic type, such as Love : V2

In concrete syntax,

- primarily: an inflection table, the collection of all forms
- secundarily: a string, i.e. a single form

Thus love, loves, loved are

- distinct words as strings
- forms of the same word as an inflection table or an abstract syntax object

2.2 Lexical categories

Part of speech = word class = lexical category

In GF, a part of speech is defined as a cat and its associated lincat.
In GF, there is no formal difference between lexical and other cats.
But in the resource grammar, we maintain a discipline of separate lexical categories.

The main lexical categories in the resource grammar

<table>
<thead>
<tr>
<th>cat</th>
<th>name</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>noun</td>
<td>house</td>
</tr>
<tr>
<td>A</td>
<td>adjective</td>
<td>small</td>
</tr>
<tr>
<td>V</td>
<td>verb</td>
<td>sleep</td>
</tr>
<tr>
<td>V2</td>
<td>two-place verb</td>
<td>love</td>
</tr>
<tr>
<td>Adv</td>
<td>adverb</td>
<td>today</td>
</tr>
</tbody>
</table>
Typical feature design

<table>
<thead>
<tr>
<th>cat</th>
<th>variable</th>
<th>inherent</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>number, case</td>
<td>gender</td>
</tr>
<tr>
<td>A</td>
<td>number, case, gender, degree</td>
<td>position</td>
</tr>
<tr>
<td>V</td>
<td>tense, number, person,...</td>
<td>auxiliary</td>
</tr>
<tr>
<td>V2</td>
<td>as V</td>
<td>complement case</td>
</tr>
<tr>
<td>Adv</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Module structure

Resource module with inflection functions as operations

```plaintext
resource MorphoEng = {oper regV : Str -> V ; ...}
```

Lexicon: abstract and concrete syntax

```plaintext
abstract Lex = {fun Walk : V ; ...}
```

```plaintext
concrete LexEng of Lex =
    open MorphoEng in {lin Walk = regV "walk" ; ...}
```

The same resource can be used (opened) in many lexica.

Abstract and concrete are top-level - they define trees, parsing, linearization.

Resource modules andopers are not top-level - they are "thrown away" after compilation (i.e. not preserved in PGF).

2.3 Example: resource module for English verb inflection

Use the library module Prelude.

Start by defining parameter types and parts of speech.

```plaintext
resource Morpho = open Prelude in {

    param
        VForm = VInf | VPres | VPast | VPastPart | VPresPart ;

    oper
        Verb : Type = {s : VForm => Str} ;

Judgement form oper: auxiliary operation.
```
Start: worst-case function

To save writing and to abstract over the Verb type

\[
\text{mkVerb} : (\_,\_,\_,\_,\_ : \text{Str}) \rightarrow \text{Verb} = \langle \text{go, goes, went, gone, going} \rightarrow \{
\begin{array}{l}
\text{VInf} \rightarrow \text{go} ; \\
\text{VPres} \rightarrow \text{goes} ; \\
\text{VPast} \rightarrow \text{went} ; \\
\text{VPastPart} \rightarrow \text{gone} ; \\
\text{VPresPart} \rightarrow \text{going}
\end{array}\}
\}
\]

Testing computation in resource modules

Import with retain option

> i -retain Morpho.gf

Use command \text{cc} = \text{compute_concrete}

> cc mkVerb "use" "uses" "used" "used" "using"
\{s : \text{Morpho.VForm} \rightarrow \text{Str}
= \text{table} \text{Morpho.VForm} \{
\begin{array}{l}
\text{Morpho.VInf} \rightarrow \text{"use"}; \\
\text{Morpho.VPres} \rightarrow \text{"uses"}; \\
\text{Morpho.VPast} \rightarrow \text{"used"}; \\
\text{Morpho.VPastPart} \rightarrow \text{"used"}; \\
\text{Morpho.VPresPart} \rightarrow \text{"using"}
\end{array}\}
\}

Defining paradigms

A paradigm is an operation of type

\[
\text{Str} \rightarrow \text{Verb}
\]

which takes a string and returns an inflection table.

Let’s first define the paradigm for regular verbs:
regVerb : Str -> Verb = \walk \rightarrow
mkVerb walk (walk + "s") (walk + "ed") (walk + "ed") (walk + "ing") ;

This will work for walk, interest, play.
It will not work for sing, kiss, use, cry, fly, stop.

More paradigms
For verbs ending with s, x, z, ch

s_regVerb : Str -> Verb = \kiss \rightarrow
mkVerb kiss (kiss + "es") (kiss + "ed") (kiss + "ed") (kiss + "ing") ;

For verbs ending with e

e_regVerb : Str -> Verb = \use \rightarrow
let us = init use
in mkVerb use (use + "s") (us + "ed") (us + "ed") (us + "ing") ;

Notice:

• the local definition \texttt{let c = d in ...}

• the operation \texttt{init} from Prelude, dropping the last character

More paradigms still
For verbs ending with y

y_regVerb : Str -> Verb = \cry \rightarrow
let cr = init cry
in mkVerb cry (cr + "ies") (cr + "ied") (cr + "ied") (cry + "ing") ;

For verbs ending with ie

ie_regVerb : Str -> Verb = \die \rightarrow
let dy = Predef.tk 2 die + "y"
in mkVerb die (die + "s") (die + "d") (die + "d") (dy + "ing") ;
What paradigm to choose

If the infinitive ends with \( s, x, z, ch \), choose \texttt{s\_regVerb}: munch, munches
If the infinitive ends with \( y \), choose \texttt{y\_regVerb}: cry, cries, cried

\begin{itemize}
  \item except if a vowel comes before: play, plays, played
\end{itemize}

If the infinitive ends with \( e \), choose \texttt{e\_regVerb}: use, used, using

\begin{itemize}
  \item except if an \( i \) precedes: die, dying
  \item or if an \( e \) precedes: free, freeing
\end{itemize}

2.4 Smart paradigms

Let GF choose the paradigm by \textbf{pattern matching on strings}

\[
\text{smartVerb} : \text{Str} \rightarrow \text{Verb} = \backslash v \rightarrow \text{case } v \text{ of } \{
  \_ \ (+ ("s"|"z"|"x"|"ch")) \Rightarrow \text{s\_regVerb } v ;
  \_ \ (+ "ie") \Rightarrow \text{ie\_regVerb } v ;
  \_ \ (+ "ee") \Rightarrow \text{ee\_regVerb } v ;
  \_ \ (+ "e") \Rightarrow \text{e\_regVerb } v ;
  \_ \ (+ ("a"|"e"|"o"|"u") + "y") \Rightarrow \text{regVerb } v ;
  \_ \ (+ "y") \Rightarrow \text{y\_regVerb } v ;
  \_ \ \Rightarrow \text{regVerb } v
\} ;
\]

Pattern matching on strings

Format: \texttt{case string of } \{ \texttt{pattern } \Rightarrow \texttt{value} \} 

Patterns:

\begin{itemize}
  \item \_ matches any string
  \item a string in quotes matches itself: "ie"
  \item + splits into substrings: \_ + "y"
  \item | matches alternatives: "a"|"e"|"o"
\end{itemize}

Common practice: last pattern a catch-all \_
Testing the smart paradigm

> cc -all smartVerb "munch"
munch munches munched munched munching

> cc -all smartVerb "die"
die dies died died dying

> cc -all smartVerb "agree"
agree agrees agreed agreed agreeing

> cc -all smartVerb "deploy"
deploy deploys deployed deployed deploying

> cc -all smartVerb "classify"
classify classifies classified classified classifying

The smart paradigm is not yet perfect

Irregular verbs are obviously not covered

> cc -all smartVerb "sing"
sing sings singed singed singing

Neither are regular verbs with consonant duplication

> cc -all smartVerb "stop"
stop stops stoped stoped stoping

The final consonant duplication paradigm

Use the Prelude function last

```
dupRegVerb : Str -> Verb = \stop ->
  let stopp = stop + last stop
  in
  mkVerb stop (stop + "s") (stopp + "ed") (stopp + "ed") (stopp + "ing") ;
```

String pattern: relevant consonant preceded by a vowel

```
_ + ("a"|"e"|"i"|"o"|"u") + ("b"|"d"|"g"|"m"|"n"|"p"|"r"|"s"|"t")
  => dupRegVerb v ;
```
Testing consonant duplication

Now it works

```plaintext
> cc -all smartVerb "stop"
  stop stops stopped stopped stopping
```

But what about

```plaintext
> cc -all smartVerb "coat"
  coat coats coatted coatted coatting
```

Solution: a prior case for diphthongs before the last char (? matches one char)

```plaintext
_ + ("ea"|"ee"|"ie"|"oa"|"oo"|"ou") + ? => regVerb v ;
```

There is no waterproof solution

Duplication depends on stress, which is not marked in English:

- *omit* [ˈomɪt]: omitted, omitting
- *vomit* [ˈvɒmɪt]: vomited, vomiting

This means that we occasionally have to give more forms than one.

We knew this already for irregular verbs. And we cannot write patterns for each of them either, because e.g. *lie* can be both *lie, lied, lied* or *lie, lay, lain*.

A paradigm for irregular verbs

Arguments: three forms instead of one.

Pattern matching done in regular verbs can be reused.

```plaintext
irregVerb : (_,_,_ : Str) -> Verb = \sing,sang,sung ->
  let v = smartVerb sing
  in
  mkVerb sing (v.s ! VPres) sang sung (v.s ! VPresPart) ;
```
Putting it all together

We have three functions:

- `smartVerb : Str -> Verb`
- `irregVerb : Str -> Str -> Str -> Verb`
- `mkVerb : Str -> Str -> Str -> Str -> Str -> Verb`

As all types are different, we can use **overloading** and give them all the same name.

### An overloaded paradigm

For documentation: variable names showing examples of arguments.

```plaintext
mkV = overload {
  mkV : (cry : Str) -> Verb = smartVerb ;
  mkV : (sing,sang,sung : Str) -> Verb = irregVerb ;
  mkV : (go,goes,went,gone,going : Str) -> Verb = mkVerb ;
}
```

Testing the overloaded paradigm

```plaintext
> cc -all mkV "lie"
lie lies lied lied lying
> cc -all mkV "lie" "lay" "lain"
lie lies lay lain lying
> cc -all mkV "omit"
omit omits omitted omitted omitting
> cc -all mkV "vomit"
vomit vomits vomitted vomitted vomiting
> cc -all mkV "vomited" "vomited" "vomited"
vomit vomits vomited vomited vomiting
```

Surely we could do better for *vomit*...

### 2.5 Phases of morphology implementation

1. Linearization type, with parametric and inherent features.
2. Worst-case function.
3. The set of paradigms, traditionally taking one argument each.
4. Smart paradigms, with relevant numbers of arguments.
5. Overloaded user function, collecting the smart paradigms.

Other parts of speech

Usually recommended order:
1. Nouns, the simplest class.
2. Adjectives, often using noun inflection, adding gender and degree.
3. Verbs, usually the most complex class, using adjectives in participles.

Morphophonemic functions

Many operations are common to different parts of speech.
Example: adding an s to an English noun or verb.

```haskell
add_s : Str -> Str = \v -> case v of {
    _ + ("s"|"z"|"x"|"ch") => v + "es" ;
    _ + ("a"|"e"|"o"|"u") + "y" => v + "s" ;
    cr + "y" => cr + "ies" ;
    _ => v + "s"
} ;
```

This should be defined separately, not directly in verb conjunctions.
Notice: pattern variable \textit{cr} matches like _ but gets bound.

2.6 Building a lexicon

Boringly, we need abstract and concrete modules even for one language.

```haskell
abstract Lex = {
    cat V ;
    fun
    play_V : V ;
    sleep_V : V ;
} ;

concrete LexEng = open Morpho in {
    lincat V = Verb ;
    lin
    play_V = mkV "play" ;
    sleep_V = mkV "sleep" "slept" "slept" ;
}
```
Fortunately, these modules can be mechanically generated from a POS-tagged word list

\[
\begin{align*}
\text{V play} \\
\text{V sleep slept slept}
\end{align*}
\]

**Bootstrapping a lexicon**

Alt 1. From a morphological POS-tagged word list: trivial

\[
\begin{align*}
\text{V play played played} \\
\text{V sleep slept slept}
\end{align*}
\]

Alt 2. From a plain word list, POS-tagged: start assuming regularity, generate, correct, and add forms by iteration

\[
\begin{align*}
\text{V play} & \implies \text{V play played played} \implies \\
\text{V sleep} & \implies \text{V sleep slept slept slept}
\end{align*}
\]

Example: Finnish nouns need 1.42 forms in average (to generate 26 forms).

### 2.7 Nonconcatenative morphology: Arabic

Semitic languages, e.g. Arabic: *kataba* has forms *kaAtib, yaktubu, ...*

Traditional analysis:

- word = **root** + **pattern**
- root = three consonants (**radicals**)
- pattern = function from root to string (notation: string with variables *F,C,L* for the radicals)

Example: *yaktubu = ktb + yaFCuLu*

Words are datastructures rather than strings!
Datastructures for Arabic

Roots are records of strings.

Root : Type = {F,C,L : Str} ;

Patterns are functions from roots to strings.

Pattern : Type = Root -> Str ;

A special case is filling: a record of strings filling the four slots in a root.

Filling : Type = {F,FC,CL,L : Str} ;

This is enough for everything except middle consonant duplication (e.g. FaCCaLa).

Applying a pattern

A pattern obtained from a filling intertwines the records:

fill : Filling -> Pattern = \p,r ->

Middle consonant duplication also uses a filling but duplicates the C consonant of the root:

dfill : Filling -> Pattern = \p,r ->

Encoding roots by strings

This is just for the ease of programming and writing lexica.

F = first letter, C = second letter, L = the rest.

getRoot : Str -> Root = \s -> case s of {
  F@? + C@? + L => {F = F ; C = C ; L = L} ;
  _ => Predef.error ("cannot get root from" ++ s)
} ;

The as-pattern x@p matches p and binds x.

The error function Predef.error stops computation and displays the string. It is a typical catch-all value.
Encoding patterns by strings

Patterns are coded by using the letters F, C, L.

\[
\text{getPattern} : \text{Str} \rightarrow \text{Pattern} = \lambda s \rightarrow \text{case } s \text{ of } \{ \\
\text{F + "F" + FC + "CC" + CL + "L" + L} => \\
\ \ \ \ \ \ \ \ \ \ \ \text{dfill } \{\text{F = F ; FC = FC ; CL = CL ; L = L}\} \\
\text{F + "F" + FC + "C" + CL + "L" + L} => \\
\ \ \ \ \ \ \ \ \ \ \ \text{fill } \{\text{F = F ; FC = FC ; CL = CL ; L = L}\} \\
\_ => \text{Predef.error ("cannot get pattern from" ++ s)} \\
\} \\
\]

A high-level lexicon building function

Dictionary entry: root + pattern.

\[
\text{getWord} : \text{Str} \rightarrow \text{Str} \rightarrow \text{Str} = \lambda r,p \rightarrow \\
\text{getPattern } p \ (\text{getRoot } r) \\
\]

Now we can try:

\[
\text{> cc getWord "ktb" "yaFCuLu"} \\
\text{"yaktubu"} \\
\text{> cc getWord "ktb" "muFaCCiLu"} \\
\text{"mukattibu"} \\
\]

Parameters for the Arabic verb type

Inflection in tense, number, person, gender.

\[
\text{param} \\
\text{Number} = \text{Sg} | \text{Dl} | \text{Pl} ; \\
\text{Gender} = \text{Masc} | \text{Fem} ; \\
\text{Tense} = \text{Perf} | \text{Impf} ; \\
\text{Person} = \text{Per1} | \text{Per2} | \text{Per3} ; \\
\]

But not in all combinations. For instance: no first person dual.
(We have omitted most tenses and moods.)
Example of Arabic verb inflection

<table>
<thead>
<tr>
<th>Persona</th>
<th>Numerus</th>
<th>Perfectum</th>
<th>Imperfectum</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. masc.</td>
<td>sing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. fem.</td>
<td>sing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. masc.</td>
<td>sing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. fem.</td>
<td>sing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>sing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. masc.</td>
<td>dual.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. fem.</td>
<td>dual.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>dual.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. masc.</td>
<td>plur.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. fem.</td>
<td>plur.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. masc.</td>
<td>plur.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. fem.</td>
<td>plur.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>plur.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Arabic verb type: implementation

We use an algebraic datatype to include only the meaningful combinations.

```plaintext
param VPer =
    Vp3  Number Gender
| Vp2Sg Gender
```
Thus $2 \times (3 \times 2 + 2 + 1 + 2 + 1 + 1) = 26$ forms, not $2 \times 3 \times 2 \times 3 = 36$.

An Arabic verb paradigm

```haskell
pattV_u : Tense -> VPer -> Pattern = \t,v -> getPattern (case t of {
  Perf => case v of {
    Vp3 Sg Masc => "FaCaLa" ;
    Vp3 Sg Fem  => "FaCaLato" ; -- o is the no-vowel sign ("sukun")
    Vp3 Dl Masc => "FaCaLaA" ;
    -- ...
  });
  Impf => case v of {
    -- ...
    Vp1Sg    => "A?aFoCuLu" ;
    Vp1Pl    => "naFoCuLu"
  }
}) ;

u_Verb : Str -> Verb = \s -> {
  s = \t,p => appPattern (getRoot s) (pattV_u t p)
} ;
```

Applying an Arabic paradigm

Testing in the resource module:

```
> cc -all u_Verb "ktb"
kataba katabato katabaA katabataA katabuwA katabona katabota katabotu katabotumaA katabotum katabotunv2a katabotu katabonaA yakotubu takotubu yakotubaAni takotubaAni yakotubuwna yakotubna takotubu takotubiyna takotubaAni takotubuwna takotubona A?aakotubu nakotubu
```

Building a lexicon:

```haskell
fun ktb_V : V ;
lin ktb_V = u_Verb "ktb" ;
```
How we did the printing (recreational GF hacking)

We defined a HTML printing operation

```haskell
oper verbTable : Verb -> Str
```

and used it in a special category `Table` built by

```haskell
fun Tab : V -> Table ;
lin Tab v = verbTable v ;
```

We then used

```bash
> l Tab ktb_V | ps -env=quotes -to_arabic | ps -to_html | wf -file=ara.html
> ! tr "\" " " <ara.html >ar.html
```

2.8 Exercises

1. Learn to use the commands `compute_concrete`, `morpho_analyse`, `morpho_quiz`.
2. Try out some smart paradigms in the resource library files `Paradigms` for some languages you know (or don’t know yet). Use the command `cc` for this.
3. Write a morphology implementation for some word class and some paradigms in your target language. Start with feature design and finish with a smart paradigm.
4. Bootstrap a GF lexicon (abstract + concrete) of 100 words in your target language.
5. (Recreational GF hacking.) Write an operation similar to `verbTable` for printing nice inflection tables in HTML.

3 Basics of a Linguistic Syntax Implementation

Contents

The key categories and rules
Morphology-syntax interface
Examples and variations in English, Italian, French, Finnish, Swedish, German, Hindi
A miniature resource grammar: Italian
Module extension and dependency graphs
Ergativity in Hindi/Urdu

*Don’t worry if the details of this lecture feel difficult! Syntax is difficult and this is why resource grammars are so useful!*

### 3.1 Syntax in the resource grammar

"Linguistic ontology": syntactic structures common to languages

80 categories, 200 functions, which have worked for all resource languages so far

Sufficient for most purposes of expressing meaning: mathematics, technical documents, dialogue systems

Must be extended by language-specific rules to permit parsing of arbitrary text (ca. 10% more in English?)

A lot of work, easy to get wrong!

### 3.2 The key categories and functions

#### The key categories

<table>
<thead>
<tr>
<th>cat</th>
<th>name</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cl</td>
<td>clause</td>
<td><em>every young man loves Mary</em></td>
</tr>
<tr>
<td>VP</td>
<td>verb phrase</td>
<td><em>loves Mary</em></td>
</tr>
<tr>
<td>V2</td>
<td>two-place verb</td>
<td><em>loves</em></td>
</tr>
<tr>
<td>NP</td>
<td>noun phrase</td>
<td><em>every young man</em></td>
</tr>
<tr>
<td>CN</td>
<td>common noun</td>
<td><em>young man</em></td>
</tr>
<tr>
<td>Det</td>
<td>determiner</td>
<td><em>every</em></td>
</tr>
<tr>
<td>AP</td>
<td>adjectival phrase</td>
<td><em>young</em></td>
</tr>
</tbody>
</table>

#### The key functions

<table>
<thead>
<tr>
<th>fun</th>
<th>name</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>PredVP</td>
<td>NP -&gt; VP -&gt; Cl</td>
<td>predication</td>
</tr>
<tr>
<td>ComplV2</td>
<td>V2 -&gt; NP -&gt; VP</td>
<td>complementation</td>
</tr>
<tr>
<td>DetCN</td>
<td>Det -&gt; CN -&gt; NP</td>
<td>determination</td>
</tr>
<tr>
<td>AdjCN</td>
<td>AP -&gt; CN -&gt; CN</td>
<td>modification</td>
</tr>
</tbody>
</table>

35
Feature design

<table>
<thead>
<tr>
<th>cat</th>
<th>variable</th>
<th>inherent</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>tense</td>
<td>-</td>
</tr>
<tr>
<td>VP</td>
<td>tense, agr</td>
<td>-</td>
</tr>
<tr>
<td>V2</td>
<td>tense, agr</td>
<td>case</td>
</tr>
<tr>
<td>NP</td>
<td>case</td>
<td>agr</td>
</tr>
<tr>
<td>CN</td>
<td>number, case</td>
<td>gender</td>
</tr>
<tr>
<td>Det</td>
<td>gender, case</td>
<td>number</td>
</tr>
<tr>
<td>AP</td>
<td>gender, number, case</td>
<td>-</td>
</tr>
</tbody>
</table>

agr = agreement features: gender, number, person

### 3.3 Predication: building clauses

Interplay between features

param Tense, Case, Agr

lincat C1 = \(s : \text{Tense} \Rightarrow \text{Str} \)
lincat NP = \(s : \text{Case} \Rightarrow \text{Str} ; a : \text{Agr} \)
lincat VP = \(s : \text{Tense} \Rightarrow \text{Agr} \Rightarrow \text{Str} \)

fun PredVP : NP -> VP -> Cl

lin PredVP np vp = \(s = \backslash t \Rightarrow np.s ! \text{subj} ++ vp.s ! t ! np.a \)

oper subj : Case

Feature passing

In general, combination rules just pass features: no case analysis (table expressions) is performed.

A special notation is hence useful:

\( \backslash p,q \Rightarrow t \equiv \text{table}\{p \Rightarrow \text{table}\{q \Rightarrow t}\} \)

It is similar to lambda abstraction (\( \backslash x,y \Rightarrow t \) in a function type).
Predication: examples

English

<table>
<thead>
<tr>
<th>np.agr</th>
<th>present</th>
<th>past</th>
<th>future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sg Per1</td>
<td>I sleep</td>
<td>I slept</td>
<td>I will sleep</td>
</tr>
<tr>
<td>Sg Per3</td>
<td>she sleeps</td>
<td>she slept</td>
<td>she will sleep</td>
</tr>
<tr>
<td>Pl Per1</td>
<td>we sleep</td>
<td>we slept</td>
<td>we will sleep</td>
</tr>
</tbody>
</table>

Italian ("I am tired", "she is tired", "we are tired")

<table>
<thead>
<tr>
<th>np.agr</th>
<th>present</th>
<th>past</th>
<th>future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masc Sg Per1</td>
<td>io sono stanco</td>
<td>io ero stanco</td>
<td>io sarò stanco</td>
</tr>
<tr>
<td>Fem Sg Per3</td>
<td>lei è stanca</td>
<td>lei era stanca</td>
<td>lei sarà stanca</td>
</tr>
<tr>
<td>Fem Pl Per1</td>
<td>noi siamo stanche</td>
<td>noi eravamo stanche</td>
<td>noi saremo stanche</td>
</tr>
</tbody>
</table>

Predication: variations

Word order:

- *will I sleep* (English), *è stanca lei* (Italian)

Pro-drop:

- *io sono stanco* vs. *sono stanco* (Italian)

Ergativity:

- ergative case of transitive verb subject; agreement to object (Hindi)

Variable subject case:

- *minä olen lapsi* vs. *minulla on lapsi* (Finnish, "I am a child" (nominative) vs. "I have a child" (adessive))
3.4 Complementation: building verb phrases

Interplay between features

\[ \text{lincat NP} = \{ s : \text{Case} \Rightarrow \text{Str} ; a : \text{Agr} \} \]
\[ \text{lincat VP} = \{ s : \text{Tense} \Rightarrow \text{Agr} \Rightarrow \text{Str} \} \]
\[ \text{lincat V2} = \{ s : \text{Tense} \Rightarrow \text{Agr} \Rightarrow \text{Str} ; c : \text{Case} \} \]

\[
\text{fun ComplV2 : V2} 
\rightarrow \text{NP} \rightarrow \text{VP} 
\]

\[
\text{lin ComplV2 v2 vp} = \{ s = \langle t,a \rangle \Rightarrow v2.s ! t ! a \Rightarrow np.s ! v2.c \}
\]

Complementation: examples

English

<table>
<thead>
<tr>
<th>v2.case</th>
<th>infinitive VP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc</td>
<td>love me</td>
</tr>
<tr>
<td>at + Acc</td>
<td>look at me</td>
</tr>
</tbody>
</table>

Finnish

<table>
<thead>
<tr>
<th>v2.case</th>
<th>VP, infinitive</th>
<th>translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accusative</td>
<td>tavata minut</td>
<td>”meet me”</td>
</tr>
<tr>
<td>Partitive</td>
<td>rakastaa minua</td>
<td>”love me”</td>
</tr>
<tr>
<td>Elative</td>
<td>pitää minusta</td>
<td>”like me”</td>
</tr>
<tr>
<td>Genitive + perään</td>
<td>katsaa minun perään</td>
<td>”look after me”</td>
</tr>
</tbody>
</table>

Complementation: variations

Prepositions: a two-place verb usually involves a preposition in addition case

\[
\text{lincat V2} = \{ s : \text{Tense} \Rightarrow \text{Agr} \Rightarrow \text{Str} ; c : \text{Case} ; \text{prep} : \text{Str} \}
\]

\[
\text{lin ComplV2 v2 vp} = \{ s = \langle t,a \rangle \Rightarrow v2.s ! t ! a \Rightarrow \text{v2.prep} \Rightarrow np.s ! v2.c \}
\]

Clitics: the place of the subject can vary, as in Italian:

- Maria ama Giovanni vs. Maria mi ama (”Mary loves John” vs. ”Mary loves me”)

38
3.5 Determination: building noun phrases

Interplay between features

\[
\begin{align*}
\text{lincat NP} &= \{ s : \text{Case} \Rightarrow \text{Str} \ ; \ a : \text{Agr} \} \\
\text{lincat CN} &= \{ s : \text{Number} \Rightarrow \text{Case} \Rightarrow \text{Str} \ ; \ g : \text{Gender} \} \\
\text{lincat Det} &= \{ s : \text{Gender} \Rightarrow \text{Case} \Rightarrow \text{Str} \ ; \ n : \text{Number} \}
\end{align*}
\]

\[
\text{fun DetCN : Det} \rightarrow \text{CN} \rightarrow \text{NP}
\]

\[
\text{lin DetCN det cn} = \\
\quad \{ \\
\quad \quad \quad \quad \quad \quad \quad s = \\backslash c \Rightarrow \text{det.s} \ ! \ \text{cn.g} \ ! \ c \ ++ \ \text{cn.s} \ ! \ \text{det.n} \ ! \ c \ ; \\
\quad \quad \quad \quad \quad \quad \quad a = \text{agr} \ \text{cn.g} \ \text{det.n} \ \text{Per3}
\}
\]

\[
\text{oper agr : Gender} \rightarrow \text{Number} \rightarrow \text{Person} \rightarrow \text{Agr}
\]

Determination: examples

English

<table>
<thead>
<tr>
<th>Det.num</th>
<th>NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sg</td>
<td>every house</td>
</tr>
<tr>
<td>Pl</td>
<td>these houses</td>
</tr>
</tbody>
</table>

Italian (”this wine”, ”this pizza”, ”those pizzas”)

<table>
<thead>
<tr>
<th>Det.num</th>
<th>CN.gen</th>
<th>NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sg</td>
<td>Masc</td>
<td>questo vino</td>
</tr>
<tr>
<td>Sg</td>
<td>Fem</td>
<td>questa pizza</td>
</tr>
<tr>
<td>Pl</td>
<td>Fem</td>
<td>quelle pizze</td>
</tr>
</tbody>
</table>

Finnish (”every house”, ”these houses”)

<table>
<thead>
<tr>
<th>Det.num</th>
<th>NP, nominative</th>
<th>NP, inessive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sg</td>
<td>jokainen talo</td>
<td>jokaisessa talossa</td>
</tr>
<tr>
<td>Pl</td>
<td>nämä talot</td>
<td>näissä taloissa</td>
</tr>
</tbody>
</table>
Determinations: variations

Systematic number variation:

- *this-these, the-the, il-i* (Italian "the-the")

"Zero" determiners:

- *talo" ("a house") vs. *talo" ("the house") (Finnish)
- *a house* vs. *houses* (English), *une maison* vs. *des maisons* (French)

Specificity parameter of nouns:

- *varje hus* vs. *det huset* (Swedish, "every house" vs. "that house")

3.6 Modification: adding adjectives to nouns

Interplay between features

\[
\begin{align*}
lincat \ AP &= \left\{ s : \text{Gender} \rightarrow \text{Number} \rightarrow \text{Case} \rightarrow \text{Str} \right\} \\
lincat \ CN &= \left\{ s : \text{Number} \rightarrow \text{Case} \rightarrow \text{Str} ; g : \text{Gender} \right\} \\
\end{align*}
\]

fun AdjCN : AP -> CN -> CN

lin AdjCN ap cn = {
    s = \n,c => ap.s ! cn.g ! n ! c ++ cn.s ! n ! c ;
    g = cn.g
}

Modification: examples

English

<table>
<thead>
<tr>
<th>CN, singular</th>
<th>CN, plural</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>new house</em></td>
<td><em>new houses</em></td>
</tr>
</tbody>
</table>

Italian ("red wine", "red house")

<table>
<thead>
<tr>
<th>CN.gen</th>
<th>CN, singular</th>
<th>CN, plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masc</td>
<td><em>vino rosso</em></td>
<td><em>vini rossi</em></td>
</tr>
<tr>
<td>Fem</td>
<td><em>casa rossa</em></td>
<td><em>case rosse</em></td>
</tr>
</tbody>
</table>
Finnish ("red house")

<table>
<thead>
<tr>
<th>CN, sg, nominative</th>
<th>CN, sg, ablative</th>
<th>CN, pl, essive</th>
</tr>
</thead>
<tbody>
<tr>
<td>punainen talo</td>
<td>punaiselta talolta</td>
<td>punaisina taloina</td>
</tr>
</tbody>
</table>

**Modification: variations**

The place of the adjectival phrase

- Italian: *casa rossa, vecchia casa* ("red house", "old house")
- English: *old house, house similar to this*

**Specificity parameter of the adjective**

- German: *ein rotes Haus vs. das rote Haus* ("a red house" vs. "the red house")

### 3.7 Lexical insertion

To "get started" with each category, use words from lexicon.

There are **lexical insertion functions** for each lexical category:

- UseN : N -> CN
- UseA : A -> AP
- UseV : V -> VP

The linearization rules are often trivial, because the lin cats match

- \text{lin UseN } n = n
- \text{lin UseA } a = a
- \text{lin UseV } v = v

However, for UseV in particular, this will usually be more complex.

**The head of a phrase**

The inserted word is the **head** of the phrases built from it:

- *house* is the head of *house, big house, big old house* etc
As a rule with many exceptions and modifications,

- variable features are passed from the phrase to the head
- inherent features of the head are inherited by the noun

This works for **endocentric** phrases: the head has the same type as the full phrase.

### What is the head of a noun phrase?

In an **NP** of form Det CN, is Det or CN the head?

Neither, really, because features are passed in both directions:

```plaintext
lin DetCN det cn = {
    s = \c => det.s ! cn.g ! c ++ cn.s ! det.n ! c ;
    a = agr cn.g det.n Per3
}
```

Moreover, this **NP** is **exocentric**: no part is of the same type as the whole.

### Structural words

**Structural words** = **function words**, words with special grammatical functions

- determiners: *the, this, every*
- pronouns: *I, she*
- conjunctions: *and, or, but*

Often members of **closed classes**, which means that new words are never (or seldom) introduces to them.

Linearization types are often specific and inflection are irregular.

### 3.8 A miniature resource grammar for Italian

We divide it to five modules - much fewer than the full resource!
abstract Grammar -- syntactic cats and funs
abstract Lang = Grammar **... -- test lexicon added to Grammar
resource ResIta -- resource for Italian
concrete GrammarIta of Grammar = open ResIta in... -- Italian syntax
concrete LangIta of Lang = GrammarIta ** open ResIta in... -- It. lexicon

Extension vs. opening

Module extension: $N = M_1, M_2, M_3 ** \{\ldots\}$

- module $N$ inherits all judgements from $M_1, M_2, M_3$

Module opening: $N = \text{open } R_1, R_2, R_3 \text{ in } \{\ldots\}$

- module $N$ can use all judgements from $R_1, R_2, R_3$ (but doesn’t inherit them)

Module dependencies

rectangle = abstract, solid ellipse = concrete, dashed ellipse = resource
Producing the dependency graph

Using the command `dg = dependency_graph` and graphviz

```
> i -retain TestIta.gf
> dependency_graph
wrote graph in file _gfdepgraph.dot
> ! dot -Tjpg _gfdepgraph.dot >testdep.jpg
```

Before calling `dot`, removed the module `Predef` to save space.

The module Grammar

```
abstract Grammar = {
  cat
  Cl ; NP ; VP ; AP ; CN ; Det ; N ; A ; V ; V2 ;
  fun
  PredVP : NP -> VP -> Cl ;
  ComplV2 : V2 -> NP -> VP ;
  DetCN : Det -> CN -> NP ;
  ModCN : CN -> AP -> CN ;
  UseV : V -> VP ;
  UseN : N -> CN ;
  UseA : A -> AP ;
  a_Det, the_Det : Det ; this_Det, these_Det : Det ;
  i_NP, she_NP, we_NP : NP ;
}
```

Parameters

Parameters are defined in `ResIta.gf`. Just 11 of the 56 verb forms.

```
Number = Sg | Pl ;
Gender = Masc | Fem ;
Case = Nom | Acc | Dat ;
Aux = Avere | Essere ; -- the auxiliary verb of a verb
Tense = Pres | Perf ;
Person = Per1 | Per2 | Per3 ;
Agr = Ag Gender Number Person ;
```
3.9 Italian verb phrases

Tense and agreement of a verb phrase, in syntax

<table>
<thead>
<tr>
<th>UseV</th>
<th>arrive_V</th>
<th>Pres</th>
<th>Perf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ag Masc Sg Per1</td>
<td>arrivo</td>
<td>sono arrivato</td>
<td></td>
</tr>
<tr>
<td>Ag Fem Sg Per1</td>
<td>arrivo</td>
<td>sono arrivata</td>
<td></td>
</tr>
<tr>
<td>Ag Masc Sg Per2</td>
<td>arrivì</td>
<td>sei arrivato</td>
<td></td>
</tr>
<tr>
<td>Ag Fem Sg Per2</td>
<td>arrivì</td>
<td>sei arrivata</td>
<td></td>
</tr>
<tr>
<td>Ag Masc Sg Per3</td>
<td>arriva</td>
<td>è arrivato</td>
<td></td>
</tr>
<tr>
<td>Ag Fem Sg Per3</td>
<td>arriva</td>
<td>è arrivata</td>
<td></td>
</tr>
<tr>
<td>Ag Masc Pl Per1</td>
<td>arriviamo</td>
<td>siamo arrivati</td>
<td></td>
</tr>
<tr>
<td>Ag Fem Pl Per1</td>
<td>arriviamo</td>
<td>siamo arrivate</td>
<td></td>
</tr>
<tr>
<td>Ag Masc Pl Per2</td>
<td>arrivate</td>
<td>siete arrivati</td>
<td></td>
</tr>
<tr>
<td>Ag Fem Pl Per2</td>
<td>arrivate</td>
<td>siete arrivate</td>
<td></td>
</tr>
<tr>
<td>Ag Masc Pl Per3</td>
<td>arrivano</td>
<td>sono arrivati</td>
<td></td>
</tr>
<tr>
<td>Ag Fem Pl Per3</td>
<td>arrivano</td>
<td>sono arrivate</td>
<td></td>
</tr>
</tbody>
</table>

The forms of a verb, in morphology

<table>
<thead>
<tr>
<th>arrive_V</th>
<th>form</th>
</tr>
</thead>
<tbody>
<tr>
<td>VInf</td>
<td>arrivare</td>
</tr>
<tr>
<td>VPres Sg Per1</td>
<td>arrivo</td>
</tr>
<tr>
<td>VPres Sg Per2</td>
<td>arrivì</td>
</tr>
<tr>
<td>VPres Sg Per3</td>
<td>arriva</td>
</tr>
<tr>
<td>VPres Pl Per1</td>
<td>arriviamo</td>
</tr>
<tr>
<td>VPres Pl Per2</td>
<td>arrivate</td>
</tr>
<tr>
<td>VPres Pl Per3</td>
<td>arrivano</td>
</tr>
<tr>
<td>VPart Masc Sg</td>
<td>arrivato</td>
</tr>
<tr>
<td>VPart Fem Sg</td>
<td>arrivata</td>
</tr>
<tr>
<td>VPart Masc Pl</td>
<td>arrivati</td>
</tr>
<tr>
<td>VPart Fem Pl</td>
<td>arrivate</td>
</tr>
</tbody>
</table>

Inherent feature: aux is essere.

The verb phrase type

Lexical insertion maps V to VP.
Two possibilities for VP: either close to Cl,

\[
\text{lincat VP} = \{s : \text{Tense} \Rightarrow \text{Agr} \Rightarrow \text{Str}\}
\]
or close to V, just adding a clitic and an object to verb,

\[
\text{lincat VP} = \{v : \text{Verb} ; \text{clit} : \text{Str} ; \text{obj} : \text{Str}\};
\]

We choose the latter. It is more efficient in parsing.

**Verb phrase formation**

Lexical insertion is trivial.

\[
\text{lin UseV v} = \{v = v ; \text{clit}, \text{obj} = []\}
\]

Complementation assumes NP has a clitic and an ordinary object part.

\[
\text{lin ComplV2} =
\begin{align*}
\text{let} & \\
\text{nps} & = \text{np.s} ! v2.c \\
\text{in} & \\
\text{v} & = \{s = v2.s ; \text{aux} = v2.aux\}; \\
\text{clit} & = \text{nps.clit} ; \\
\text{obj} & = \text{nps.obj}
\end{align*}
\]

### 3.10 Italian noun phrases

Being clitic depends on case

\[
\text{lincat NP} = \{s : \text{Case} \Rightarrow \{\text{clit, obj} : \text{Str}\} ; \text{a} : \text{Agr}\};
\]

Examples:

\[
\text{lin she_NP} = \{
\begin{align*}
\text{s} & = \text{table} \{
\text{Nom} & \Rightarrow \{\text{clit} = [] ; \text{obj} = "lei"\} ; \\
\text{Acc} & \Rightarrow \{\text{clit} = "la" ; \text{obj} = []\} ; \\
\text{Dat} & \Rightarrow \{\text{clit} = "le" ; \text{obj} = []\}
\end{align*}
\}
\]

46
Noun phrases: alternatively

Use a feature instead of separate fields,

```haskell
lincat NP = {s : Case => {s : Str ; isClit : Bool} ; a : Agr} ;
```

The use of separate fields is more efficient and scales up better to multiple clitic positions.

Determination

No surprises

```haskell
lincat Det = {s : Gender => Case => Str ; n : Number} ;
```

```haskell
lin DetCN det cn = {
  s = \c => {obj = det.s ! cn.g ! c ++ cn.s ! det.n ; clit = []} ;
    a = Ag cn.g det.n Per3
  } ;
```

Building determiners

Often from adjectives:

```haskell
lin this_Det = adjDet (mkA "questo") Sg ;
lin these_Det = adjDet (mkA "questo") Pl ;
```

```haskell
oper prepCase : Case -> Str = \c -> case c of {
  Dat => "a" ;
  _ => []
```
\oper\ \text{adjDet} : \text{Adj} \rightarrow \text{Number} \rightarrow \text{Determiner} = \adj,n \rightarrow \{ \\
\quad s = \\\g,c \Rightarrow \text{prepCase} \ c ++ \adj.s ! g ! n ; \\
\quad n = n \\
\}\ ; \\

\text{Articles: see GrammarIta.gf} \\

\textbf{Adjectival modification} \\

Recall the inherent feature for position \\

\text{lincat AP = \{}s : \text{Gender} \Rightarrow \text{Number} \Rightarrow \text{Str} ; \text{isPre} : \text{Bool}\}\{ \\
\text{lin ModCN cn ap = } \{ \\
\quad s = \\n \Rightarrow \text{preOrPost} \ \text{ap.isPre} \ (\text{ap.s} ! \text{cn.g} ! \text{n}) \ (\text{cn.s} ! \text{n}) ; \\
\quad g = \text{cn.g} \\
\}\ ; \\

\text{Obviously, separate pre- and post- parts could be used instead.} \\

\textbf{Italian morphology} \\

Complex but mostly great fun: \\

\text{regNoun : Str \rightarrow Noun = \{}\text{vino} \rightarrow \text{case vino of } \{ \\
\quad \text{fuo} + \ c@("c"\|"g") + "o" \Rightarrow \text{mkNoun} \ \text{vino} \ (\text{fuo} + \ c + "hi") \ \text{Masc} ; \\
\quad \text{ol} + "io" \Rightarrow \text{mkNoun} \ \text{vino} \ (\text{ol} + "i") \ \text{Masc} ; \\
\quad \text{vin} + "o" \Rightarrow \text{mkNoun} \ \text{vino} \ (\text{vin} + "i") \ \text{Masc} ; \\
\quad \text{cas} + "a" \Rightarrow \text{mkNoun} \ \text{vino} \ (\text{cas} + "e") \ \text{Fem} ; \\
\quad \text{pan} + "e" \Rightarrow \text{mkNoun} \ \text{vino} \ (\text{pan} + "i") \ \text{Masc} ; \\
\quad _ \Rightarrow \text{mkNoun} \ \text{vino} \ \text{vino} \ \text{Masc} \\
\}\ ; \\

\text{See ResIta for more details.} \\

\textbf{3.11 Predication, at last} \\

Place the object and the clitic, and select the verb form.
lin PredVP np vp =
  let
    subj = (np.s ! Nom).obj ;
    obj = vp.obj ;
    clit = vp.clit ;
    verb = table {
      Pres => agrV vp.v np.a ;
      Perf => agrV (auxVerb vp.v.aux) np.a ++ agrPart vp.v np.a
    }
  in {
    s = \t => subj ++ clit ++ verb ! t ++ obj
  } ;

Selection of verb form

We need it for the present tense

oper agrV : Verb -> Agr -> Str = \v,a -> case a of {
  Ag _ n p => v.s ! VPres n p
} ;

The participle agrees to the subject, if the auxiliary is essere

oper agrPart : Verb -> Agr -> Str = \v,a -> case v.aux of {
  Avere => v.s ! VPart Masc Sg ;
  Essere => case a of {
    Ag g n _ => v.s ! VPart g n
  }
} ;

3.12 To do

Full details of the core resource grammar are in ResIta (150 loc) and GrammarIta (80 loc).

One thing is not yet done correctly: agreement of participle to accusative clitic object: now it gives io la ho amato, and not io la ho amata.

This is left as an exercise!

3.13 Ergativity in Hindi/Urdu

Normally, the subject is nominative and the verb agrees to the subject.
However, in the perfective tense:

- the subject of a transitive verb is in an ergative “case” (particle *ne*)
- the verb agrees to the object

Example: "the boy/girl eats the apple/bread"

<table>
<thead>
<tr>
<th>subj</th>
<th>obj</th>
<th>gen. present</th>
<th>perfective</th>
</tr>
</thead>
</table>
| Masc | Masc | *ladka: seb Ka:ta: hai* | *ladke ne seb Ka:ya:*
| Masc | Fem | *ladka: roTi: Ka:ta: hai* | *ladke ne roTi: Ka:yi:*
| Fem  | Masc | *ladki: seb Ka:ti: hai* | *ladki: ne seb Ka:ya:*
| Fem  | Fem | *ladki: roTi: Ka:ti: hai* | *ladki: ne roTi: Ka:yi:*

A Hindi clause in different tenses

<table>
<thead>
<tr>
<th>Verb Phrase</th>
<th>Hindi</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPGenPres True</td>
<td>लड़की सेब खाती है</td>
</tr>
<tr>
<td>VPGenPres False</td>
<td>लड़की सेब नहीं खाती है</td>
</tr>
<tr>
<td>VPImpPast True</td>
<td>लड़की सेब खाती थी</td>
</tr>
<tr>
<td>VPImpPast False</td>
<td>लड़की सेब नहीं खाती थी</td>
</tr>
<tr>
<td>VPContPres True</td>
<td>लड़की सेब खा रही है</td>
</tr>
<tr>
<td>VPContPres False</td>
<td>लड़की सेब नहीं खा रही है</td>
</tr>
<tr>
<td>VPContPast True</td>
<td>लड़की सेब खा रही थी</td>
</tr>
<tr>
<td>VPContPast False</td>
<td>लड़की सेब नहीं खा रही थी</td>
</tr>
<tr>
<td>VPPartPres True</td>
<td>लड़की ने सेब खाया</td>
</tr>
<tr>
<td>VPPartPres False</td>
<td>लड़की ने सेब नहीं खाया</td>
</tr>
<tr>
<td>VPPartPast True</td>
<td>लड़की सेब खायी है</td>
</tr>
<tr>
<td>VPPartPast False</td>
<td>लड़की सेब नहीं खायी है</td>
</tr>
<tr>
<td>VPSubj True</td>
<td>लड़की सेब खाये</td>
</tr>
<tr>
<td>VPSubj False</td>
<td>लड़की सेब न खाये</td>
</tr>
<tr>
<td>VPFut True</td>
<td>लड़की सेब खायेगी</td>
</tr>
<tr>
<td>VPFut False</td>
<td>लड़की सेब न खायेगी</td>
</tr>
</tbody>
</table>
3.14 Exercises

1. Learn the commands `dependency_graph`, `print_grammar`, system escape !, and system pipe ?.

2. Write tables of examples of the key syntactic functions for your target languages, trying to include all possible forms.

3. Implement `Grammar` and `Test` for your target language.

4. Even if you don’t know Italian, you may try this: add a parameter or something in `GrammarIta` to implement the rule that the participle in the perfect tense agrees in gender and number with an accusative clitic. Test this with the sentences `lei la ha amata` and `lei ci ha amati` (where the current grammar now gives `amato` in both cases).


4 Using the Resource Grammar Library in Applications

Contents

Software libraries: programmer’s vs. users view
Semantic vs. syntactic grammars
Example of semantic grammar and its implementation
Interfaces and parametrized modules
Free variation
Overview of the Resource Grammar API

4.1 Software libraries

Collections of reusable functions/types/classes

API = Application Programmer’s Interface

- show enough to enable use
- hide details

Example: maps (lookup tables, hash maps) in Haskell, C++, Java, ...
Hidden: the definition of the type `Map` and of the functions `lookup` and `update`.

**Advantages of software libraries**

Programmers have

- less code to write (e.g. *how* to look up)
- less techniques to learn (e.g. efficient Map datastructures)

Improvements and bug fixes can be inherited

**Grammars as software libraries**

Smart paradigms as API for morphology

```
mkN : (talo : Str) -> N
```

Abstract syntax as API for syntactic combinations

```
PredVP : NP -> VP -> Cl
Comp1V2 : V2 -> NP -> VP
NumCN : Num -> CN -> NP
```

### 4.2 Using the library: natural language output

Task: in an email program, generate phrases saying *you have n message(s)*

Problem: avoid *you have one messages*

Solution: use the library

```
PredVP youSg_NP (Comp1V2 have_V2 (NumCN two_Num (UseN (mkN "message"))))
```

```plaintext
===> you have two messages
```

```
PredVP youSg_NP (Comp1V2 have_V2 (NumCN one_Num (UseN (mkN "message"))))
```

```plaintext
===> you have one message
```
Software localization

Adapt the email program to Italian, Swedish, Finnish...

\[
\text{PredVP youSg\_NP (ComplV2 have\_V2 (NumCN two\_Num (UseN (mkN "messaggio"))))} \\
\text{====> hai due messaggi}
\]

\[
\text{PredVP youSg\_NP (ComplV2 have\_V2 (NumCN two\_Num (UseN (mkN "meddelande"))))} \\
\text{====> du har två meddelanden}
\]

\[
\text{PredVP youSg\_NP (ComplV2 have\_V2 (NumCN two\_Num (UseN (mkN "viesti"))))} \\
\text{====> sinulla on kaksi viestiä}
\]

The new languages are more complex than English - but only internally, not on the API level!

Correct number in Arabic

(From “Implementation of the Arabic Numerals and their Syntax in GF” by Ali Dada, ACL workshop on Arabic, Prague 2007)

Use cases for grammar libraries

Grammars need very much very special knowledge, and a lot of work - thus an excellent topic for a software library!

Some applications where grammars have shown to be useful:

- software localization
• natural language generation (from formalized content)
• technical translation
• spoken dialogue systems

4.3 Two kinds of grammarians

Application grammarians vs. resource grammarians

<table>
<thead>
<tr>
<th>grammar</th>
<th>applications</th>
<th>resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>expertise</td>
<td>application domain</td>
<td>linguistics</td>
</tr>
<tr>
<td>programming skills</td>
<td>programming in general</td>
<td>GF programming</td>
</tr>
<tr>
<td>language skills</td>
<td>practical use</td>
<td>theoretical knowledge</td>
</tr>
</tbody>
</table>

We want a division of labour.

4.4 Meaning-preserving translation

Translation must preserve meaning.
It need not preserve syntactic structure.
Sometimes it is even impossible:

- John likes Mary in Italian is Maria piace a Giovanni

The abstract syntax in the semantic grammar is a logical predicate:

fun Like : Person -> Person -> Fact
lin Like x y = x ++ "likes" ++ y -- English
lin Like x y = y ++ "piace" ++ "a" ++ x -- Italian
Translation and resource grammar

To get all grammatical details right, we use resource grammar and not strings

\[
\text{lin\ cat Person = NP ; Fact = Cl ;}
\]

\[
\begin{align*}
\text{lin Like x y} &= \text{PredVP x (ComplV2 like_V2 y)} \quad \text{-- English} \\
\text{lin Like x y} &= \text{PredVP y (ComplV2 piacere_V2 x)} \quad \text{-- Italian}
\end{align*}
\]

From syntactic point of view, we perform transfer, i.e. structure change.

GF has compile-time transfer, and uses interlingua (semantic abstrac syntax) at run time.

Domain semantics

"Semantics of English", or of any other natural language as a whole, has never been built.

It is more feasible to have semantics of fragments - of small, well-understood parts of natural language.

Such languages are called domain languages, and their semantics, domain semantics.

Domain semantics = ontology in the Semantic Web terminology.

Examples of domain semantics

Expressed in various formal languages

- mathematics, in predicate logic
- software functionality, in UML/OCL
- dialogue system actions, in SISR
- museum object descriptions, in OWL

GF abstract syntax can be used for any of these!
4.5 Example: abstract syntax for a "Face" community

What messages can be expressed on the community page?

abstract Face = {

flags startcat = Message ;

cat
   Message ; Person ; Object ; Number ;

fun
   Have : Person -> Number -> Object -> Message ; -- p has n o's
   Like : Person -> Object -> Message ; -- p likes o
   You : Person ;
   Friend, Invitation : Object ;
   One, Two, Hundred : Number ;
}

Notice the startcat flag, as the start category isn’t S.

Presenting the resource grammar

In practice, the abstract syntax of Resource Grammar is inconvenient

- too deep structures, too much code to write
- too many names to remember

We do the same as in morphology: overloaded operations, named mkC where C is the value category.

The resource defines e.g.

mkCl : NP -> V2 -> NP -> C1 = \subj,verb,obj ->
       PredVP subj (ComplV2 verb obj)
mkCl : NP -> V -> C1 = \subj,verb ->
       PredVP subj (UseV verb)

Relevant part of Resource Grammar API for "Face"

These functions (some of which are structural words) are used.
<table>
<thead>
<tr>
<th>Function</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>mkCl : NP -&gt; V2 -&gt; NP -&gt; Cl</td>
<td><em>John loves Mary</em></td>
</tr>
<tr>
<td>mkNP : Numeral -&gt; CN -&gt; NP</td>
<td><em>five cars</em></td>
</tr>
<tr>
<td>mkNP : Quant -&gt; CN -&gt; NP</td>
<td><em>that car</em></td>
</tr>
<tr>
<td>mkNP : Pron -&gt; NP</td>
<td><em>we</em></td>
</tr>
<tr>
<td>mkCN : N -&gt; CN</td>
<td><em>car</em></td>
</tr>
<tr>
<td>this_Quant : Quant</td>
<td><em>this, these</em></td>
</tr>
<tr>
<td>youSg_Pron : Pron</td>
<td><em>you (singular)</em></td>
</tr>
<tr>
<td>n1_Numeral, n2_Numeral : Numeral</td>
<td><em>one, two</em></td>
</tr>
<tr>
<td>n100_Numeral : Numeral</td>
<td><em>one hundred</em></td>
</tr>
<tr>
<td>have_V2 : V2</td>
<td><em>have</em></td>
</tr>
</tbody>
</table>

Concrete syntax for English

How are messages expressed by using the library?

custom FaceEng of Face = open SyntaxEng, ParadigmsEng in {
  lincat
    Message = Cl ;
    Person = NP ;
    Object = CN ;
    Number = Numeral ;
  lin
    Have p n o = mkCl p have_V2 (mkNP n o) ;
    Like p o = mkCl p like_V2 (mkNP this_Quant o) ;
    You = mkNP youSg_Pron ;
    Friend = mkCN friend_N ;
    Invitation = mkCN invitation_N ;
    One = n1_Numeral ;
    Two = n2_Numeral ;
    Hundred = n100_Numeral ;
  oper
    like_V2 = mkV2 "like" ;
    invitation_N = mkN "invitation" ;
    friend_N = mkN "friend" ;
}

Concrete syntax for Finnish

Exactly the same rules of combination, only different words:

custom FaceFin of Face = open SyntaxFin, ParadigmsFin in {
  lincat

4.6 Functors and interfaces

English and Finnish: the same combination rules, only different words!

Can we avoid repetition of the lincat and lin code? Yes!

New module type: functor, a.k.a. incomplete or parametrized module

    incomplete concrete FaceI of Face = open Syntax, LexFace in ...

A functor may open interfaces.

An interface has oper declarations with just a type, no definition.

Here, Syntax and LexFace are interfaces.

The domain lexicon interface

Syntax is the Resource Grammar interface, and gives

- combination rules
- structural words

Content words are not given in Syntax, but in a domain lexicon
interface LexFace = open Syntax in {

oper
  like_V2 : V2 ;
  invitation_N : N ;
  friend_N : N ;
}

Concrete syntax functor "FaceI"

incomplete concrete FaceI of Face = open Syntax, LexFace in {

lincat
  Message = Cl ;
  Person = NP ;
  Object = CN ;
  Number = Numeral ;
lin
  Have p n o = mkCl p have_V2 (mkNP n o) ;
  Like p o = mkCl p like_V2 (mkNP this_Quant o) ;
  You = mkNP youSg_Pron ;
  Friend = mkCN friend_N ;
  Invitation = mkCN invitation_N ;
  One = n1_Numeral ;
  Two = n2_Numeral ;
  Hundred = n100_Numeral ;
}

An English instance of the domain lexicon

Define the domain words in English

instance LexFaceEng of LexFace = open SyntaxEng, ParadigmsEng in {

oper
  like_V2 = mkV2 "like" ;
  invitation_N = mkN "invitation" ;
  friend_N = mkN "friend" ;
}
Put everything together: functor instantiation

Instantiate the functor FaceI by giving instances to its interfaces

```latex
--# -path=.:present

congcrete FaceEng of Face = FaceI with
  (Syntax = SyntaxEng),
  (LexFace = LexFaceEng);
```

Also notice the domain search path.

Porting the grammar to Finnish

1. Domain lexicon: use Finnish paradigms and words

```latex
instance LexFaceFin of LexFace = open SyntaxFin, ParadigmsFin in {
  oper
    like_V2 = mkV2 (mkV "pitää") elative ;
    invitation_N = mkN "kutsu" ;
    friend_N = mkN "ystävä" ;
}
```

2. Functor instantiation: mechanically change Eng to Fin

```latex
--# -path=.:present

congcrete FaceFin of Face = FaceI with
  (Syntax = SyntaxFin),
  (LexFace = LexFaceFin);
```

4.7 Modules of a domain grammar: ”Face” community

1. Abstract syntax, Face
2. Parametrized concrete syntax: FaceI
3. Domain lexicon interface: LexFace
4. For each language $L$: domain lexicon instance LexFace$L$
5. For each language $L$: concrete syntax instantiation Face$L$
Module dependency graph

Porting the grammar to Italian

1. Domain lexicon: use Italian paradigms and words

```
instance LexFaceIta of LexFace = open SyntaxIta, ParadigmsIta in {
    oper
        like_V2 = mkV2 (mkV (piacere_64 "piacere")) dative ;
        invitation_N = mkN "invito" ;
        friend_N = mkN "amico" ;
    }
```

2. Functor instantiation: restricted inheritance, excluding Like

```
concrete FaceIta of Face = FaceI - [Like] with
    (Syntax = SyntaxIta),
    (LexFace = LexFaceIta) ** open SyntaxIta in {
    lin Like p o =
        mkCl (mkNP this_Quant o) like_V2 p ;
    }
```

4.8 Free variation

There can be many ways of expressing a given semantic structure.
This can be expressed by the variant operator |.
fun BuyTicket : City -> City -> Request

lin BuyTicket x y =
   (("I want" ++ (("to buy" | []) ++ ("a ticket"))) | "to go")
   |
   ("can you" | []) ++ "give me" ++ "a ticket")
   |
   [] ++ "from" ++ x ++ "to" ++ y

The variants can of course be resource grammar expressions as well.

4.9 Overview of the resource grammar API

For the full story, see the resource grammar synopsis in grammaticalframework.org/lib/doc/synopsis.html

Main division:

- Syntax, common to all languages
- Paradigms $L$, specific to language $L$

Main categories and their dependencies
Categories of complex phrases

<table>
<thead>
<tr>
<th>Category</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>sequence of utterances</td>
<td><em>Does John walk? Yes.</em></td>
</tr>
<tr>
<td>Utt</td>
<td>utterance</td>
<td><em>does John walk</em></td>
</tr>
<tr>
<td>Imp</td>
<td>imperative</td>
<td><em>don't walk</em></td>
</tr>
<tr>
<td>S</td>
<td>sentence (fixed tense)</td>
<td><em>John wouldn’t walk</em></td>
</tr>
<tr>
<td>QS</td>
<td>question sentence</td>
<td><em>who hasn’t walked</em></td>
</tr>
<tr>
<td>Cl</td>
<td>clause (variable tense)</td>
<td><em>John walks</em></td>
</tr>
<tr>
<td>QCl</td>
<td>question clause</td>
<td><em>who doesn’t walk</em></td>
</tr>
<tr>
<td>VP</td>
<td>verb phrase</td>
<td><em>love her</em></td>
</tr>
<tr>
<td>AP</td>
<td>adjectival phrase</td>
<td><em>very young</em></td>
</tr>
<tr>
<td>CN</td>
<td>common noun phrase</td>
<td><em>young man</em></td>
</tr>
<tr>
<td>Adv</td>
<td>adverbial phrase</td>
<td><em>in the house</em></td>
</tr>
</tbody>
</table>

Lexical categories for building predicates

<table>
<thead>
<tr>
<th>Cat</th>
<th>Explanation</th>
<th>Compl</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>one-place adjective</td>
<td>-</td>
<td><em>smart</em></td>
</tr>
<tr>
<td>A2</td>
<td>two-place adjective</td>
<td>NP</td>
<td><em>married (to her)</em></td>
</tr>
<tr>
<td>Adv</td>
<td>adverb</td>
<td>-</td>
<td><em>here</em></td>
</tr>
<tr>
<td>N</td>
<td>common noun</td>
<td>-</td>
<td><em>man</em></td>
</tr>
<tr>
<td>N2</td>
<td>relational noun</td>
<td>NP</td>
<td><em>friend (of John)</em></td>
</tr>
<tr>
<td>NP</td>
<td>noun phrase</td>
<td>-</td>
<td><em>the boss</em></td>
</tr>
<tr>
<td>V</td>
<td>one-place verb</td>
<td>-</td>
<td><em>sleep</em></td>
</tr>
<tr>
<td>V2</td>
<td>two-place verb</td>
<td>NP</td>
<td><em>love (her)</em></td>
</tr>
<tr>
<td>V3</td>
<td>three-place verb</td>
<td>NP, NP</td>
<td><em>show (it to me)</em></td>
</tr>
<tr>
<td>VS</td>
<td>sentence-complement verb</td>
<td>S</td>
<td><em>say (that I run)</em></td>
</tr>
<tr>
<td>VV</td>
<td>verb-complement verb</td>
<td>VP</td>
<td><em>want (to run)</em></td>
</tr>
</tbody>
</table>
### Functions for building predication clauses

<table>
<thead>
<tr>
<th>Fun</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>mkCl</td>
<td>NP -&gt; V -&gt; Cl</td>
<td>John walks</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; V2 -&gt; NP -&gt; Cl</td>
<td>John loves her</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; V3 -&gt; NP -&gt; NP -&gt; Cl</td>
<td>John sends it to her</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; VV -&gt; VP -&gt; Cl</td>
<td>John wants to walk</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; VS -&gt; S -&gt; Cl</td>
<td>John says that it is good</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; A -&gt; Cl</td>
<td>John is old</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; A -&gt; NP -&gt; Cl</td>
<td>John is older than Mary</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; A2 -&gt; NP -&gt; Cl</td>
<td>John is married to her</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; AP -&gt; Cl</td>
<td>John is very old</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; N -&gt; Cl</td>
<td>John is a man</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; CN -&gt; Cl</td>
<td>John is an old man</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; NP -&gt; Cl</td>
<td>John is the man</td>
</tr>
<tr>
<td>mkCl</td>
<td>NP -&gt; Adv -&gt; Cl</td>
<td>John is here</td>
</tr>
</tbody>
</table>

### Noun phrases and common nouns

<table>
<thead>
<tr>
<th>Fun</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>mkNP</td>
<td>Quant -&gt; CN -&gt; NP</td>
<td>this man</td>
</tr>
<tr>
<td>mkNP</td>
<td>Numeral -&gt; CN -&gt; NP</td>
<td>five men</td>
</tr>
<tr>
<td>mkNP</td>
<td>PN -&gt; NP</td>
<td>John</td>
</tr>
<tr>
<td>mkNP</td>
<td>Pron -&gt; NP</td>
<td>we</td>
</tr>
<tr>
<td>mkNP</td>
<td>Quant -&gt; Num -&gt; CN -&gt; NP</td>
<td>these (five) man</td>
</tr>
<tr>
<td>mkCN</td>
<td>N -&gt; CN</td>
<td>man</td>
</tr>
<tr>
<td>mkCN</td>
<td>A -&gt; N -&gt; CN</td>
<td>old man</td>
</tr>
<tr>
<td>mkCN</td>
<td>AP -&gt; CN -&gt; CN</td>
<td>very old Chinese man</td>
</tr>
<tr>
<td>mkNum</td>
<td>Numeral -&gt; Num</td>
<td>five</td>
</tr>
<tr>
<td>n100_Numeral</td>
<td>Numeral</td>
<td>one hundred</td>
</tr>
<tr>
<td>p1Num</td>
<td>Num</td>
<td>(plural)</td>
</tr>
</tbody>
</table>
Questions and interrogatives

<table>
<thead>
<tr>
<th>Fun</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>mkQC1</td>
<td>Cl -&gt; QC1</td>
<td>does John walk</td>
</tr>
<tr>
<td>mkQC1</td>
<td>IP -&gt; V -&gt; QC1</td>
<td>who walks</td>
</tr>
<tr>
<td>mkQC1</td>
<td>IP -&gt; V2 -&gt; NP -&gt; QC1</td>
<td>who loves her</td>
</tr>
<tr>
<td>mkQC1</td>
<td>IP -&gt; NP -&gt; V2 -&gt; QC1</td>
<td>whom does she love</td>
</tr>
<tr>
<td>mkQC1</td>
<td>IP -&gt; AP -&gt; QC1</td>
<td>who is old</td>
</tr>
<tr>
<td>mkQC1</td>
<td>IP -&gt; NP -&gt; QC1</td>
<td>who is the boss</td>
</tr>
<tr>
<td>mkQC1</td>
<td>IP -&gt; Adv -&gt; QC1</td>
<td>who is here</td>
</tr>
<tr>
<td>mkIP</td>
<td>CN -&gt; IP</td>
<td>which car</td>
</tr>
<tr>
<td>who_IP</td>
<td>IP</td>
<td>who</td>
</tr>
<tr>
<td>why_IAdv</td>
<td>IAdv</td>
<td>why</td>
</tr>
<tr>
<td>where_IAdv</td>
<td>IAdv</td>
<td>where</td>
</tr>
</tbody>
</table>

Sentence formation, tense, and polarity

<table>
<thead>
<tr>
<th>Fun</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>mkS</td>
<td>Cl -&gt; S</td>
<td>he walks</td>
</tr>
<tr>
<td>mkS</td>
<td>(Tense) -&gt; (Ant) -&gt; (Pol) -&gt; Cl -&gt; S</td>
<td>he wouldn’t have walked</td>
</tr>
<tr>
<td>mkQS</td>
<td>QC1 -&gt; QS</td>
<td>does he walk</td>
</tr>
<tr>
<td>mkQS</td>
<td>(Tense) -&gt; (Ant) -&gt; (Pol) -&gt; QC1 -&gt; QS</td>
<td>wouldn’t he have walked</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>conditionalTense</td>
<td>Tense</td>
<td>(he would walk)</td>
</tr>
<tr>
<td>futureTense</td>
<td>Tense</td>
<td>(he will walk)</td>
</tr>
<tr>
<td>pastTense</td>
<td>Tense</td>
<td>(he walked)</td>
</tr>
<tr>
<td>presentTense</td>
<td>Tense</td>
<td>(he walks) [default]</td>
</tr>
<tr>
<td>anteriorAnt</td>
<td>Ant</td>
<td>(he has walked)</td>
</tr>
<tr>
<td>negativePol</td>
<td>Pol</td>
<td>(he doesn’t walk)</td>
</tr>
</tbody>
</table>
Utterances and imperatives

<table>
<thead>
<tr>
<th>Fun</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>mkUtt</td>
<td>CI -&gt; Utt</td>
<td>he walks</td>
</tr>
<tr>
<td>mkUtt</td>
<td>S -&gt; Utt</td>
<td>he didn’t walk</td>
</tr>
<tr>
<td>mkUtt</td>
<td>QS -&gt; Utt</td>
<td>who didn’t walk</td>
</tr>
<tr>
<td>mkUtt</td>
<td>Imp -&gt; Utt</td>
<td>walk</td>
</tr>
<tr>
<td>mkImp</td>
<td>V -&gt; Imp</td>
<td>walk</td>
</tr>
<tr>
<td>mkImp</td>
<td>V2 -&gt; NP -&gt; Imp</td>
<td>find it</td>
</tr>
<tr>
<td>mkImp</td>
<td>AP -&gt; Imp</td>
<td>be brave</td>
</tr>
</tbody>
</table>

More


Relative clauses: man who owns a donkey

Adverbs: in the house

Subjunction: if a man owns a donkey

Coordination: John and Mary are English or American

4.10 Exercises

1. Compile and make available the resource grammar library, latest version. Compilation is by make in GF/lib/src. Make it available by setting GF_LIB_PATH to GF/lib.

2. Compile and test the grammars face/FaceL (available in course source files).

3. Write a concrete syntax of Face for some other resource language by adding a domain lexicon and a functor instantiation.

4. Add functions to Face and write their concrete syntax for at least some language.

5. Design your own domain grammar and implement it for some languages.

5 Developing a GF Resource Grammar

Contents

Module structure

Statistics
How to start building a new language
How to test a resource grammar
The Assignment

5.1 The principal module structure

Division of labour

Written by the resource grammarian:

- concrete of the row from Structural to Verb
- concrete of Cat and Lexicon
- Paradigms
- abstract and concrete of Extra, Irreg

Already given or derived mechanically:

- all abstract modules except Extra, Irreg
- concrete of Common, Grammar, Lang, All
- Constructors, Syntax, Try
Roles of modules: Library API

Syntax: syntactic combinations and structural words
Paradigms: morphological paradigms
Try: (almost) everything put together
Constructors: syntactic combinations only
Irreg: irregularly inflected words (mostly verbs)

Roles of modules: Top-level grammar

Lang: common syntax and lexicon
All: common grammar plus language-dependent extensions
Grammar: common syntax
Structural: lexicon of structural words
Lexicon: test lexicon of 300 content words
Cat: the common type system
Common: concrete syntax mostly common to languages

Roles of modules: phrase categories

<table>
<thead>
<tr>
<th>module</th>
<th>scope</th>
<th>value categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>adjectives</td>
<td>AP</td>
</tr>
<tr>
<td>Adverb</td>
<td>adverbial phrases</td>
<td>AdN, Adv</td>
</tr>
<tr>
<td>Conjunction</td>
<td>coordination</td>
<td>Adv, AP, NP, RS, S</td>
</tr>
<tr>
<td>Idiom</td>
<td>idiomatic expressions</td>
<td>Cl, QCl, VP, Utt</td>
</tr>
<tr>
<td>Noun</td>
<td>noun phrases and nouns</td>
<td>Card, CN, Det, NP, Num, Ord</td>
</tr>
<tr>
<td>Numeral</td>
<td>cardinals and ordinals</td>
<td>Digit, Numeral</td>
</tr>
<tr>
<td>Phrase</td>
<td>suprasentential phrases</td>
<td>PConj, Phr, Utt, Voc</td>
</tr>
<tr>
<td>Question</td>
<td>questions and interrogatives</td>
<td>IAdv, IComp, IDet, IP, QCl</td>
</tr>
<tr>
<td>Relative</td>
<td>relat. clauses and pronouns</td>
<td>RCl, RP</td>
</tr>
<tr>
<td>Sentence</td>
<td>clauses and sentences</td>
<td>Cl, Imp, QS, RS, S, SC, SSslash</td>
</tr>
<tr>
<td>Text</td>
<td>many-phrase texts</td>
<td>Text</td>
</tr>
<tr>
<td>Verb</td>
<td>verb phrases</td>
<td>Comp, VP, VPSlash</td>
</tr>
</tbody>
</table>
Type discipline and consistency

Producers: each phrase category module is the producer of value categories listed on previous slide.

Consumers: all modules may use any categories as argument types.

Contract: the module Cat defines the type system common for both consumers and producers.

Different grammarians may safely work on different producers.

This works even for mutual dependencies of categories:

Sentence.UseCl : Temp -> Pol -> Cl -> S -- S uses Cl
Sentence.PredVP : VP -> NP -> Cl -- uses VP
Verb.ComplVS : VS -> S -> VP -- uses S

Auxiliary modules

resource modules provided by the library:

- Prelude and Predef: string operations, booleans
- Coordination: generic formation of list conjunctions
- ParamX: commonly used parameter, such as Number = Sg | Pl

resource modules up to the grammarian to write:

- Res: language specific parameter types, morphology, VP formation
- Morpho, Phono,....: possible division of Res to more modules

Dependencies

Most phrase category modules:

concrete VerbGer of Verb = CatGer ** open ResGer, Prelude in ...

Conjunction:

concrete ConjunctionGer of Conjunction = CatGer **
open Coordination, ResGer, Prelude in ...
Lexicon:

```plaintext
concrete LexiconGer of Lexicon = CatGer **
open ParadigmsGer, IrregGer in {
```

**Functional programming style**

The Golden Rule: *Whenever you find yourself programming by copy and paste, write a function instead!*

- Repetition inside one definition: use a `let` expression
- Repetition inside one module: define an `oper` in the same module
- Repetition in many modules: define an `oper` in the `Res` module
- Repetition of an entire module: write a functor

**Functors in the Resource Grammar Library**

Used in families of languages

- Romance: Catalan, French, Italian, Spanish
- Scandinavian: Danish, Norwegian, Swedish

Structure:

- **Common**, a common resource for the family
- **Diff**, a minimal interface extended by interface `Res`
- **Cat** and phrase structure modules are functors over `Res`
- **Idiom, Structural, Lexicon, Paradigms** are ordinary modules

**Example: DiffRomance**

Words and morphology are of course different, in ways we haven’t tried to formalize. In syntax, there are just eight parameters that fundamentally make the difference: Prepositions that fuse with the article (Fre, Spa *de, a*; Ita also *con, da, in, su*).

```plaintext
param Prepos ;
```
Which types of verbs exist, in terms of auxiliaries. (Fre, Ita *avoir*, *être*, and refl; Spa only *haber* and refl).

    param VType ;

Derivatively, if/when the participle agrees to the subject. (Fre *elle* *est* *partie*, Ita *lei è partita*, Spa not)

    oper partAgr : VType -> VP Agr ;

Whether participle agrees to foregoing clitic. (Fre *je* *l’ai* *vue*, Spa *yo* *la he visto*)

    oper vpAgrClit : Agr -> VP Agr ;

Whether a preposition is repeated in conjunction (Fre *la somme de 3 et de 4*, Ita *la somma di 3 e 4*).

    oper conjunctCase : NP Form -> NP Form ;

How infinitives and clitics are placed relative to each other (Fre *la voir*, Ita *vederla*). The Bool is used for indicating if there are any clitics.

    oper clitInf : Bool -> Str -> Str -> Str ;

To render pronominal arguments as clitics and/or ordinary complements. Returns True if there are any clitics.

    oper pronArg : Number -> Person -> CAgr -> CAgr -> Str * Str * Bool ;

To render imperatives (with their clitics etc).

    oper mkImperative : Bool -> Person -> VPC -> {s : Polarity => AAgr => Str} ;

**Pros and cons of functors**

+ intellectual satisfaction: linguistic generalizations
+ code can be shared: of syntax code, 75% in Romance and 85% in Scandinavian
+ bug fixes and maintenance can often be shared as well
+ adding a new language of the same family can be very easy
- difficult to get started with proper abstractions
- new languages may require extensions of interfaces

Workflow: don’t start with a functor, but do one language normally, and refactor it to an interface, functor, and instance.
Suggestions about functors for new languages

Romance: Portuguese probably using functor, Romanian probably independent
Germanic: Dutch maybe by functor from German, Icelandic probably independent
Slavic: Bulgarian and Russian are not functors, maybe one for Western Slavic (Czech, Slovak, Polish) and Southern Slavic (Bulgarian)
Fenno-Ugric: Estonian maybe by functor from Finnish
Indo-Aryan: Hindi and Urdu most certainly via a functor
Semitic: Arabic, Hebrew, Maltese probably independent

5.2 Effort statistics, completed languages

<table>
<thead>
<tr>
<th>language</th>
<th>syntax</th>
<th>morpho</th>
<th>lex</th>
<th>total</th>
<th>months</th>
<th>started</th>
</tr>
</thead>
<tbody>
<tr>
<td>common</td>
<td>413</td>
<td>-</td>
<td>-</td>
<td>413</td>
<td>2</td>
<td>2001</td>
</tr>
<tr>
<td>abstract</td>
<td>729</td>
<td>-</td>
<td>468</td>
<td>1197</td>
<td>24</td>
<td>2001</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>1200</td>
<td>2329</td>
<td>502</td>
<td>4031</td>
<td>3</td>
<td>2008</td>
</tr>
<tr>
<td>English</td>
<td>1025</td>
<td>772</td>
<td>506</td>
<td>2303</td>
<td>6</td>
<td>2001</td>
</tr>
<tr>
<td>Finnish</td>
<td>1471</td>
<td>1490</td>
<td>703</td>
<td>3664</td>
<td>6</td>
<td>2003</td>
</tr>
<tr>
<td>German</td>
<td>1337</td>
<td>604</td>
<td>492</td>
<td>2433</td>
<td>6</td>
<td>2002</td>
</tr>
<tr>
<td>Russian</td>
<td>1492</td>
<td>3668</td>
<td>534</td>
<td>5694</td>
<td>18</td>
<td>2002</td>
</tr>
<tr>
<td>Romance</td>
<td>1346</td>
<td>-</td>
<td>-</td>
<td>1346</td>
<td>10</td>
<td>2003</td>
</tr>
<tr>
<td>Catalan</td>
<td>521</td>
<td>*9000</td>
<td>518</td>
<td>*10039</td>
<td>4</td>
<td>2006</td>
</tr>
<tr>
<td>French</td>
<td>468</td>
<td>1789</td>
<td>514</td>
<td>2771</td>
<td>6</td>
<td>2002</td>
</tr>
<tr>
<td>Italian</td>
<td>423</td>
<td>*7423</td>
<td>500</td>
<td>*8346</td>
<td>3</td>
<td>2003</td>
</tr>
<tr>
<td>Spanish</td>
<td>417</td>
<td>*6549</td>
<td>516</td>
<td>*7482</td>
<td>3</td>
<td>2004</td>
</tr>
<tr>
<td>Scandinavian</td>
<td>1293</td>
<td>-</td>
<td>-</td>
<td>1293</td>
<td>4</td>
<td>2005</td>
</tr>
<tr>
<td>Danish</td>
<td>262</td>
<td>683</td>
<td>486</td>
<td>1431</td>
<td>2</td>
<td>2005</td>
</tr>
<tr>
<td>Norwegian</td>
<td>281</td>
<td>676</td>
<td>488</td>
<td>1445</td>
<td>2</td>
<td>2005</td>
</tr>
<tr>
<td>Swedish</td>
<td>280</td>
<td>717</td>
<td>491</td>
<td>1488</td>
<td>4</td>
<td>2001</td>
</tr>
<tr>
<td>total</td>
<td>12545</td>
<td>*36700</td>
<td>6718</td>
<td>*55963</td>
<td>103</td>
<td>2001</td>
</tr>
</tbody>
</table>

Lines of source code in April 2009, rough estimates of person months. * = generated code.

5.3 How to start building a language, e.g. Marathi

1. Create a directory GF/lib/src/marathi
2. Check out the ISO 639-3 language code: Mar
3. Copy over the files from the closest related language, e.g. hindi
4. Rename files marathi/*Hin.gf to marathi/*Mar.gf
5. Change imports of Hin modules to imports of Mar modules
6. Comment out every line between header { and the final }
7. Now you can import your (empty) grammar: i marathi/LangMar.gf

Suggested order for proceeding with a language

1. ResMar: parameter types needed for nouns
2. CatMar: lincat N
3. ParadigmsMar: some regular noun paradigms
4. LexiconMar: some words that the new paradigms cover
5. (1.-4.) for V, maybe with just present tense
6. ResMar: parameter types needed for Cl, CN, Det, NP, Quant, VP
7. CatMar: lincat Cl, CN, Det, NP, Quant, VP
9. VerbMar: lin UseV
10. SentenceMar: lin PredVP

Character encoding for non-ASCII languages

GF internally: 32-bit unicode
Generated files (.gfo, .pgf): UTF-8
Source files: whatever you want, but use a flag if not isolatin-1.
UTF-8 and cp1251 (Cyrillic) are possible in strings, but not in identifiers. The module must contain

```haskell
flags coding = utf8 ; -- OR coding = cp1251
```

Transliterations are available for many alphabets (see help unicode_table).

Using transliteration

This is what you have to add in GF/src/GF/Text/Transliterations.hs

```haskell
transHebrew :: Transliteration
transHebrew = mkTransliteration allTrans allCodes where
```
allTrans = words $ 
  "A b g d h w z H T y K k l M m N" ++ 
  "n S O P p Z z q r s t -- -- --" ++ 
  "w2 w3 y2 g1 g2"
allCodes = [0x05d0..0x05f4]

Also edit a couple of places in GF/src/GF/Command/Commands.hs.
You can later convert the file to UTF-8 (see help put_string).

Diagnosis methods along the way

Make sure you have a compilable LangMar at all times!
Use the GF command pg -missing to check which functions are missing.
Use the GF command gr -cat=C | l -table to test category C

Regression testing with a treebank

Build and maintain a treebank: a set of trees with their linearizations:
1. Create a file test.trees with just trees, one by line.
2. Linearize each tree to all forms, possibly with English for comparison.

> i english/LangEng.gf
> i marathi/LangMar.gf
> rf -lines -tree -file=trs | l -all -treebank | wf test.treebank

3. Create a gold standard gold.treebank from test.treebank by manually correcting
the Marathi linearizations.
4. Compare with the Unix command diff test.treebank gold.treebank
5. Rerun (2.) and (4.) after every change in concrete syntax; extend the tree set and the
gold standard after every new implemented function.

Sources

A good grammar book

- lots of inflection paradigms
- reasonable chapter on syntax
• traditional terminology for grammatical concepts

A *good* dictionary

• inflection information about words
• verb subcategorization (i.e. case and preposition of complements)

Wikipedia article on the language

Google as ”gold standard”: is it *rucola* or *ruccola*?

Google translation for suggestions (can’t be trusted, though!)

**Compiling the library**

The current development library sources are in *GF/lib/src*.

Use `make` in this directory to compile the libraries.

Use `runghc Make lang api langs=Mar` to compile just the language *Mar*.

### 5.4 Assignment: a good start

1. Build a directory and a set of files for your target language.
2. Implement some categories, morphological paradigms, and syntax rules.
3. Give the `lin` rules of at least 100 entries in *Lexicon*.
4. Send us: your source files and a treebank of 100 trees with linearizations in English and your target language. These linearizations should be correct, and directly generated from your grammar implementation.