

This page is intentionally left (almost) blank.

CONTENTS

1	INTRODUCTION	1
1.1	An alternative user interface	1
1.2	Problem description	1
1.3	A proposed solution	2
1.4	Grammatical Framework	2
1.4.1	A simple example	3
1.5	GF resource grammar library	7
1.5.1	Example usage of RGL in a grammar	7
1.5.2	Generalizing the concrete syntax	8
2	APPLICATION DEVELOPMENT	12
2.1	Requirements specification	12
2.1.1	Generation of mock data	12
2.1.2	Grammar development	12
2.1.3	Suggestions	12
2.1.4	Runtime environment	13
2.2	Grammar development	14
2.2.1	Supported instructions	14
2.2.2	Names	14
2.2.3	More instructions	15
2.2.4	Extending the grammar	16
2.3	Boolean operators	20
2.4	Suggestion engine	22
2.4.1	Populating the Solr index	23
2.5	Generation of mock data	27
3	RESULTS	29
3.1	Translation	29
3.2	Suggestions	34
4	CONCLUSION	38
4.1	Implications for future research	38
4.1.1	Inspiration to GF-developers how to handle suggestions	38
4.1.2	Not use RGL when focusing on one natural language	38
4.2	Known issues	38
4.2.1	Multiple equal abstract syntax trees and linearizations	38
4.2.2	Incorrect English grammar	38
4.2.3	Name suggestions	38
4.3	Limitations	38
4.4	Related work	39
4.5	Future work	39
4.5.1	Speech to text	39

4.5.2	Proper handling of ambiguous instructions	39
4.5.3	Prediction function in PGF in combination with "previous words"	39
4.5.4	Generate Solr-queries to store in the index	39
4.5.5	ad	39
5	DISCUSSION	40
5.1	Contributions	40
5.2	To dos	40
5.3	Negative things	40
5.4	install the environment	40
5.5	Issues	41
5.6	Future work	41
5.7	Related work	41
A	GF RUNTIME SYSTEMS AND LIBRARIES	42
A.1	GF runtime systems	42
A.1.1	Portable Grammar Format	42
A.1.2	GF libraries	42
A.1.3	Building and installing the C to Java wrapper library	43
A.1.4	Using the Java-runtime	44
B	IMPLEMENTATION DETAILS	47
B.1	Mock data generation	47
B.2	Web application	47
	BIBLIOGRAPHY	49

INTRODUCTION

1.1 AN ALTERNATIVE USER INTERFACE

In this thesis we will investigate how we can create a user interface which allows us to execute queries in a query language by expressing instructions in natural language. In other terms, we want to translate from a natural language into a query language. A query language is a computer language which is used to query a database or index. A natural language is a language that humans use to communicate with each other.

1.2 PROBLEM DESCRIPTION

How can one retrieve information from a computer by writing instructions in a natural language? The inspiration for this thesis came from Facebook graph search¹, which is a service that allows users to search for entities by asking the server for information in a natural language.

In this project, we have chosen examine how a similar service can be realized. We have limited the project to handle instructions that can occur naturally in the intranet of a software development company. We assume that there exists an intranet which consists of a database with information about employees, customers and projects. A typical instruction in this environment could be

people who know Java

The answer would be a list of all employees in the database who have some degree of expertise of the programming language Java. However, when using search engines, expert users do not use instructions as the one above. They simply rely purely on keywords[1]. The following instruction is more suited for expert users

people know java

How can we create a user interface that is sufficient for both regular- and expert users? And how can we translate these instructions into machine readable queries?

¹ <https://www.facebook.com/about/graphsearch>

1.3 A PROPOSED SOLUTION

Query languages require precise syntax, we therefore need precise translation from a natural language into a query language. Since we have a limited scope of instructions, we know all instructions that the program shall support and we know how their machine readable representation shall look like. We only need a tool which we can use to make the mapping between natural language into query language. A very flexible tool we can use to accomplish this is a multilingual grammar.

A grammar is a set of structural rules which decide how words can be created and be combined into clauses and phrases. By expressing how words can be combined into an instruction in one language one can also use the same logic to express how the same instruction can be produced in another language. A *multilingual* grammar is a special type of grammar which can translate between two or more languages.

Grammars have been used since the 1950's in the field of computer science [2, p. 4], and they have played a main role in the development of compilers where they are used to translate source code into machine readable instructions.

1.4 GRAMMATICAL FRAMEWORK

Grammatical Framework (GF) is an open source functional programming language for creating grammars for natural languages.[2, p. 1] GF features a strong type system, separate abstract and concrete syntax rules and reusable libraries to facilitate design of grammars. For a reader with a background within compilers, one can see that GF is very much based on the theory of programming languages as they also make use of abstract- and concrete syntaxes[3, pp. 69-70].

Abstract syntax is a tree representation which captures the *meaning* of a sentence, and leaves out anything irrelevant. The concrete syntax represents a natural language. It describes how an abstract syntax tree is represented as a string in the natural language.

With both abstract and concrete syntaxes, GF is able to create a *parser* and a *linearizer* for all given concrete languages. The parser translates strings into abstract syntax and the linearizer translates abstract syntax trees into string representations for a specified concrete syntax. In addition, GF can also derive a *generator* for the abstract syntax. The generator can create all possible abstract syntax trees.

Because GF separates abstract and concrete syntax, one can easily add a new concrete syntax (a natural language) to an existing abstract syntax. This advantage also makes it easy to translate between many languages.

1.4.1 A simple example

The following section presents an example of how GF can be used to create a grammar that can generate and translate the sentence *which people know Java?* into Apache Solr query language and vice versa. Apache Solr is a search platform based on Apache Lucene. We will explain more about Solr query language later in the thesis.

Abstract syntax

To model the meaning of sentences, GF adopts the use of functions and *categories*. A category (cat) in GF is the same as a data type. We start by listing the categories we need. We then define how our data types can take on values. This is achieved by using functions. The functions in an abstract syntax are not implemented, we can therefore only see the function declarations. The purpose for this is to allow the concrete syntaxes to choose how to represent the semantics. Two concrete syntaxes can therefore implement the same semantics differently.

In [Figure 1](#), we define a function `Java : Object` which means that Java is a constant and returns a value of type `Object`. `Know` takes one argument of the type `Object` and returns a value of type `Relation`.

An instruction can be created by obtaining a value of the type `Instruction`. Only `MkInstruction` returns the desired type and it takes two arguments, one of type `Subject` the other of type `Relation`.

A function without arguments is called a constant in lazy functional programming languages.

```

1 abstract Instrucs = {
2   flags startcat = Instruction;
3   cat
4     Instruction ; -- An Instruction
5     Subject ; -- The subject of an instruction
6     Relation ; -- A verb phrase
7     Object ; -- an object
8
9   fun
10    MkInstruction : Subject -> Relation -> Instruction ;
11    People : Subject ;
12    Know : Object -> Relation ;
13    Java : Object ;
14 }
```

Figure 1: Abstract syntax

We can now use this abstract syntax to create an abstract syntax tree as seen in [Figure 2](#)

```
MkInstruction People (Know Java)
```

Figure 2: Abstract syntax tree

Concrete syntax

We are now going to implement the function declarations we just defined in the abstract syntax. This implementation makes it possible to linearize abstract syntax trees into concrete syntax. We will start by defining the concrete syntax for English.

English concrete syntax

Figure 3 shows the implementation of the concrete syntax for English. Categories are linearized by the keyword `lincat`, which literally means the linearization of categories. A category is linearized by assigning a data type to it. Here we assign all categories to be strings. The functions are linearized by using the keyword `lin`. We linearize Java by returning the string "Java", as it is a constant function. Analogously, "people" is returned by `People`. The function `Know` takes one parameter. This parameter is appended on the string "know". Finally, `MkInstruction` takes two arguments, where `subject` is prepended and `relation` is appended on "who". One can easily see how these functions can be used to construct the sentence *people who know Java*.

```

1 concrete InstrucsEng of Instrucs = {
2   lincat
3     Instruction = Str ;
4     Subject = Str ;
5     Relation = Str ;
6     Object = Str ;
7   lin
8     MkInstruction subject relation = subject ++ "who" ++ relation ;
9     People = "people" ;
10    Know object = "know" ++ object ;
11    Java = "Java" ;
12 }
```

Figure 3: English concrete syntax

Solr concrete syntax

The final step in this example is to linearize abstract syntax into Solr concrete syntax. As Figure 4 shows, the categories are strings as in English. The function linearizations are however different. `People` returns "object_type : Person", we assume that the Solr-schema has a field with the name `object_type`

which represents which type a document is. Similarly, we make another assumption about Know. MkInstruction is also implemented differently, here we can see that the result is going to be a query string² by looking at the first part "q=" which is prepended on the subject. We then append "AND" together with relation in order to create a valid Solr query.

```

1 concrete InstrucsSolr of Instrucs = {
2   lincat
3     Instruction = Str ;
4     Subject = Str ;
5     Relation = Str ;
6     Object = Str ;
7
8   lin
9     MkInstruction subject relation = "q=" ++ subject ++ "AND" ++ relation ;
10    People = "object_type : Person" ;
11    Know object = "expertise : " ++ object ;
12    Java = "Java" ;
13 }

```

Figure 4: Solr concrete syntax

Translation

In order to make any translations, we need to use the GF runtime system. The runtime system we will use in this section is the shell application, which allows us to load our GF source code and use parsers, linearizers and generators. In addition to the shell application, there also exists programming libraries for GF in C, Haskell, Java and Python. These libraries can be used to build a translation application which not requires the user to have GF installed.

² http://en.wikipedia.org/wiki/Query_string


```
$ gf InstrucsEng.gf InstrucsSwe.gf

      * * *
    *           *
  *             *
 *
*
*
*      * * * * *
*      *           *
*      * * * * *
 *      *           *
    *      *           *
      * * *

This is GF version 3.5.12-darcs.
No detailed version info available
Built on linux/x86_64 with ghc-7.6, flags: interrupt server
License: see help -license.
Bug reports: http://code.google.com/p/grammatical-framework/issues/list
- compiling Instrucs.gf...   write file Instrucs.gfo
- compiling InstrucsEng.gf... write file InstrucsEng.gfo
- compiling InstrucsSolr.gf... write file InstrucsSolr.gfo
linking ... OK
Languages: InstrucsEng InstrucsSolr
Instrucs>
```

Figure 5: GF shell prompt

A string can be parsed into an abstract syntax tree.

```
Instrucs> parse -lang=InstrucsEng "people who know Java"
MkInstruction People (Know Java)
```

Figure 6: Parse a string

Abstract syntax trees can be linearized into concrete syntaxes, here we linearize one abstract syntax tree into all known concrete syntaxes.

```
Instrucs> linearize MkInstruction People (Know Java)
people who know Java
q= object_type : Person AND expertise : Java
```

Figure 7: Linearize an abstract syntax tree

Finally, a string can be translated from one concrete syntax into another. Here we translate from InstrucsEng into InstrucsSolr. We use a *pipeline*³ to pass the result of the parsing as an argument to the linearizing function.

³ [http://en.wikipedia.org/wiki/Pipeline_\(Unix\)](http://en.wikipedia.org/wiki/Pipeline_(Unix))

Note how we use `p` instead of `parse` and `l` instead of `linearize`. They are just shorthands of their longer representations.

```
Instrucs> p -lang=InstrucsEng "people who know Java" | l -lang=InstrucsSolr
q= object_type : Person AND expertise : Java
```

Figure 8: Translate between concrete syntaxes

1.5 GF RESOURCE GRAMMAR LIBRARY

The previous example is fairly easy to understand, but it is also very small. As a grammar grows to support translation of more sentences, the complexity grows as well. GF has the power to make distinctions between singular and plural forms, genders, tenses and anteniors. However, in order to correctly develop a grammar to translate these sentences one needs to have knowledge of linguistics. It is also very time consuming to implement basic morphologies over and over again. Instead, one can use GF Resource Grammar Library (RGL)[4]. The RGL contains at the time of writing grammars for 29 natural languages. These grammars includes categories and functions which can be used to represent all kinds of different words and sentences. A developer needs therefore only knowledge of her *domain* and do not need to worry about linguistic problems. By domain, we mean specific words which may have special grammatical rules, e.g. fish in English which is the same in singular as in plural form.

1.5.1 Example usage of RGL in a grammar

In this section, we will present how the previous concrete syntax for English can be implemented by using the RGL. We will also show how this grammar can be further generalized into an incomplete concete syntax which can be used by both English and Swedish.

Figure 9 shows the concrete syntax for English by using the RGL. Instead of just concatenating strings, we now use functions to create specific type of words and sentences. The categories are now set to be built in types that exists in the RGL and the functions are now using the RGL in order to create values of the correct types.

The most simple function in this case is `People`, which shall return a noun (N). A noun can be created by using the *operation* `mkN`. We create a noun which has the singular form "person" and plural form "people", we will never use the singular form in this grammar, but it will become handy later in the thesis to use both singular and plural forms.

Java creates a noun similarly, except that it only gives `mkN` one argument. By doing this, GF applies an algorithm in order to find the plural form automati-

An operation in GF is a function which can be called by linearization functions.

cally. This does however only work on regular nouns. We then create a noun phrase (NP) from the noun received. Know returns a relative sentence (RS). A relative sentence can for example be *who know Java*. A relative sentence is constructed by first creating a *verb phrase* (VP) from a verb and an object. This verb phrase is then used together with a constant operation *which_RP* to create a *relative clause*. We can then finally convert the relative clause into a relative sentence. This is achieved by using a self made operation named *mkRS'*, the purpose of this operation is to make the code easier to read and also in the future reuse code.

The only thing that is left is to combine a noun with a relative sentence, e.g. combine *people* with *who know Java*. This is done by using the operation *mkCN* to create a common noun (CN). As CN's do not have any determiners, we have to construct a noun phrase together with the determiner *aPL_Det* in order to only allow translation of plural forms.

```

1 concrete InstrucsEng of Instrucs = open SyntaxEng, ParadigmsEng in {
2   lincat
3     Instruction = NP ;
4     Subject = N ;
5     Relation = RS ;
6     Object = NP ;
7
8   lin
9     MkInstruction subject relation = mkNP aPL_Det (mkCN subject relation) ;
10    People = mkN "person" "people" ;
11    Know object = mkRS' (mkVP (mkV2 (mkV "know") object)) ;
12    Java = mkNP (mkPN "Java") ;
13
14  oper
15    mkRS' : VP -> RS = \vp -> mkRS (mkRCl which_RP vp) ;
16 }

```

Figure 9: English concrete syntax using the RGL

1.5.2 Generalizing the concrete syntax

The English concrete syntax which uses the [RGL](#) is more complicated than the first version which only concatenates strings. In order to motivate why one shall use the RGL, this section describes how the concrete syntax can be generalized into an *incomplete concrete syntax* and then be instantiated by two concrete syntaxes, one for English and one for Swedish.

An incomplete concrete syntax

As we already have designed the concrete syntax for English, we can fairly easy convert it to a generalised version. The incomplete concrete syntax can be

seen in [Figure 10](#). We no longer have any strings defined, as we want to keep the syntax generalised. Constant operations are used in place of strings, and they are imported from the lexicon interface `LexInstrucs`.

```

1 incomplete concrete InstrucsI of Instrucs = open Syntax, LexInstrucs in {
2   lincat
3     Instruction = NP ;
4     Subject = N ;
5     Relation = RS ;
6     Object = NP ;
7
8   lin
9     MkInstruction subject relation = mkNP aPL_Det (mkCN subject relation) ;
10    People = people_N ;
11    Know object = mkRS' (mkVP (know_V2 object)) ;
12    Java = java_NP ;
13
14   oper
15     mkRS' : VP -> RS = \vp -> mkRS (mkRCl which_RP vp) ;
16 }

```

Figure 10: Incomplete concrete syntax

`LexInstrucs` is an *interface*, which means that it only provides declarations. [Figure 11](#) shows that we have one operation declaration for each word we want to use in the concrete syntax. Because we do not implement the operations, it is possible to create multiple instances of the lexicon where each one can implement the lexicon differently.

```

1 interface LexInstrucs = open Syntax in {
2   oper
3     person_N : N ;
4     know_V2 : V2 ;
5     java_NP : NP ;
6 }

```

Figure 11: Lexicon interface

[Figure 12](#) shows how the operations defined in `LexInstrucs` are implemented in `LexInstrucsEng`, we represent the words in the same way as in the old version of the concrete syntax for English.

```

1 instance LexInstrucsEng of LexInstrucs = open SyntaxEng, ParadigmsEng in {
2   oper
3     person_N = mkN "person" "people" ;
4     know_V2 = mkV2 (mkV "know") ;
5     java_NP = mkNP (mkPN "Java");
6 }

```

Figure 12: Lexicon instantiation of English

Figure 13 shows another instance of LexInstrucs, the lexicon for Swedish.

```

1 instance LexInstrucsSwe of LexInstrucs = open SyntaxSwe, ParadigmsSwe in {
2   oper
3     person_N = mkN "person" "personer" ;
4     know_V2 = mkV2 (mkV "kunna" "kan" "kunna" "kunde" "kunnat" "kunna") ;
5     java_NP = mkNP (mkPN "Java");
6 }

```

Figure 13: Lexicon instantiation of Swedish

Why can't we define know_V2 as just mkV2 (mkV "kan")?

We are now ready to instantiate the incomplete concrete syntax. The code below describes how InstrucsI is instantiated as InstrucsEng. Note how we override Syntax with SyntaxEng and LexInstrucs with LexInstrucsEng.

```

1 concrete InstrucsEng of Instrucs = InstrucsI with
2                                     (Syntax = SyntaxEng),
3                                     (LexInstrucs = LexInstrucsEng)
4                                     ** open ParadigmsEng in {}

```

Figure 14: English instantiation of the incomplete concrete syntax

Analogously, we create an instance for Swedish concrete syntax by instantiating InstrucsI and overriding with different files.

```

1 concrete InstrucsSwe of Instrucs = InstrucsI with
2                                     (Syntax = SyntaxSwe),
3                                     (LexInstrucs = LexInstrucsSwe)
4                                     ** open ParadigmsSwe in {}

```

Figure 15: Swedish instantiation of the incomplete concrete syntax

If we load the GF-shell with InstrucsEng.gf and InstrucsSwe.gf we can make the following translation from English to Swedish.

```
1 Instrucs> p -lang=InstrucsEng "people who know Java" | l -lang=InstrucsSwe  
2 personer som kan Java  
3  
4 personer som kan Java  
5  
6 personer som kan Java
```

Figure 16: Swedish instantiation of the incomplete concrete syntax

Why we get three results is unknown at the moment. Whats really interesting is that we can translate Instrucs formulated in English and Swedish into Solr-syntax.

APPLICATION DEVELOPMENT

2.1 REQUIREMENTS SPECIFICATION

2.1.1 *Generation of mock data*

As described in [Section 1.2](#), we want to develop an application which can translate natural language questions that refers to entities in a database or index owned by a software development company. This project has been made with strong collaboration with Findwise, a company with focus on search driven solutions. Findwise has an index with information about employees, projects and customers, however, it is not possible to use their information because it is confidential and cannot be published in a master thesis. A different approach to get hold of relevant data is to generate mock data that is inspired by Findwise's data model. Mock data in this project is simply generated data from files that can be used to simulate a real world example application.

2.1.2 *Grammar development*

The grammar in [Section 1.4.1](#) can only translate the question people who know Java in English and Swedish into Solr query language. The grammar needs to be extended to handle *any* programming language that exists in the mock data, not only Java. In addition, it also needs to support other questions involving not only people, but customers and projects.

2.1.3 *Suggestions*

If a user has no idea of which instructions the application can translate, how can she use the application? This thesis uses a narrow application grammar, which means that it only covers specified sentences. Therefore, if a sentence has one character in the wrong place, GF will not be able to translate anything. This problem can be solved by using a wide coverage grammar, an example of an application that adopts this technique is the GF android app [5, p. 41]. This project, however, does not.

GF has the power to suggest valid words of an incomplete sentence by using incremental parsing [6]. This means that even though a user do not know what to type, the application can suggest valid words to use. If the user chooses to add one of those words, the suggestion engine can show a new list of words that will match the new partial sentence.

However, this method does not support the use of only keywords, since one cannot start a sentence with for example the word Java. It is also inflexible since it does not support the use of words outside the grammar, e.g. all people who know Java.

The parser would not be able to parse the word all

This thesis takes a different approach on a suggestion engine. Instead of suggesting one word at a time, one can suggest a whole sentence based on what the user has typed so far. This is achieved by generating all possible sentences that the application can translate and indexing them in Apache Solr. This makes it very easy to search on matching sentences, we also gain powerful techniques such as approximate string matching.

By using this approach, if a user types a sentence in the application, it will search in the index on instructions related to the string and retrieve the most relevant instructions.

As the suggestion engine uses a search platform, it is possible to type anything and get suggestions, even only keywords like 'people java' will suggest instructions that can be formulated with these two words.

2.1.4 Runtime environment

The chosen programming language for this project is Java. The main reason is because it is Findwise's primary programming language. It is also very well known among many companies in the world. Many professional Java-developers adopts a specific development platform, Java Enterprise Edition (Java EE). This platform provides many libraries that can be scaled to work in an enterprise environment. This project also adopts Java EE.

Handling dependencies

A typical Java EE project make use of several libraries, in computer science terms we say that a project can have other libraries as *dependencies*. It is not unusual that these libraries also have their own dependencies. Larger projects can therefore have a lot of dependencies, so many that it becomes hard to keep track of them. This project make use of an open source tool called Apache Maven to handle dependencies. One simply list all libraries the project shall have access to, then Maven will automatically fetch them and their dependencies. This also makes the application more flexible, as it do not have to include the needed libraries in the application.

Input and output presentation

Besides handling translation and suggestions, the application also needs to handle input and present its results in some way. This application takes input and presents output by using a web gui¹.

To summarize, this application is a web application built in Java EE and uses Apache Maven for library management.

2.1.4.1 *Running the application*

Web applications built in Java are usually has the WAR file format. It is a special JAR-file which includes classes, dependencies and web pages. This project uses an open source web server called Apache Tomcat to host a web application by exporting our application as a WAR-file. Apache Tomcat will make the application available by using HTTP-requests and spawn a new thread for each request.

Details about the runtime environment can be found in [Appendix A](#).

2.2 GRAMMAR DEVELOPMENT

This section continues the work on the grammar introduced in [Section 1.4.1](#).

2.2.1 *Supported instructions*

The example grammar could only translate one instruction. This instruction in English is *people who know Java*. It is easy to extend this grammar to support more programming languages, for example, to support *Python* one can add a function `Python : Object` in the abstract syntax and implement it as `Python = "python"` in the concrete syntaxes. However, this approach makes the grammar inflexible because we need to extend the grammar every time we want to add a new programming language.

2.2.2 *Names*

Defining a new function for each programming language is not a good idea, it forces us to update the grammar every time we want it to support a new programming language. A better solution would be to make one function that could be used by any programming language.

One intuitive approach to solve this problem is to create a function `MkObject : String -> Object`. The implementation for this function would be

¹ http://en.wikipedia.org/wiki/Graphical_user_interface

```

1 -- Abstract syntax
2 MkObject : String -> Object ;
3 -- RGL implementation
4 MkObject str = mkNP (mkPN str.s) ; -- PN = Proper Name
5 -- Solr implementation
6 MkObject str = str.s ;

```

Figure 17: Intuitive approach on names

The GF-code compiles, and the parsing and linearization by using Solr query language works. Unfortunately, this approach does not work with the RGL. GF cannot create a proper name by using an arbitrary string. **Why?**

Fortunately, there exists a built in category which can be used for exactly these situations. We use the category `Symb`, along with the function `MkSymb : String -> Symb` to represent arbitrary strings. We can then use the function `SymbPN` to create a proper name and finally create a noun phrase.

```

1 -- Abstract syntax
2 MkObject : Symb -> Object ;
3 -- RGL implementation
4 MkObject symb = mkNP (SymbPN symb) ; -- PN = Proper Name
5 -- Solr implementation
6 MkObject symb = symb.s ; -- Symb has the type { s : Str }

```

Figure 18: Working approach on names

By using this solution, we can translate the sentence *people who know foo*, where *foo* can be anything.

2.2.3 More instructions

We have only covered translation of one sentence in the application so far. [Figure 19](#) shows all sentences that the end application supports.

English	Solr query language
people who know Java	q= object_type : Person AND KNOWS : Java
people who work in London	q= object_type : Person AND WORKS_IN : London
people who work with Unicef	q= object_type : Person AND WORKS_WITH : Unicef
customers who use Solr	q= object_type : Customer AND USES : Solr
projects who use Solr	q= object_type : Project AND USES : Solr

Figure 19: All sentences supported by the application

Two more cases has been added to instructions regarding *people*. In addition, two new type of instructions has been added, translations about customers

and projects. Note that [Figure 19](#) only shows instances of instructions. As described in the previous section, the grammar uses names that can take on any string for specific parts of the sentences. This means that the words *Java*, *London*, *Unicef* and *Solr* can be exchanged into anything.

2.2.4 Extending the grammar

It is not trivial to extend the grammar to support the instructions described in the previous section.. One has to take into account that it shall not be possible to translate invalid instructions like *projects who work in London*.

2.2.4.1 Abstract syntax

The first step towards a solution to this problem is to modify the abstract syntax. We begin by removing the category *Subject* and replacing it with three new categories: *Internal*, *External* and *Resource*. The function *People* will return a value of the type *Internal* and *Customer* and *Project* will return values of the types *External* and *Resource* respectively.

```

1 -- Instructions.gf
2 cat
3   Internal ;
4   External ;
5   Resource ;
6 fun
7   People   : Internal ;
8   Customer : External ;
9   Project  : Resource ;

```

Figure 20: Abstract syntax with new categories and functions for subjects

In addition to adding new subject categories, three new categories for relations are also introduced: *InternalRelation*, *ExternalRelation* and *ResourceRelation* (*Relation* is removed). The idea is to link subject values to the correct relation types. For instance, we link a value of the type *Internal* with a value of type *InternalRelation*.

All relation functions are changed to return the correct type. For example, *Know* is changed to return a value of the type *InternalRelation*. This means that only *People* can be used together with *Know*, as desired. [Figure 21](#) also shows the new supported relations.

```

1 cat
2   InternalRelation ;
3   ExternalRelation ;
4   ResourceRelation ;
5 fun
6   Know      : Object -> InternalRelation ;
7   UseExt    : Object -> ExternalRelation ;
8   UseRes    : Object -> ResourceRelation ;
9   WorkIn    : Object -> InternalRelation ;
10  WorkWith  : Object -> InternalRelation ;

```

Figure 21: Abstract syntax with new categories and functions for relations

The last thing to modify is how subjects and relations are combined. The function `MkInstruction` is replaced by three new functions: `InstrucInternal`, `InstrucExternal` and `InstrucResource`. However, as we do not need to make any distinction between different type of instructions at this level, all three functions returns a value of the type `Instruction`.

```

1 cat
2   Instruction ;
3 fun
4   InstrucInternal : Internal -> InternalRelation -> Instruction ;
5   InstrucExternal : External -> ExternalRelation -> Instruction ;
6   InstrucResource : Resource -> ResourceRelation -> Instruction ;

```

Figure 22: Abstract syntax with new categories and functions for instructions

2.2.4.2 Concrete syntax for RGL

As the abstract syntax has changed, the concrete syntaxes has to be modified as well. This section explains how the generalised concrete syntax which uses the RGL is implemented.

Figure 23 shows how the categories has been implemented. The new categories are implemented in the same way as the previous.

```

1 lincat
2   Instruction = NP ;
3   Internal, External, Resource = N ;
4   InternalRelation, ExternalRelation, ResourceRelation = VP ;

```

Figure 23: RGL concrete syntax with new category implementations

The new subject functions are implemented in the same way as `People`.

```

1 lin
2   People = person_N ;
3   Customer = customer_N ;
4   Project = project_N ;

```

Figure 24: RGL concrete syntax with new subject implementations

Four new relation functions are added. Line 5-6 in Figure 25 shows how we use the verb `work_V` together with two prepositions, `in_Prep` and `with_Prep` in order correctly linearize into *work in foo* and *work with foo* respectively (*foo* is the value of object). The relation implementations make use of an operation `mkRS'` to reuse code.

```

1 lin
2   Know object = mkRS' (mkVP know_V2 object) ;
3   UseExt object = mkRS' (mkVP use_V2 object) ;
4   UseRes object = mkRS' (mkVP use_V2 object) ;
5   WorkIn object = mkRS' (mkVP (mkV2 work_V in_Prep) object) ;
6   WorkWith object = mkRS' (mkVP (mkV2 work_V with_Prep) object) ;
7
8 oper
9   -- Make a relative sentence
10  mkRS' : VP -> RS = \vp -> mkRS (mkRCl which_RP vp) ;

```

Figure 25: RGL concrete syntax with new relation implementations

Subjects and relations are combined as before, but as this solution has three functions instead of one, a new operation `mkQ` has been defined in order to reuse code.

```

1 lin
2   InstrucInternal internal relation = mkI internal relation ;
3   InstrucExternal external relation = mkI external relation ;
4   InstrucResource resource' relation = mkI resource' relation ;
5
6 oper
7   mkI : N -> RS -> NP = \noun,rs -> mkNP aPl_Det
8                                     (mkCN noun rs) ;

```

Figure 26: RGL concrete syntax with new instruction implementations

2.2.4.3 Concrete syntax for Solr

This section describes how the concrete syntax for Solr is modified to work with the new abstract syntax.

The new categories are all defined as strings.

```

1 lincat
2   Instruction = Str ;
3   Internal, External, Resource = Str ;
4   InternalRelation, ExternalRelation, ResourceRelation = Str ;
5   Object = Str ;

```

Figure 27: Solr concrete syntax with new implementation of categories

Subject types are hard coded into strings. We assume that these strings exists in the Solr index.

```

1 lin
2   People = "Person" ;
3   Customer = "Organization" ;
4   Project = "Project" ;

```

Figure 28: Solr concrete syntax with new subject implementations

We also make an assumption about how the relations are defined in the Solr index.

```

1 lin
2   Know obj = "KNOWS" ++ ":" ++ obj ;
3   UseExt obj = "USES" ++ ":" ++ obj ;
4   UseRes obj = "USES" ++ ":" ++ obj ;
5   WorkWith obj = "WORKS_WITH" ++ ":" ++ obj ;
6   WorkIn obj = "WORKS_IN" ++ ":" ++ obj ;

```

Figure 29: Solr concrete syntax with new relation implementations

As in the concrete syntax for RGL, also an operation is defined and used by the three functions.

```

1 lin
2   InstrucInternal internal relation = select internal relation ;
3   InstrucExternal external relation = select external relation;
4   InstrucResource resource' relation = select resource' relation;
5
6 oper
7   select : Str -> Str -> Str = \subj,relation ->
8       "select?q=:*&wt=json&fq=" ++ "object_type :"
9       ++ subj ++ "AND" ++ relation ;

```

Figure 30: Solr concrete syntax with new instruction implementations

2.3 BOOLEAN OPERATORS

The grammar is now powerful enough to translate a variety of questions. To make it even more powerful, one could use boolean operators in order to combine relations. For example, an instruction that could be useful is *people who know Java and Python*. Another useful instruction is *people who know Java and work in Gothenburg*. This section explains how the grammar can be extended to support these kind of instructions.

In addition to the previous example with the boolean operator *and*, we will also add support for the boolean operator *or*. We begin by adding functionality to support boolean operators to combine values of the type `Object`. As seen in [Figure 31](#), two new functions are defined in the abstract syntax to handle these cases, one for each operator.

```
1 lin
2   And_0 : Object -> Object -> Object ;
3   Or_0  : Object -> Object -> Object ;
```

Figure 31: Abstract syntax for boolean operators and objects

The RGL implementation is shown in [Figure 32](#).

```
1 lin
2   And_0 o1 o2 = mkNP and_Conj o1 o2 ;
3   Or_0  o1 o2 = mkNP or_Conj o1 o2 ;
```

Figure 32: RGL concrete syntax for boolean operators and objects

The Solr implementation is shown in [Figure 33](#). We add AND or OR between the two objects.

```
1 lin
2   And_0 o1 o2 = "(" ++ o1 ++ "AND" ++ o2 ++ ")" ;
3   Or_0  o1 o2 = "(" ++ o1 ++ "OR"  ++ o2 ++ ")" ;
```

Figure 33: Solr concrete syntax for boolean operators and objects

It is now possible to express *people who know Java and Python*. In order to use boolean operators with whole relations like *people who know Java and work in Gothenburg*, the grammar has to be further extended.

We must also take into account that it shall only be possible to combine relationship that are possible to express in the current sentence. Therefore, we need to define the boolean logic three times, as we have three different types of instructions.

```

1 lin
2   InternalAnd : InternalRelation -> InternalRelation -> InternalRelation ;
3   InternalOr : InternalRelation -> InternalRelation -> InternalRelation ;
4
5   ExternalAnd : ExternalRelation -> ExternalRelation -> ExternalRelation ;
6   ExternalOr : ExternalRelation -> ExternalRelation -> ExternalRelation ;
7
8   ResourceAnd : ResourceRelation -> ResourceRelation -> ResourceRelation ;
9   ResourceOr : ResourceRelation -> ResourceRelation -> ResourceRelation ;

```

Figure 34: Abstract syntax for boolean operators and relations

Instead of combining noun phrases as in [Figure 32](#), here we combine relative sentences in the RGL implementation.

```

1 lin
2   InternalAnd rs1 rs2 = mkRS and_Conj rs1 rs2 ;
3   InternalOr rs1 rs2 = mkRS or_Conj rs1 rs2 ;
4
5   ExternalAnd rs1 rs2 = mkRS and_Conj rs1 rs2 ;
6   ExternalAnd rs1 rs2 = mkRS or_Conj rs1 rs2 ;
7
8   ResourceAnd rs1 rs2 = mkRS and_Conj rs1 rs2 ;
9   ResourceOr rs1 rs2 = mkRS or_Conj rs1 rs2 ;

```

Figure 35: RGL concrete syntax for boolean operators and relations

The Solr implementation is fairly straight forward, similarly as with values of the type `Object`, we also add AND or OR between the strings. The only difference is that we do it three times as we have three different subject types.

```

1 lin
2   InternalAnd s1 s2 = "(" ++ s1 ++ "AND" ++ s2 ++ ")";
3   InternalOr s1 s2 = "(" ++ s1 ++ " OR " ++ s2 ++ ")";
4
5   ExternalAnd s1 s2 = "(" ++ s1 ++ "AND" ++ s2 ++ ")";
6   ExternalOr s1 s2 = "(" ++ s1 ++ " OR " ++ s2 ++ ")";
7
8   ResourceAnd s1 s2 = "(" ++ s1 ++ "AND" ++ s2 ++ ")";
9   ResourceOr s1 s2 = "(" ++ s1 ++ " OR " ++ s2 ++ ")";

```

Figure 36: Solr concrete syntax for boolean operators and relations

2.4 SUGGESTION ENGINE

This section continues where [Section 2.1.3](#) left off. The suggestion engine shall generate all possible instructions in all natural languages and store them in an Apache Solr index. It is suitable to use the *generator* which GF creates to generate abstract syntax trees and linearize them into the specified concrete languages (English and Swedish in this case).

The generator can be accessed through the GF-shell. [Figure 37](#) shows how the function `generate_trees` is executed to generate all trees with the *depth* 4. By depth, we mean the maximum number edges between a leaf and the root element.

```
Instrucs> generate_trees
InstrucExternal Customer (UseExt (MkObject (MkSymb "Foo")))
InstrucInternal People (Know (MkObject (MkSymb "Foo")))
InstrucInternal People (WorkIn (MkObject (MkSymb "Foo")))
InstrucInternal People (WorkWith (MkObject (MkSymb "Foo")))
InstrucResource Project (UseRes (MkObject (MkSymb "Foo")))
```

Figure 37: `generate_trees` is used to create all abstract syntax trees with max depth 4

[Figure 37](#) shows 5 trees, but there exists many more trees. The reason GF only generates 5 trees is because of the default depth setting is 4. If we increase the depth we will obtain more trees, as it then will include trees containing boolean operators. By increasing the max depth to 5, GF will generate 36 trees. With depth 6, GF will generate 1653 trees.

Actually, there exists infinitely many trees, as we have recursive functions in the abstract syntax.

It is often good to visualize trees to understand them better. [Figure 51](#) shows two abstract syntax trees, the first one with depth 4 and the second with depth 5. One can easily see that the former has maximum 4 edges between root and leaf, and the latter has maximum 5 edges between root and leaf.

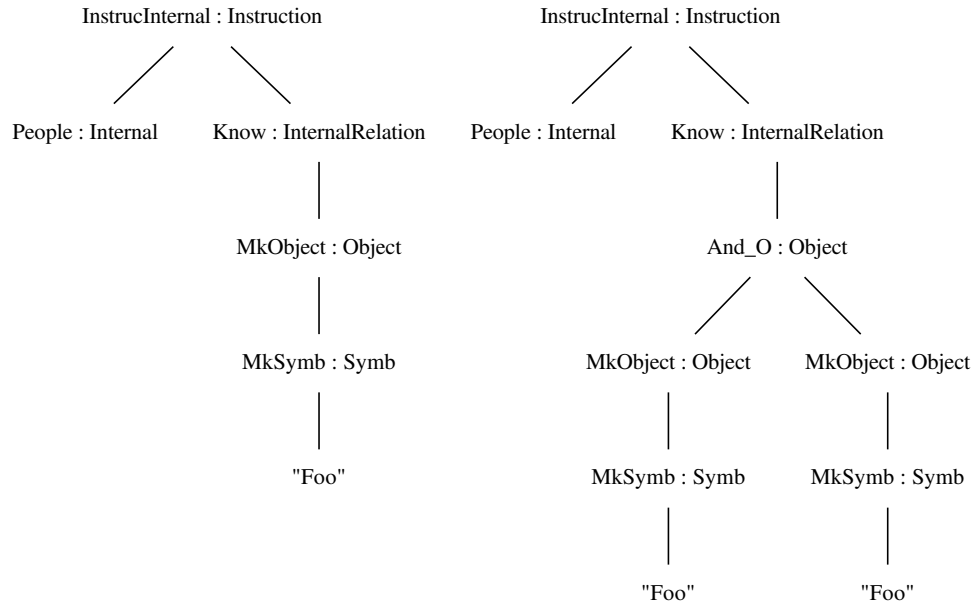


Figure 38: Visualization of abstract syntax trees with depth 4 and 5

2.4.1 Populating the Solr index

It is now time to linearize the generated trees and store them in a Solr index that is dedicated to store linearizations. By doing so, we will be able to search on instructions by using words that exists in the instructions. What we cannot do is to search on names, because the names does not occur in the instructions, instead they only contain a placeholder for a name ("Foo"). It will therefore only suggest instructions like *people who know Foo and Foo* which is a useless suggestion.

We want to be able to suggest relevant names, if the user database contains a person which knows Java, then we also want to suggest instructions based on that name. This requirement forces us to change the application once more.

A naïve solution would be to fetch all distinct names from the database and create all possible instructions that a user shall be able to express with these names. For an instruction like *people who know Foo or Foo and works in Foo* and if the database contains 10 programming languages and 10 cities we would then have to generate $2 * 10 * 10 = 200$ instructions. This is clearly not suitable, as GF generates 1653 trees (with depth 6).

A better approach on the problem is to store all distinct names in a separate index. When a user begins to type an instruction, the application checks each word the user has typed. If a word exists in the name-index, then treat it as a name and replace it with Foo, then query the linearizations index. Re-

trieve the results, and change back *Foo* to the original name. However, as the application do not make any distinction between different type of names, we could end up with suggestions like *people who work in Python*, because *Python* was replaced by *Foo*. Luckily, this problem can be resolved by introducing a distinction between different types of names in the grammar.

This application uses four different kind of names. Programming languages are used together with the *Know* relation. Organizations are used with the *Work_With* relation. Locations are used with the *Work_In* relation. Lastly, modules are used with the *Use* relation. We extend the grammar to support these new name types.

Figure 39 shows the new abstract syntax. Line 3 defines new name types. Line 7-10 defines how the names are instantiated. Lines 13-23 defines how names can be combined by using boolean operators. Note how we use the type *Skill* for programming languages.

```

1 cat
2   -- Names
3   Skill ; Organization ; Location ; Module ;
4
5 fun
6   -- Create unknown names
7   MkSkill : Symb -> Skill ;
8   MkOrganization : Symb -> Organization ;
9   MkModule: Symb -> Module ;
10  MkLocation : Symb -> Location ;
11
12  -- Boolean operators for Organizations
13  And_S : Skill -> Skill -> Skill ;
14  Or_S : Skill -> Skill -> Skill ;
15
16  And_O : Organization -> Organization -> Organization ;
17  Or_O : Organization -> Organization -> Organization ;
18
19  And_L : Location -> Location -> Location ;
20  Or_L : Location -> Location -> Location ;
21
22  And_M : Module -> Module -> Module ;
23  Or_M : Module -> Module -> Module ;

```

Figure 39: Abstract syntax with new name types

The concrete syntax for these new functions are implemented in the same way as the ones we removed (category *Object* and functions *MkObject*, *And_O* and *Or_O*). We have omitted the new concrete syntaxes from the thesis as they do not contribute to anything new.

As the grammar has changed, also the abstract syntax trees has changed. Figure 40 shows how GF now generate all trees with depth 4.

```
Instrucs> generate_trees
InstrucExternal Customer (UseExt (MkModule (MkSymb "Foo")))
InstrucInternal People (Know (MkSkill (MkSymb "Foo")))
InstrucInternal People (WorkIn (MkLocation (MkSymb "Foo")))
InstrucInternal People (WorkWith (MkOrganization (MkSymb "Foo")))
InstrucResource Project (UseRes (MkModule (MkSymb "Foo")))
```

Figure 40: Abstract syntax trees with name types

As the index stores the linearizations of each tree it is important to post process each linearization in order to preserve the type information. Currently, the type of each name would be lost when linearized, for example `MkSkill (MkSymb "Foo")` is linearized into just `"Foo"` if using the english concrete syntax.

To preserve type information, we replace each `"Foo"` with the current name type + an index. [Figure 41](#) shows an example.

```
InstrucInternal People (InternalAnd (Know
  (And_S (MkSkill (MkSymb "Skill0")) (MkSkill (MkSymb "Skill1"))))
  (WorkIn (MkLocation (MkSymb "Location0"))))
```

Figure 41: An abstract syntax tree with name types and changed names

The previous example linearizes into *people who know Skill0 and Skill1 and who work in Location0* by using the concrete syntax for English. We generate all trees again, linearize them into all natural languages and store them in the linearizations index.

In addition, as we store each name in a separate index, we associate each name with its type. By doing so, we can easily find out if a word is a name and then find out what type that name has.

The grammar and the Solr indices are now complete. The application can now suggest relevant sentences even though the user only has typed names (provided that the names exists in the name index). [Figure 42](#) shows the pseudocode for the suggestion algorithm with informative comments. This algorithm is executed by the Java-application when a user has typed a partial question.

```

1 suggestions(sentence) {
2   // sentence = "persons who knew java or python and work in London"
3   // names[] = {Java, Python, London}
4   names[] = extractNames(sentence);
5
6   // "persons who knew java or python and work in London" ==>
7   // "persons who knew Skill0 or Skill1 and work in Location0"
8   sentence = replaceNamesWithTypes(sentence, names);
9
10  // suggestions =
11  //   { "people who know Skill0 or Skill1 and who work in Location0", ... }
12  suggestions[] = findSentences(sentence);
13
14  foreach suggestion in suggestions {
15    // "people who know Skill0 or Skill1 and who work in Location0" ==>
16    // "people who know Java or Python and who work in London"
17    suggestion = restoreNames(names, suggestion);
18  }
19  return suggestions;
20 }

```

Figure 42: Pseudocode for the suggestion algorithm

2.4.1.1 Supporting invalid sentences

This section aims to add better accuracy of the suggestion engine by linearizing different variations of a sentence. While the linearization function in GF returns a string representation of a tree, there also exists a parameter `-list` which allows GF to return all semantically equivalent ways of expressing a sentence.

```

Languages: InstrucsEng
Instrucs> p "people who know Java" | l -list
people who know Java, people's who know Java

```

Figure 43: Linearization to InstrucsEng using `-list`

Here we actually get four results, three duplicates! Fix, but how?

Figure 45 shows that *people who know Java* actually can be expressed in two ways. We add both linearizations as a list to the same object in the linearizations index, and we list the one we think is the most suitable in the beginning of the list. This makes it possible to search for everything the grammar can parse, but get a more suitable suggestion. For instance, if type *people's who know Java* then the application will match the lesser suitable sentence in the index, but return *people who know Java* as a suggestion.

It is possible to add more linearizations than those listed in Figure 45. GF

adopts the pipe symbol to express that two elements are semantically equivalent. The following sections describes how the pipe symbol can be used to further extend the grammar.

We extend the grammar to treat a subject in singular form the same as a subject in plural form. I.e. the sentences *a person which knows Java* is the same as *people who know Java*. As we treat them equivalent, they will have the same abstract syntax tree, and this tree will linearize to only one Solr query, hence if we translate *a person who knows Java* into Solr, we will obtain a query that retrieves multiple results. The implementation of this extension is achieved by modifying the operation mkI.

```

1 oper
2   mkI : N -> RS -> NP = \noun_N,rs_RS -> mkNP (aPl_Det | aSg_Det)
3                                     (mkCN noun_N rs_RS) ;

```

Figure 44: Treating singular form equivalent to plural form

The second extension is to treat the *progressive form* of a verb as the same as the regular form. A progressive form of a verb indicates that something is happening right now or was happening or will be happening. The grammar in this project only supports present tense, so it will only cover progressive verbs which expresses that something is happening right now. This is achieved by using the variance function again.

```

1 oper
2   mkRS' : VP -> RS = \vp -> mkRS (mkRCl which_RP (vp | progressiveVP vp)) ;

```

Figure 45: Treating progressive verb equivalent normal form

However, not all verbs can be used in progressive form. An example of such verb is *know*. An example of *know* in progressive form is *people who are knowing Java* which clearly is incorrect. Although it might look ugly, it is not an issue for our application as the suggestion engine only suggest the most suitable suggestion. The extension will contribute to make the suggestion engine richer as it allows users to use incorrect instructions.

Also add that we let `people_N = mkN "person" ("people" | "persons")`

2.5 GENERATION OF MOCK DATA

This section describes how the data used by the application is generated. In order to generate data, we must know what to generate. From [Section 2.2.3](#) we know the Solr queries our index must support. We have three different types of objects: *Customer*, *Person* and *Project*. Each of these types of objects has their

own data. A customer shall have one value, a list of technologies it uses. A person shall have three values, a list of programming languages it knows, a list of cities it works in and a list of organizations it has been working with. A project shall have one value, the same as a customer, a list of technologies it uses. In addition to these requirements by the grammar, customers and persons also have names. Figure 46 shows an example of data that could exist in the Solr index.

```

1 { // Customer
2   "name" : "Amnesty"
3   "object_type" : "Customer",
4   "USES" : ["Solr", "Tomcat", "GF"]
5 },
6 { // Person
7   "name" : "Jane Doe"
8   "object_type" : "Person",
9   "KNOWS" : ["Java", "Python", "C"],
10  "WORKS_IN" : ["London", "Gothenburg"],
11  "WORKS_WITH" : ["Amnesty", "Unicef"]
12 },
13 { // Project
14   "object_type" : "Project",
15   "USES" : ["Android", "GF"]
16 }

```

Figure 46: Example data in JSON

Mock data is generated by combining data from text-files. We have two text-files concerning personal names (`first_names.txt` and `last_names.txt`). One text-file with organization names (`charity_organizations.txt`). One text file about programming languages (`programming_languages.txt`). Lastly, one text file containing cities (`cities.txt`).

An object is generated by taking one value from each text file that the object needs. For instance, to create an object of the type `Person`, we fetch values from `first_names.txt`, `last_names.txt`, `programming_languages.txt`, `cities.txt` and `charity_organizations.txt`. The object is then exported to Solr to be stored in the index.

RESULTS

This chapter presents results from the end application.

3.1 TRANSLATION

We begin by demonstrating how a few sentences are parsed into an abstract syntax which is linearized into all possible concrete syntaxes. The input is shown in the first box in each figure and the result in second (larger) box. The result of a translation is JSON-data, given by the application.

The first five figures shows the most simplest translations. They include all subjects and all types of names. Each one contains a subject, one verb and one name of a certain type.

people who know Java

```

1 [
2   {
3     'ast': 'InstrucInternal People (Know_R (MkSkill (MkSymb "Java")))',
4     'linearizations': [
5       {
6         'query': 'people who know Java',
7         'language': 'InstrucsEng'
8       },
9       {
10        'query': 'select?q=*&wt=json&fq= object_type : Person AND KNOWS : ( Java )',
11        'language': 'InstrucsSolr'
12      },
13      {
14        'query': 'personer som kan Java',
15        'language': 'InstrucsSwe'
16      }
17    ]
18  }
19 ]

```

Figure 47: Translation including people and a name of type Skill

people who work in London

```

1 [
2 {
3   'ast': 'InstrucInternal People (WorkIn_R (MkLocation (MkSymb "London")))',
4   'linearizations': [
5     {
6       'query': 'people who work in London',
7       'language': 'InstrucsEng'
8     },
9     {
10      'query': 'select?q=*&wt=json&fq= object_type : Person AND WORKS_IN : ( London )',
11      'language': 'InstrucsSolr'
12    },
13    {
14      'query': 'personer som arbetar i London',
15      'language': 'InstrucsSwe'
16    }
17  ]
18 }
19 ]

```

Figure 48: Translation including people and a name of type Location

people who work with Amnesty

```

1 [
2 {
3   'ast': 'InstrucInternal People (WorkWith_R (MkOrganization (MkSymb "Amnesty")))',
4   'linearizations': [
5     {
6       'query': 'people who work with Amnesty',
7       'language': 'InstrucsEng'
8     },
9     {
10      'query': 'select?q=*&wt=json&fq= object_type : Person AND WORKS_WITH : ( Amnesty )',
11      'language': 'InstrucsSolr'
12    },
13    {
14      'query': 'personer som arbetar med Amnesty',
15      'language': 'InstrucsSwe'
16    }
17  ]
18 }
19 ]

```

Figure 49: Translation including people and a name of type Organization

customers who use Solr

```

1 [
2 {
3   'ast': 'InstrucExternal Customer (UseExt_R (MkModule (MkSymb "Solr")))',
4   'linearizations': [
5     {
6       'query': 'customers who use Solr',
7       'language': 'InstrucsEng'
8     },
9     {
10      'query': 'select?q=*&wt=json&fq= object_type : Organization AND USES : ( Solr )',
11      'language': 'InstrucsSolr'
12    },
13    {
14      'query': 'kunder som använder Solr',
15      'language': 'InstrucsSwe'
16    }
17  ]
18 }
19 ]

```

Figure 50: Translation including customer and a name of type Module

projects who use Solr

```

1 [
2 {
3   'ast': 'InstrucResource Project (UseRes_R (MkModule (MkSymb "Solr")))',
4   'linearizations': [
5     {
6       'query': 'projects who use Solr',
7       'language': 'InstrucsEng'
8     },
9     {
10      'query': 'select?q=*&wt=json&fq= object_type : Project AND USES : ( Solr )',
11      'language': 'InstrucsSolr'
12    },
13    {
14      'query': 'projekt som använder Solr',
15      'language': 'InstrucsSwe'
16    }
17  ]
18 }
19 ]

```

Figure 51: Translation including project and a name of type Module

Figure 52 shows how the applications handles the boolean operator *and*. The application handles the case for *or* similarly.

people who know Java and Python

```

1 [
2 {
3   'ast': 'InstrucInternal People (Know_R (And_S (MkSkill (MkSymb "Java"))
4                                     (MkSkill (MkSymb "Python"))))',
5   'linearizations': [
6     {
7       'query': 'people who know Java and Python',
8       'language': 'InstrucsEng'
9     },
10    {
11      'query': 'select?q=*&wt=json&fq= object_type : Person AND
12                                     KNOWS : ( ( Java ) AND ( Python ) )',
13      'language': 'InstrucsSolr'
14    },
15    {
16      'query': 'personer som kan Java och Python',
17      'language': 'InstrucsSwe'
18    }
19  ]
20 }
21 ]

```

Figure 52: Translation including people and two names of the type Skill

Figure 53 shows how the application handles an ambiguous instruction. Two abstract syntax trees are seen in the result, as there are two ways of interpreting the instruction. The different interpretations can be modelled as follows: *people who know (Java and Python) or Haskell* or *people who know Java and (Python or Haskell)*.

```

people who know Java and Python or Haskell

1 [
2 {
3   'ast': 'InstrucInternal People (Know_R (Or_S (And_S (MkSkill (MkSymb "Java"))
4           (MkSkill (MkSymb "Python")))) (MkSkill (MkSymb "Haskell"))))',
5   'linearizations': [
6     {
7       'query': 'people who know Java and Python or Haskell',
8       'language': 'InstrucsEng'
9     },
10    {
11      'query': 'select?q=*&wt=json&fq= object_type : Person AND
12              KNOWS : ( ( Java ) AND ( Python ) ) OR ( Haskell ) )',
13      'language': 'InstrucsSolr'
14    },
15    {
16      'query': 'personer som kan Java och Python eller Haskell',
17      'language': 'InstrucsSwe'
18    }
19  ]
20 },
21 {
22   'ast': 'InstrucInternal People (Know_R (And_S (MkSkill (MkSymb "Java"))
23           (Or_S (MkSkill (MkSymb "Python")) (MkSkill (MkSymb "Haskell")))))',
24   'linearizations': [
25     {
26       'query': 'people who know Java and Python or Haskell',
27       'language': 'InstrucsEng'
28     },
29     {
30       'query': 'select?q=*&wt=json&fq= object_type : Person AND
31              KNOWS : ( ( Java ) AND ( ( Python ) OR ( Haskell ) ) )',
32       'language': 'InstrucsSolr'
33     },
34     {
35       'query': 'personer som kan Java och Python eller Haskell',
36       'language': 'InstrucsSwe'
37     }
38  ]
39 }
40 ]

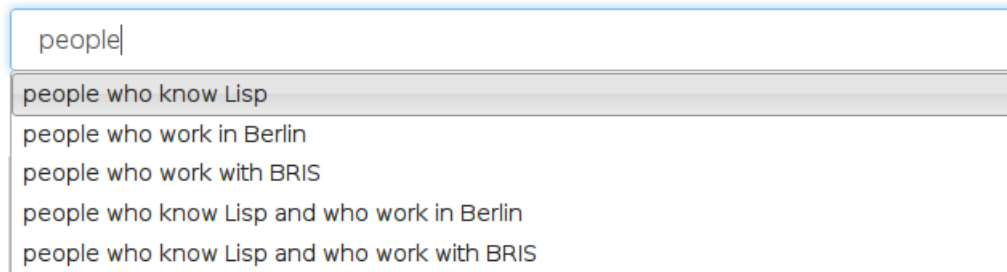
```

Figure 53: Translation of an ambiguous instruction involving people and three names of the type Skill

3.2 SUGGESTIONS

This section shows how the application reacts to input before the user has chosen to translate the instruction. Each figure shows an image of how the application suggest valid sentences from a partial instruction.

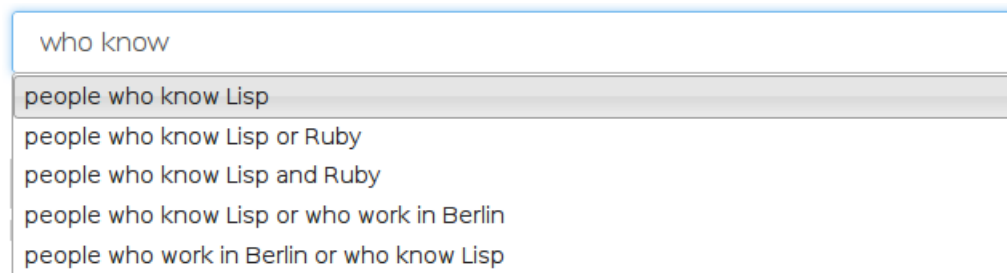
Figure 54 shows suggestions based on the first word in a valid sentence.



A screenshot of a web application interface. At the top is a text input field containing the word 'people'. Below the input field is a dropdown menu with a light gray background. The first item in the dropdown is 'people who know Lisp', which is highlighted with a darker gray background. Below it are five more items: 'people who work in Berlin', 'people who work with BRIS', 'people who know Lisp and who work in Berlin', and 'people who know Lisp and who work with BRIS'.

Figure 54: Suggestions based on the word people

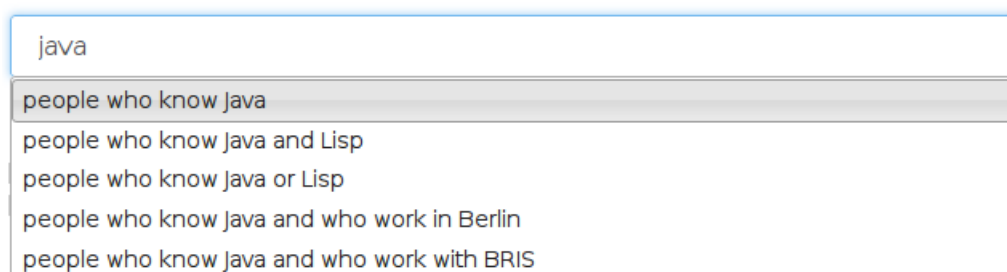
Figure 55 shows that it is possible to use a combination of words to retrieve suggestions. It also demonstrates that the words do not have to be the first words in an instruction.



A screenshot of a web application interface. At the top is a text input field containing the words 'who know'. Below the input field is a dropdown menu with a light gray background. The first item in the dropdown is 'people who know Lisp', which is highlighted with a darker gray background. Below it are five more items: 'people who know Lisp or Ruby', 'people who know Lisp and Ruby', 'people who know Lisp or who work in Berlin', and 'people who work in Berlin or who know Lisp'.

Figure 55: Suggestions based on the words who know

The next four figures shows how it is possible to retrieve suggestions based on only names.



A screenshot of a web application interface. At the top is a text input field containing the word 'java'. Below the input field is a dropdown menu with a light gray background. The first item in the dropdown is 'people who know java', which is highlighted with a darker gray background. Below it are five more items: 'people who know java and Lisp', 'people who know java or Lisp', 'people who know java and who work in Berlin', and 'people who know java and who work with BRIS'.

Figure 56: Suggestions based on a name of type Skill

london
people who work in London
people who work in London and Berlin
people who work in London or Berlin
people who work in London and who know Lisp
people who work in London and who work with BRIS

Figure 57: Suggestions based on a name of type Location

amnesty
people who work with Amnesty
people who work with Amnesty or BRIS
people who work with Amnesty and BRIS
people who know Lisp or who work with Amnesty
people who work with Amnesty or who know Lisp

Figure 58: Suggestions based on a name of type Organization

solr
customers who use Solr
projects who use Solr
customers who use Solr and Neo4j
customers who use Solr or Neo4j
projects who use Solr and Neo4j

Figure 59: Suggestions based on a name of type Module

Figure 60 shows suggestions based on two names of the same type.

java python
people who know Java or Python
people who know Java and Python
people who know Java and Python or Lisp
people who know Java or Python and Lisp
people who know Java or Python or Lisp

Figure 60: Suggestions based on two names of type Skill

Figure 61 shows that names do not have to be of the same type to suggest relevant instructions.

A screenshot of a search interface. The input field contains the text "Haskell London". Below the input field, a list of suggestions is displayed. The first suggestion, "people who know Haskell or who work in London", is highlighted with a grey background. The other suggestions are "people who work in London or who know Haskell", "people who work in London and who know Haskell", "people who know Haskell and who work in London", and "people who know Haskell".

Figure 61: Suggestions based on two names of different types

Figure 62 shows how we can use the word persons in order to get suggestions. The word persons does not exist in any suggestion, but as it is a synonym to people, the application suggest relevant instructions.

A screenshot of a search interface. The input field contains the text "persons". Below the input field, a list of suggestions is displayed. The first suggestion, "people who know Lisp", is highlighted with a grey background. The other suggestions are "people who work in Berlin", "people who work with BRIS", "people who know Lisp and who work in Berlin", and "people who know Lisp and who work with BRIS".

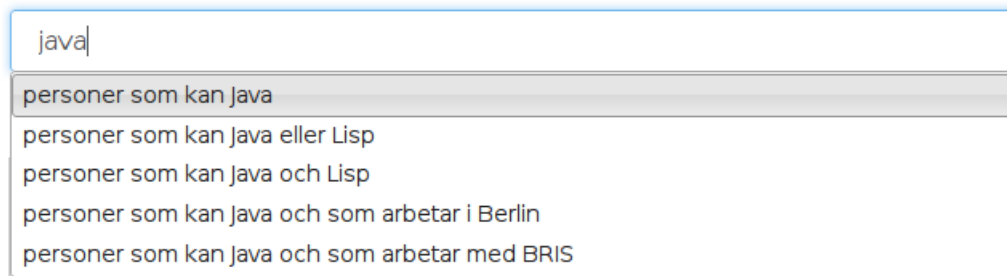
Figure 62: Suggestions combining

Figure 63 shows how the application shows relevant suggestion based on a misspelled word.

A screenshot of a search interface. The input field contains the text "Tjockholm". Below the input field, a list of suggestions is displayed. The first suggestion, "people who work in Stockholm", is highlighted with a grey background. The other suggestions are "people who work in Stockholm and Berlin", "people who work in Stockholm or Berlin", "people who work in Stockholm and who know Lisp", and "people who work in Stockholm and who work with BRIS".

Figure 63: Suggestions based on a misspelled word

Lastly, Figure 64 demonstrates that the application also can translate give suggestions in Swedish. The application can translate valid Swedish sentences as long as the user has chosen Swedish as input language in the application.



A screenshot of a search interface. At the top is a text input field containing the word "java". Below the input field is a list of five suggestions, each on a new line. The first suggestion, "personer som kan java", is highlighted with a light gray background. The other suggestions are "personer som kan java eller Lisp", "personer som kan java och Lisp", "personer som kan java och som arbetar i Berlin", and "personer som kan java och som arbetar med BRIS".

java
personer som kan java
personer som kan java eller Lisp
personer som kan java och Lisp
personer som kan java och som arbetar i Berlin
personer som kan java och som arbetar med BRIS

Figure 64: Suggestions in Swedish

CONCLUSION

4.1 IMPLICATIONS FOR FUTURE RESEARCH

4.1.1 *Inspiration to GF-developers how to handle suggestions*

A hint to GF-developers that it is crucial for an application to handle incorrect sentences properly.

4.1.2 *Not use RGL when focusing on one natural language*

When only using one natural language and have relatively little valid sentences, I feel that RGL is unnessecary. It makes it harder to express valid sentences, but also express invalid sentences.

4.2 KNOWN ISSUES

4.2.1 *Multiple equal abstract syntax trees and linearizations*

4.2.2 *Incorrect English grammar*

project who uses Solr.

4.2.3 *Name suggestions*

Cannot suggest names based on two words.

4.3 LIMITATIONS

Can be applied in any context where one want to instruct a computer by using a natural language.

4.4 RELATED WORK

4.5 FUTURE WORK

4.5.1 *Speech to text*

4.5.2 *Proper handling of ambiguous instructions*

4.5.3 *Prediction function in PGF in combination with "previous words"*

This will also make it possible to suggest infinitely many suggestions in recursive functions as we can suggest based on a sentence and continue to suggest on the resulting sentence.

4.5.4 *Generate Solr-queries to store in the index*

Map each English or Swedish instruction to the corresponding Solr instruction, then no need of translating with GF anymore.

DISCUSSION

5.1 CONTRIBUTIONS

My work has improved PGF, at least a bit

Issue with the suggestion algorithm, cannot take more than one word..

5.2 TO DO'S

How we add a special behavior of names with spaces in it. Write about how we get the most related results by counting names Write that we use fuzzy search Write how to install GF Write about how we remove sentences like people who know java and who know java Fördelar med att använda Solr framför ett gigantiskt lexikon (we get fuzzy search and keyword only suggestions, maybe faster too and more flexible)

Add "more sophisticated suggestion engine with graph database. Eg. if one type "people" then it shall recognize it as a node and check all distinct edges from that node as relationships to other nodes and suggest to autocomplete all those nodes.

5.3 NEGATIVE THINGS

Generates three trees for each instruction (p "people who know Java or Python and who work in London" Can just generate all instructions in all languages and add them to solr. No need for translating then. Generates the instruction "people who know Skillo and who work in Locationo" but also "people who work in Locationo and who know Skillo" they're equivalent. Write about how the suggestions are not optimal.

5.4 INSTALL THE ENVIRONMENT

Prereqs: Tomcat8, Maven. Install GF through cabal (needed to generate instructions). Write howto install generate instructions and mock data. Write how to add the gf.pgf.jar as a maven dependency

```
mvn install:install-file -Dfile=org.grammaticalframework.pgf.jar -DgroupId=org.grammaticalframe
-DartifactId=pgf -Dversion=1.0 -Dpackaging=jar
```

5.5 ISSUES

Bug causes gf to parse strings into multiple ASTS and multiple linearizations, and the application post process alot. Must have capital letter on each name projects who use Java is an incorrect sentence. It shall be projects which/that use Java.

5.6 FUTURE WORK

Speech to text Handle ambiguous instructions Make application know which language that is used Use prediction function in pgf to give suggestions of sentences (<https://groups.google.com/forum/#!topic/gf-dev/uoQDQ0DSzoo>) I thought about one approach: As we still want to have keyword only suggestions we can implement a "backward completion function" by mapping each word with all other words that can be used in front of it. When we have the first word we can use the completion function in PGF. This solution would make it possible to suggest a sentence based on a sentence. Instead of storing the names in solr, store them as a lexicon in GF (maybe) Not use RGL, instead use string concatenation, since we only support a few sentences

Support more variations of questions

5.7 RELATED WORK

adasd

A.1 GF RUNTIME SYSTEMS

While the GF-shell is a powerful tool, it is not very convenient to interact with when programming an application. Luckily, the creators of GF has thought about this and built embeddable runtime systems for a few programming languages [7, p. 3]. These runtime systems makes it possible to interact with a grammar directly through language specific data types. We have chosen to work with the Java-runtime system in this project.

A.1.1 *Portable Grammar Format*

The GF-shell interacts with grammars by interpreting the GF programming language. This allows us to write our grammars in an simple and convenient syntax. Interpreting the GF programming language directly is however a heavy operation[7][p. 13], especially with larger grammars. This is where the Portable Grammar Format (PGF)[7][p. 14] comes in. PGF is a custom made machine language which is dynamically created by compiling a grammar with GF into a PGF-file. The runtime systems works exclusively with PGF-files.

A.1.2 *GF libraries*

In order to use the Java-runtime, we first need to generate a few libraries which is used by the runtime system. The Java-runtime system depends on the C-runtime system and a special wrapper between the C- and the Java-runtime. The libraries are platform dependent and at the time of writing, no pre-generated libraries exists. We therefore need to generate the libraries by ourselves. We will start by generating and installing the C-libraries. We will then go through how we can generate the wrapper library.

A.1.2.1 *Building and installing the C-runtime*

Start by fetching the needed dependencies

```
sudo apt-get install gcc autoconf libtool
```

Download the latest source code of GF from GitHub.

```
git clone https://github.com/GrammaticalFramework/GF.git
```

It is also possible to download the project as an archive by visiting the repository url.

You will receive a directory GF/. Change the current working directory to the C-runtime folder.

```
cd GF/src/runtime/c
```

Generate a configuration file

```
autoreconf -i
```

Check that all dependencies are met

```
./configure
```

If there exists a dependency that is not fulfilled, try to install an appropriate package using your package-manager.

Build the program

```
make
```

Install the libraries you just built

```
sudo make install
```

Make sure the installed libraries are installed into `/usr/local/lib`. It is crucial that they exists in that directory in order for the program to work.

A.1.3 *Building and installing the C to Java wrapper library*

Start by installing the needed dependency

```
sudo apt-get install g++
```

The wrapper is built by using a script which is executed in Eclipse. This step assumes that you have Eclipse installed with the CDT-plugin. If you don't have Eclipse, you can download it with your package manager, just do not forget to install the CDT-plugin.

Start Eclipse and choose `File > Import...` in the menu. Choose `Import Existing Projects into Workspace` and click on the `Next` button. Select `Browse...` and navigate to the location where you downloaded GF from GitHub and press enter. Uncheck everything except `jpgf` and click on `Finish`. You have now imported the project which can build the Java-runtime system.

Unfortunately, the build-configuration for the `jpgf`-project is not complete at time of writing. We therefore need to make additional adjustments in order to build the project.

Right-click on the project and choose Properties. Expand the C/C++ Build menu, click on Settings. Click on Includes which is located below GCC C Compiler. You will see one directory listed in the textbox. You need to check that this directory exists. If not, change it to the correct one. For instance, this tutorial was written using Debian 7 amd64 with Oracle Java 8, hence the correct directory is

```
/usr/lib/jvm/java-8-oracle/include
```

In addition, one more directory is also needed by the project to build properly.

```
/usr/lib/jvm/java-8-oracle/include/linux
```

The project also needs another flag in order to build properly. In the Properties window, click on Miscellaneous below GCC C Compiler. Add `-fPIC` to the text field next to Other flags. Click on Ok to save the settings.

You can now build the project by choosing Project > Build Project in the menu. If everything went well you shall have generated a file `libjpgf.so` in `Release (posix)/`. You can check that the dependencies of `libjpgf.so` is fulfilled (i.e. it finds the C-runtime) by executing the following in a terminal

```
ldd libjpgf.so
```

If you cannot see 'not found' anywhere in the results, all dependencies are met. However if the C-runtime libraries are missing then `LD_LIBRARY_PATH` is probably not set. This is achieved by executing the following in the terminal:

```
export LD_LIBRARY_PATH=/usr/local/lib
```

This is only a one time setting and the variable will not exist for the next terminal session. This is however not a problem, since the libraries will be used by Apache Tomcat which will set the variable at startup.

Finish the tutorial by moving the wrapper library to the correct location.

```
mv libjpgf.so /usr/local/lib
```

A.1.4 Using the Java-runtime

This application can be executed by using any application server that supports WAR-files. The WAR-file in this project is built using Maven, which also can upload the file to an instance of the application server Tomcat. This method is very convenient since it automates a lot work. The following section describes how to install and configure Tomcat and Maven to work with this project.

Download and install Tomcat 8 and Maven (here by using aptitude package-manager).

```
$ sudo apt-get install tomcat8 tomcat8-admin maven
```

Tomcat requires an uploader to have the correct permissions. Edit `/etc/tomcat8/tomcat-users.xml` and add the following:

```
/etc/tomcat8/tomcat-users.xml
-----
<tomcat-users>
  <role rolename="manager-gui"/>
  <role rolename="manager-script"/>
  <user username="admin" password="secr3t" roles="manager-gui,manager-script"/>
</tomcat-users>
```

As the application will use the generated wrapper library `libjpgf.so`, we need to make a proper reference to this library and its dependencies (the C-libraries). This is achieved by creating a new file `setenv.sh` in the directory `/usr/share/tomcat8/bin/`, the location of this directory can differ on different Linux-distributions. The directory shall contain the file `catalina.sh`, so a search on the file should show the correct directory.

Create the file `setenv.sh` and add the following

```
/usr/share/tomcat8/bin/setenv.sh
-----
#!/bin/sh
export LD_LIBRARY_PATH=/usr/local/lib:$LD_LIBRARY_PATH
```

Restart Tomcat for the changes to take effect.

```
$ sudo service tomcat8 restart
```

The next thing we would like to do is to allow Maven to upload applications to Tomcat. As Tomcat now has an admin user with a password, we can use this to setup a server definition in Maven.

Add the following to `/etc/maven/settings.xml`

```
/etc/maven/settings.xml
-----
<servers>
  <server>
    <id>localTomcatServer</id>
    <username>admin</username>
    <password>secr3t</password>
  </server>
</servers>
```

The field `id` is used by the application to define that it shall be uploaded to the server we just specified.

The project can now be uploaded to tomcat by executing the following.


```
$ cd <project_directory>/nlparser  
$ mvn tomcat7:deploy
```

The application shall now be accessible through the URL
`http://localhost:8080/nlparser`.

IMPLEMENTATION DETAILS

B.1 MOCK DATA GENERATION

B.2 WEB APPLICATION

ACRONYMS

GF	Grammatical Framework
RGL	Resource Grammar Library
Java EE	Java Enterprise Edition
PGF	Portable Grammar Format

BIBLIOGRAPHY

- [1] Marissa Mayer. Seattle Conference on Scalability: Scaling Google for Every User. <https://www.youtube.com/watch?v=Syc3axgRsBw>, 2007. [Online; accessed 1-July-2014].
- [2] Aarne Ranta. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford, 2011. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).
- [3] Alfred V. Aho, Ravi Sethi, and Jeffrey D. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1986. ISBN 0-201-10088-6.
- [4] Aarne Ranta. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2009.
- [5] Krasimir Angelov, Björn Bringert, and Aarne Ranta. Speech-enabled hybrid multilingual translation for mobile devices. *EACL 2014*, page 41, 2014.
- [6] Krasimir Angelov. Incremental parsing with parallel multiple context-free grammars. In *European Chapter of the Association for Computational Linguistics*, 2009.
- [7] Krasimir Angelov. *The Mechanics of the Grammatical Framework*. Chalmers University of Technology, Göte, 2011. ISBN-13: 978-91-7385-605-8.