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 Generation of mock data 14
- 2.3 Grammar development 14
  - 2.3.1 Supported instructions 14
  - 2.3.2 Names 14
  - 2.3.3 More questions 15
  - 2.3.4 Extending the grammar 16
- 2.4 Boolean operators 20
- 2.5 Suggestion engine 22
  - 2.5.1 Populating the Solr index 23

3 DISCUSSION 27
- 3.1 Contributions 27
- 3.2 Negative things 27

A GF RUNTIME SYSTEMS AND LIBRARIES 28
- A.1 GF runtime systems 28
  - A.1.1 Portable Grammar Format 28
  - A.1.2 GF libraries 28
  - A.1.3 Building and installing the C to Java wrapper library 29
  - A.1.4 Using the Java-runtime 30

BIBLIOGRAPHY 32

1 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 search[1], 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. The 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 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 has 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 transaltes 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 verca. 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`.

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

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`.

```
abstract Instrucs = {
  flags startcat = Instruction;
  cat
    Instruction ; -- An Instruction
    Subject ;  -- The subject of an instruction
    Relation ; -- A verb phrase
    Object ;   -- an object

  fun
    MkInstruction : Subject -> Relation -> Instruction ;
    People : Subject ;
    Know : Object -> Relation ;
    Java : Object ;
}
```

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*.

```
concrete InstrucsEng of Instrucs = {
  lincat
    Instruction = Str ;
    Subject = Str ;
    Relation = Str ;
    Object = Str ;
  lin
    MkInstruction subject relation = subject ++ "who" ++ relation ;
    People = "people" ;
    Know object = "know" ++ object ;
    Java = "Java" ;
}
```

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 string[2] 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.

```
concrete InstrucsSolr of Instrucs = {
    lincat
      Instruction = Str ;
      Subject = Str ;
      Relation = Str ;
      Object = Str ;

    lin
      MkInstruction subject relation = "q=" ++ subject ++ "AND" ++ relation ;
      People = "object_type : Person" ;
      Know object = "expertise : " ++ object ;
      Java = "Java" ;
}
```

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 abstact 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)

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 anteriors. 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.

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

`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 automatically. 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 to only allow translation of plural forms.

```
concrete InstrucsEng of Instrucs = open SyntaxEng, ParadigmsEng in {
  lincat
    Instruction = NP ;
    Subject = N ;
    Relation = RS ;
    Object = NP ;

  lin
    MkInstruction subject relation = mkNP aPl_Det (mkCN subject relation) ;
    People = mkN "person" "people" ;
    Know object = mkRS' (mkVP (mkV2 (mkV "know") object)) ;
    Java = mkNP (mkPN "Java") ;

   oper
      mkRS' : VP -> RS = \vp -> mkRS (mkRCl which_RP vp) ;
}
```

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 concerete 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`.

```
incomplete concrete InstrucsI of Instrucs = open Syntax, LexInstrucs in {
  lincat
    Instruction = NP ;
    Subject = N ;
    Relation = RS ;
    Object = NP ;

  lin
    MkInstruction subject relation = mkNP aPl_Det (mkCN subject relation) ;
    People = people_N ;
    Know object = mkRS’ (mkVP (know_V2 object)) ;
    Java = java_NP ;

   oper
      mkRS’ : VP -> RS = \vp -> mkRS (mkRCl which_RP vp) ;
}
```

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.

```
interface LexInstrucs = open Syntax in {
   oper
     person_N : N ;
     know_V2 : V2 ;
     java_NP : NP ;
}
```

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.

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

Figure 12: Lexicon instantiation of English

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

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

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`.

```
concrete InstrucsEng of Instrucs = InstrucsI with
                                         (Syntax = SyntaxEng),
                                         (LexInstrucs = LexInstrucsEng)
                                         ** 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.

```
concrete InstrucsSwe of Instrucs = InstrucsI with
                                             (Syntax = SyntaxSwe),
                                             (LexInstrucs = LexInstrucsSwe)
                                             ** 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.

```
Instrucs> p -lang=InstrucsEng "people who know Java" | l -lang=InstrucsSwe
personer som kan Java

personer som kan Java

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.

2 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 questions the application can translate, how can she use the application? GF has the power to suggest valid words of an incomplete sentence by using incremental parsing. [5]. 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 it 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

One solution to this problem is ; write about how to use chunks and stuff

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 gui[1].

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

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

2.1.4.1 *Running the application*

Web applications built in Java are usually has the WAR file format. It 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 GENERATION OF MOCK DATA

This section describes how the data used by the application is generated. In order to generate data, one must know what to generate. In this context, this means that

2.3 GRAMMAR DEVELOPMENT

This section continues the work on the grammar introduced in Section 1.4.1.

2.3.1 *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.

2.3.2 *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

```
-- Abstract syntax
MkObject : String -> Object ;
-- RGL implementation
MkObject str = mkNP (mkPN str.s) ; -- PN = Proper Name
-- Solr implementation
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.

```
-- Abstract syntax
MkObject : Symb -> Object ;
-- RGL implementation
MkObject symb = mkNP (SymbPN symb) ; -- PN = Proper Name
-- Solr implementation
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.

2.3.3 *More questions*

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.

2.3.4 *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*.

2.3.4.1 *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.

```
-- Instructions.gf
cat
  Internal ;
  External ;
  Resource ;
fun
  People   : Internal ;
  Customer : External ;
  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.

```
cat
  InternalRelation ;
  ExternalRelation ;
  ResourceRelation ;
fun
  Know     : Object -> InternalRelation ;
  UseExt   : Object -> ExternalRelation ;
  UseRes   : Object -> ResourceRelation ;
  WorkIn   : Object -> InternalRelation ;
  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`.

```
cat
  Instruction ;
fun
  InstrucInternal : Internal -> InternalRelation -> Instruction ;
  InstrucExternal : External -> ExternalRelation -> Instruction ;
  InstrucResource : Resource -> ResourceRelation -> Instruction ;
```

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

2.3.4.2 *Concrete syntax for RGL*

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

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

```
lincat
  Instruction = NP ;
  Internal, External, Resource = N ;
  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`.

```
lin
  People = person_N ;
  Customer = customer_N ;
  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`.

```
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) ;
```

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.

```
lin
  InstrucInternal internal relation = mkI internal relation ;
  InstrucExternal external relation = mkI external relation ;
  InstrucResource resource' relation = mkI resource' relation ;

oper
  mkI : N -> RS -> NP = \noun,rs -> mkNP aPl_Det
                            (mkCN noun rs) ;
```

Figure 26: RGL concrete syntax with new instruction implementations

2.3.4.3 *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.

```
lincat
  Instruction = Str ;
  Internal, External, Resource = Str ;
  InternalRelation, ExternalRelation, ResourceRelation = Str ;
  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.

```
lin
  People = "Person" ;
  Customer = "Organization" ;
  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.

```
lin
  Know obj = "KNOWS" ++ ":" ++ obj ;
  UseExt obj = "USES" ++ ":" ++ obj ;
  UseRes obj = "USES" ++ ":" ++ obj ;
  WorkWith obj = "WORKS_WITH" ++ ":" ++ obj ;
  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.

```
lin
  InstrucInternal internal relation = select internal relation ;
  InstrucExternal external relation = select external relation;
  InstrucResource resource’ relation = select resource’ relation;

oper
  select : Str -> Str -> Str = \subj,relation ->
                  "select?q=*:*&wt=json&fq=" ++ "object_type :"
                  ++ subj ++ "AND" ++ relation ;
```

Figure 30: Solr concrete syntax with new instruction implementations

2.4 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.

```
lin
  And_O : Object -> Object -> Object ;
  Or_O : Object -> Object -> Object ;
```

Figure 31: Abstract syntax for boolean operators and objects

The RGL implementation is shown in Figure 32.

```
lin
  And_O o1 o2 = mkNP and_Conj o1 o2 ;
  Or_O 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.

```
lin
  And_O o1 o2 = "(" ++ o1 ++ "AND" ++ o2 ++ ")" ;
  Or_O 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.

```
lin
  InternalAnd : InternalRelation -> InternalRelation -> InternalRelation ;
  InternalOr : InternalRelation -> InternalRelation -> InternalRelation ;

  ExternalAnd : ExternalRelation -> ExternalRelation -> ExternalRelation ;
  ExternalOr : ExternalRelation -> ExternalRelation -> ExternalRelation ;

  ResourceAnd : ResourceRelation -> ResourceRelation -> ResourceRelation ;
  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.

```
lin
  InternalAnd rs1 rs2 = mkRS and_Conj rs1 rs2 ;
  InternalOr rs1 rs2 = mkRS or_Conj rs1 rs2 ;

  ExternalAnd rs1 rs2 = mkRS and_Conj rs1 rs2 ;
  ExternalAnd rs1 rs2 = mkRS or_Conj rs1 rs2 ;

  ResourceAnd rs1 rs2 = mkRS and_Conj rs1 rs2 ;
  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.

```
lin
  InternalAnd s1 s2 = "(" ++ s1 ++ "AND" ++ s2 ++ ")";
  InternalOr s1 s2 = "(" ++ s1 ++ " OR " ++ s2 ++ ")";

  ExternalAnd s1 s2 = "(" ++ s1 ++ "AND" ++ s2 ++ ")";
  ExternalOr s1 s2 = "(" ++ s1 ++ " OR " ++ s2 ++ ")";

  ResourceAnd s1 s2 = "(" ++ s1 ++ "AND" ++ s2 ++ ")";
  ResourceOr s1 s2 = "(" ++ s1 ++ " OR " ++ s2 ++ ")";
```

Figure 36: Solr concrete syntax for boolean operators and relations

2.5 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 38 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.

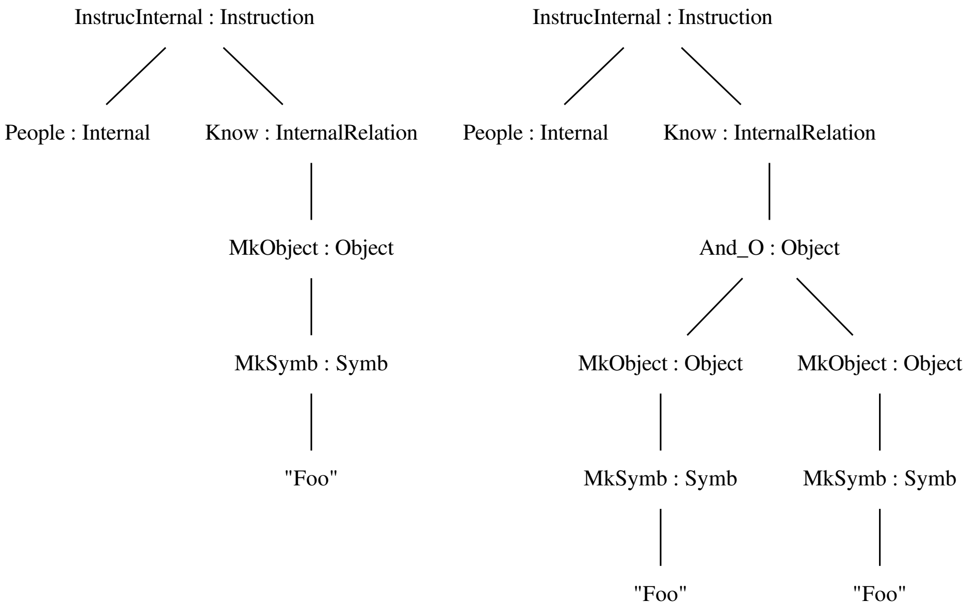

Figure 38: Visualization of abstract syntax trees with depth 4 and 5

2.5.1 *Populating the Solr index*

It is now time to linearize the generated trees and store them in a Solr index. 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 tree index. Retrieve 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: Abstract syntax trees with name types

In addition, as we store each name in a separate index, we also want to associate each name in that index with a type.

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.

```
suggestions(question) {
    // question = "persons who knew java or python and work in London"
    names[] = extractNames(question);

    // "persons who knew java or python and work in London" ===>
    // "persons who knew Skill0 or Skill1 and work in Location0"
    question = replaceNamesWithTypes(question, names);

    // suggestions =
    //   { "people who know Skill0 or Skill1 and who work in Location0", ... }
    suggestions[] = findQuestions(question);

    foreach suggestion in suggestions {
        // "people who know Skill0 or Skill1 and who work in Location0" ===>
        // "people who know Java or Python and who work in London"
        suggestion = restoreNames(names, suggestion);
    }
    return suggestions;
}
```

Figure 42: Pseudocode for the suggestion algorithm

3 DISCUSSION

3.1 CONTRIBUTIONS

My work has improved PGF, at least a bit

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

3.2 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.

A GF RUNTIME SYSTEMS AND LIBRARIES

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 [6, 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[6][p. 13], especially with larger grammars. This is where the Portable Grammar Format (PGF)[6][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
```

The C-runtime for PGF is now installed.

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`. 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 Ubuntu 14.04 amd64 with OpenJDK 7, hence the correct directory is

```
/usr/lib/jvm/java-7-openjdk-amd64/include
```

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 some dependencies are missing, try to locate the files and move them to `/usr/local/lib` (or `/usr/lib` in some distros).

The last step is to move `libjpgf.so` into the correct directory.

```
mv libjpgf.so /usr/local/lib (or /usr/lib)
```

You have now installed the wrapper library.

A.1.4 *Using the Java-runtime*

Have not started on this yet... Don't forget how to set java lib path when using tomcat!

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. Incremental parsing with parallel multiple context-free grammars. In *European Chapter of the Association for Computational Linguistics*, 2009.

[6] Krasimir Angelov. *The Mechanics of the Grammatical Framework*. Chalmers University of Technology, GÃ¶teborg, 2011. ISBN-13: 978-91-7385-605-8.