Writing a parser for an artificial language, Atasi cnoba

Johan Holmberg
Department of Computer Science
Institute of Technology, Lund University
d01jh@student.lth.se

Abstract

This report describes an attempt to build a parser for a small and coherent artifical language. It presents a fairly complete grammar denoted on the Panini-Backus form, accomanied with a sample thesaurus, all implemented using the Prolog language. The project is part of the course Language Processing and Computational Linguistics given at Lund university in the autumn semester of 2006.

1 Introduction

While the World has a phletora of languages, some linguistics claim there to be 6500 languages thoughout the World, there still seems to be a need for artificially created languages. Those languages differ in their seriousness, longevity and span, but they all fill a void needed to be filled. Most of those languages will not prevail, and the actual need of the languages can thus be questioned. Nonetheless do they exist, and probably will for a long, long time.

This is an attempt to outline a possible way to automatically parse and validate such a language.

2 Rationale for an artificial language

There are mainly three different reasons for developing an artifical language. These are, in no particular order:

Auxiliarity. Languages, such as *Interlingua*, *Volapük* or *Esperanto*, resides in this category. They are constructed with the purpose of being a common language, either by merely being a *lingua franca*, or by even replacing existing languages.

Experimentation. Languages in this category are usually created as a mean of trying out ideas, usually regarding logic and philosophy. Proponents of the Sapir–Whorf hypothesis commonly uses this as a tool for exploring how languages inflicts on the human mind.

Artistic ambitions. These are languages intended to add an extra dimension to artistics works, such as books or movies. Famous languages, such as Tolkien's *Quenya* or *Klingon* of Star Trek fame are part of this group.

Apart from these main reasons, there is another one: *obfuscation*. Languages in this category are devised to hide information from a general public, while keeping it open to the speakers of the language in question. A lot of informal languages are created this way, but since they are secretive by nature, they tend to fall in oblivion as soon as their inventors and speakers drop them.

3 The Atasi cnoba language

The Atasi cnoba language was originally constructed for usage in refugee role playing games set up by Swedish scouts, as an integral part of the act. The rationale for this was that the participants, who act as refugees, were to feel unsecure and disorientated, but it was also meant to be a useful tool for the game leader – staff communication. As such, it has only been used a couple of times, due to the late date of maturity for the language.

The premises for developing a langauge like that, were clear and simple: The language had to

be different enough to be percieved as a foreign language, yet simple enough to learn in a matter of days, or even hours, if needed. This was accomplished by cutting back on grammatical nuances, relying on the fact that the language will not ever be used as a real world language. The developers were also keen of grouping words with similar meaning into more general words, relying on descriptors to further describe objects and phenomena as needed. The result is a, within it's domain, fairly useful language with no more than a hundred words, coupled with a handful of modifiers. Due to the compactness and regularity of the language, it turned out to be well-suited for computational handling.

3.1 Vocabulary

The vocabulary of Atasi cnoba is heavily influenced by the author's past experiences in the field of linguistics, coupled with the premise that the language should be hard to grasp for an outsider. Hence, it's main sources are the Romance, German and Slavonic language families. The words coming from these families are usually distorted in order to make them unintelligble by non-speakers.

Some words, as in the case of the numerals, are direct loanwords from Georigan, a Kartvelian language. In fact, *atasi cnoba* is Georgian for *a thousand words*. Those words are usually not distorted, as a discreet homage to the language's Georgian roots.

3.2 Sound system

In order to make the language sound foreign and strange, the need of a distinct sound system was needed. We chose the Georgian sound system as a basis to build on, mainly because the many consonant sounds in the language and the relative lack of vowels are radically different from Swedish. The distinct sounds are however relatively easy to mimic for a non-native speaker, and are thus easy to comprehend.

Georgian features 33 consonant sounds and 5 vowel sounds. Since quite a few of the western European sounds doesn't appear in the Georigan sound system, imported words and names has to be retrofitted before being used.

Unlike most modern day European languages, aspirated and unaspirated plosives and affricatives aren't allophones in the Georgian language. As a result, some distinct words in Atasi cnoba may seem to be homonyms, while they are in fact distinct words on their own merits, and will consequently be treated as different entities by the parser.

Some simplifications had to be done in order for Atasi cnoba to be easy to grasp. This is especially true in regards to the somewhat perverse consonant clusters used by the Georgians. Clusters of up to six consonats are not uncommon, peaking at eight consecutive consonants, whereof some consonats would be rendered as two or more separate sounds by a native Swedish speaker. The Swedish language, as a comparison, features no more than three consonants in a row, not counting possibly appended suffixes.

Atasi cnoba is to be stressed using trochee feet, but as this has no consequences in regards to spelling, this fact can be easily dismissed in the scope of this project.

3.3 Writing system

Since the sound system is effectively Georgian, we also use the Mkhedrulian alphabet, which is used for writing Georgian, as well as the other Kartvelian languages. The alphabet works flawlessy in conjucture with the Georgian sound system. Furthermore, it has the property of looking radically different from the European scripts, although those alphabets share the same ancestor as the Mkhedrulian: the classical Greek script.

The Mkhedrulian alphabet provides a $1 : 1$ mapping for each sound in Georgian, and as a consequence also for Atasi cnoba. Another useful trait of the Mhedrulian alphabet is the lack of capital letters. What this means in terms of this project, is that we don't have to take conversion of upper case letters into lower case letters for lookups in the thesaurus into account while developing the system.

The Mkhedruli alphabet is covered by the Unicode standard, as discussed later.

3.4 Grammar

The grammar of Atasi cnoba is, as previously mentioned, thought to be simplistic in it's approach. It lacks ways of morphologically denote the words' roles in the sentence. Instead, it relies on an absolute subject-verb-object order throughout the languages. This is even true in questions, different from, say English and Swedish, where questions usually are constructed in a verb-subject-object order. This is possible due to a type

of question mark, which surrounds the whole sentence, not unlike Portuguese or Spanish, the difference here being that the question mark is actually realized as a distinct phoneme.

3.4.1 Morphology

Atasi cnoba is an agglutinative language, meaning that new words are created by the means of affixes added to a stem. By doing this, a verb might turn into a noun, and a noun turn into a descriptor. More importantly, affixes are used for verb conjugation, descriptor comparison and noun inflection.

Some affixes, eg. *-ni-* which acts as a negator, are universal for the major lexical classes.

3.4.2 Nouns

Atasi cnoba lacks both grammatical numbers and genders, and thus keeping the different noun forms to an absolute minimum. The nouns are, however, slightly inflicted in order to accompany verbs, nouns and pronouns as descriptors. There are three cases, not counting the nominative case, into which a noun can be inflicted. These are:

The possessive case is used to denote dependencies and possession. A simple example would be *ia ioaniav*, meaning *Johan's flower* or *the flower of Johan*. The case is marked by adding the *-av* suffix to the possessing noun.

The locative case denotes where something is going on. It marks a position, not a direction. The *-la* suffix is used for this, as in *ek es lundla*, meaning *you are in Lund*, while *ik ial do lund* translates into *I'm going to Lund*.

The temporal case is somewhat similar to the locative case, but denotes a specific moment in time, rather than a position. The case marker used here is *-tem*, as seen in *ik vaial do ek hok'ratem*, which literally means *I will be going to you near-object-time*, or *I'll be going to you soon*.

Nouns, aswell as proper names and pronouns, can also be used for comparison in hierarchies, using the prefixes *sup'-*, *hok'-* and *sub-*, meaning *over*, *near, at the same level as* and *below*. This treat is used in constructed words, such as *sup'utur*, meaning *mother, goddess, female superior* and *hok'ekla*, meaning *near where you are*.

It is possible to create nouns out of verb stems by adding suffixes. Such a constructed noun may take on one out of two roles: an agent, using the *-ari* suffix, or an essence, using the *-de* suffix. Thus, the verb *dokt'* (*to instruct*), would transform into the agent form *dokt'ari* (*teacher*), and into the essence form *dokt'de* (*instruction, learning*).

3.4.3 Proper names

Proper names are denoted by adding an extra *-i* at the end of the word. This is a remnant from the Georgian nominative case.

3.4.4 Pronouns

Atasi cnoba features nine pronouns, whereof seven are personal:

	singular	*plural*
1st person	ik	ikek (inclusive) ikak (exclusive)
2nd person	ek	ekek
3rd person	ak	akak

The remaining pronouns are *k'iak*, which roughly translates into the English pronouns *anybody, somebody*, and *tot'ak*, which roughly translates into the English pronouns *all, everybody*.

The inclusive and exclusive versions of the 1st person plural are used to denote whether the speaker includes it's audience into the "we collective" or not.

3.4.5 Verbs

The verb system in Atasi cnoba is very simple and completely regular. Each verb has three conjucations: the past, the present and the future tenses. The future tense also acts as an imperative tense, while the present acts as the verb stem and infinitive.

past	*present*	*future*
a-stem	0-stem	va-stem

3.4.6 Descriptors

In Atasi cnoba, there is usually no reason to distinguish between adjectives and adverbs, and as a result, those groups are usually treated as a single, larger group: the descriptors. Descriptors are composed either by inflected nouns, nouns with the *-(sh)ko* suffix (roughly translated *-like*), or by proper adjectives.

Semantically, the non-nominative cases of the nouns in Atasi cnoba are treated as descriptors, as they are solely used to describe either a verb or a noun.

The descriptors are given in a certain order, answering the questions *whose*, *how*, *where* and *when*.

3.4.7 Numbers

Numbers in Atasi cnoba can take on two roles, depending on their placement within the sentence. A number following a noun is an ordinal number, while a number in front of a noun is a cardinal number.

The counting system in Atasi cnoba is a simple ten-based, additive system, where each multiplicient above 1000 is realized as a power of ten.

3.4.8 Prepositions

The Atasi cnoba prepositional system is rather simplistic, and is currently made up of only two words *do* (*to*) and *od* (*from*). They denote directions, and nothing more.

4 System overview

The purpose of the parser is mainly to validate a small corpus written in Atasi cnoba using a manually engineered grammar. A secondary goal is to explore the possibilites for a translator from Atasi cnoba to Swedish.

The parser is made up of two separate tools: the chunker and the actual parser. The chunker preprocesses the data, and breaks affixes away from their word stems, making the data easier to handle for the parser.

4.1 Chunker

The chunker is written i PHP, and is invoked from the command line. Being written in PHP, it also means that the chunker is easily embeddable in a web application, which is part of the rationale for writing it in PHP. However, being command line based allows it to handle large amount of data, which would otherwise be hard, due a feature in the web server version of PHP, which limits the allowed execution time. This constraint does not apply for the standalone version.

The design of the chunker is rather naïve and simplistic. Technically speaking, it loops through the corpus three times, one loop for each major group of lexical classes. These are, in given order, descriptors, objects (nouns, proper names and pronouns) and verbs. Please note that descriptors derived from nouns, are handled as nouns for the sake of simplicity. The order was selected in regards to the way that words are constructed within the Atasi cnoba language. The parser also separates punctuation characters from their adjacent words, making them true atoms.

In order to safely separate affixes from the word stems, the chunker has to have some rudimentary knowledge of the Atasi cnoban vocabulary, since some words, eg. numerals, might seem to collide with some apparently ambigious rules.

When the chunker is done separating the affixes from the word stems, the corpus is transformed into a Prolog query and written to a file. This file, which is linked to the parser, is later compliled and run within the Prolog environment, as we will see next.

4.2 Parser

The parser is written in the Prolog language. It uses a Panini-Backus-like grammar to describe the Atasi cnoban syntax. Using this grammar, we are able to traverse the incoming data, validating each and every word according to the syntax.

The parsing process is simply put a giant pattern matching, requiring each atom to be in it's correct location. The parser is accompanied by a thesaurus, containing a great portion of the words found in Atasi cnoba. The thesaurus also contains rough Swedish translations of the words, making the parser able to very roughly translate the Atasi cnoban text, rather than just parsing and validating it.

The output is also very simplistic. The Prolog system merely tells the use whether the query passed the grammar checks or not. In case it did pass, it also leaves a printout of the rough translation.

No effort was put into pointing out where the errors occured, nor is the printout beautifully performed in this version.

4.3 Data formats

As the parser is supposed to merely validate a small corpus, there is no need for a specific data format. Hence, the input data is given in plain text.

4.4 Handling Unicode characters

The Mkhedrulian alphabet is fully covered by the Unicode, range 10D0–10FF. While SWI–Prolog is able to transparently handle Unicode, the practical means of using non-Latin characters within the development tools are rather sparse. Due to this, and the fact that Mkhedrulian input methods are not supported by standard keyboards, we have

chosen to transliterate the texts from Mkhedrulian script to Latin script, mostly according to the National translitteration rules. These rules, paired with the Mkhedrulian alphabet, are described in Appendix B.

4.5 Grammatical features not covered by the parser

Even though most of the grammar is covered, there is still one feature that is not covered by the parser. This is the compositional infix, *-na-*, used for gluing two or more nouns together, cf. the *-a-* in Scanian words such as *hunn-a-mad* (*dog food*). The feature was left out not as much on purpose as out of sheer neglect. It would, however, leave a significant impact to the structure of the chunker, since it would need to know each single word in the language in order to correctly split the nouns into their stems. The parser would also have to be slightly modified, although this would be a signficantly smaller operation, fully realizable in adding one simple rule.

5 Results

As part of the evaluation of the system, a number of pre-written texts were used. As Atasi cnoba isn't thought of as being a literary language, the number of texts written in it is rather sparse. In the texts that were availabe, we were able to detect several errors, proving the robustness of the parser. Those errors were of the syntactical kind as well as simple spelling errors, meaning that the parser could possibly double as a spellchecker.

The development process itself shed light over some ambiguousities in the language itself. Since the language itself isn't written in stone, those ambiguousities were taken care of, making the process somewhat a two-way process.

6 Possible future enhancements

An interesting side effect of the parser is, as mentioned above, the spellchecking function. Given a more elaborate and robust way of detecting those errors, the system might be useful for an interactive language validator, not unlike the spellchecking and grammarchecking functions in mainstream products, such as Microsoft Word.

While not being complete as of today, it would be interesting to see a more intelligent translating mechanism, possibly by changing the internal Prolog output by making it more elaborate, and then reverse that into Swedish.

It would aslo be interesting to see the translator from above implemented as a web service. Given that both of the programming languages used in the project are interpreted languages, this would not be too har to accomplish.

7 Conclusion

We have proven that it is possible to write a static parser for a constructed language, as shown by the results. It is quite possible to write a static parser, which covers the vast majority of a fairly small and well formed constructed language. It should however be noted that even for such a small language, the grammatical rules can be rather complicated. However, his does not imply that writing such a parser would be impossible for a real world language, as proven by the Sanskrit grammatician Panini as early as in the 1st millenum BC.

8 Acknowledgements

We wish to thank Pierre Nugues and Richard Johansson for their support during this project. We would also like to thank Marcus Agbrant for proofreading and Johan Muskala for inspiration.

References

H. N. Shenton, E. Sapir and O. Jespersen. 1925. *The Function of an International Auxiliary Language*. International Communication: A Symposium on the Language Problem, London 1931, pp. 65–94.

The Unicode Standard, Version 5.0. 2006. *Georigan*, Range: 10A0–10FF. http://unicode.org/charts/PDF/U10A0.pdf.

The national system of romanization. February 2002. State Department of Geodesy and Cartography of Georgia and the Institute of Linguistics, Georgian Academy of Sciences.

The Ashtadhyayi. 2003. Microsoft Corporation. http://www.bhashaindia.com/Patrons/LanguageTech/Ashtadhyayi.aspx.

T.R.N. Rao. 2006. The Panini-Backus Form in Syntax of Formal Languages. http://www.infinityfoundation.com/mandala/t_es/t_es_rao-t_syntax.htm.

A Running instructions

Start off by opening a terminal and navigating to the directory where you installed the program files. Make sure you have placed your input file inside this directory. Then invoke the chunker by typing

```
$> php chunk.php input.txt output.pl
```

You should now have a Prolog source file called `output.pl` in your directory. Start your Prolog environment, and then type

```
1 ?- qcompile(output).
```

Start the parser by typing

```
2 ?- translate(Translation).
```

If your corpus was valid, you should now get an `Yes` from the Prolog interpreter. Otherwise, you will get a `No`.
Good luck!

B The Mkhedrulian alphabet

The Mkhedrulian alphabet, accompanied by the translitteration rules, as advised by the Georgian Academy of Sciences.

Georgian	Latin	Our notation
ა	a	
ბ	b	
გ	g	
დ	d	
ე	e	
ვ	v	
ზ	z	
თ	t'	
ი	i	
კ	k	
ლ	l	
მ	m	
ნ	n	
ო	o	
პ	p	
ჟ	zh	
რ	r	
ს	s	
ტ	t	
უ	u	
ფ	p'	
ქ	k'	
ღ	gh	
ყ	q'	
შ	sh	
ჩ	ch'	
ც	ts'	
ძ	dz	
წ	ts	
ჭ	c	
ხ	kh	x
ჰ	h	

C A sample Atasi cnoba text

A sample corpus written in Atasi cnoba using the Mkhedrulian alphabet, coupled with a translitterated version of the same text.

ქარ იკ ეს უთურ ეკავ, და ეკ ეს ფენ იკავ. ქიაკთემ ეკ ნეც' იკ, იკ ვაფაქ ტოტაკ იკ პოვ. იკაკ ვაიალ დო ქიაკ, ქიაკლა იკეკ ნიაეს. ქიაკთემ იკ ეს ნიბრავ, შას იკ ეს პარი დო ამ, ქარ ფოთ ამდეავ.

k'ar ik es utur ekav, da ek es p'en ikav. k'iaktem ek nec' ik, ik vap'ak' t'ot'ak ik pov. ikak vaial do k'iak, k'iakla ikek niaes. k'iaktem ik es nibrav, (sh)as ik es pari do am, k'ar p'ot amdeav