Lex: a Software Project for Linguists

Chris Wilson
Aptivate
chris+lexreport@qwirx.com

Reference: Chris Wilson, "Lex: a Software Project for Linguists." *Technical Report SEE-J Hiphil* 6 [http://www.see-j.net/index.php/hiphil] (2009). Read April 2 2009.

Abstract

This paper describes a project called Lex, developed initially to assist a linguist with analysis of the Hebrew Bible and now being extended to other languages.

Lex is an implementation of the Role Lexical Module (RLM) by Winther-Nielsen [NWN 08]. It integrates with the tagged Biblical Hebrew corpus of the Workgroep Informatica (WIVU) and provides corpus navigation, and display of morphological and syntactic markup from that corpus. We summarise our use of this corpus and the steps needed to extend our approach to other languages and corpora.

We introduce the Emdros database [USP 08] used by Lex and the WIVU corpus, and explain the reasons for its choice in this project and the advantages that it appears to offer to corpus and computational linguists.

We present the rule-based active chart parser developed for Lex, and its extensions to support free word-order languages, including Dyirbal. We describe the features of attributes and unification which enable arbitrary restrictions on rule combination, to simulate the linguistic template structures of RRG, and generation of logical structures and focus structure. Winther-Nielsen has developed grammatical rules for Biblical Hebrew using this parser, and we believe that it should be powerful enough to work with any written language, and facilitate computational linguistics and fully automated machine translation.

We present an idea for a rule-based morphological analysis system that would work with Lex to enable parsing of other languages while avoiding the need to store each morphological word form in the lexicon. We give examples of the use of this system with Biblical Hebrew.

We present the new database-driven transliteration system in Lex, illustrated with examples from Biblical Hebrew, and describe its potential in developing and testing

transliterations of other languages using a test-driven development [KB 02] approach borrowed from software engineering.

We present our ideas for machine translation based on Role and Reference Grammar [VVLP], with optimisations for conversion of the parse tree from the source to the target language.

Introduction

Lex is a tool designed to help linguists to perform detailed analysis of text corpora and to publish their results:

- Browsing the corpus
- Searching for words (simple)
- Searching for grammatical constructs (advanced)
- Parsing (using a database of parser rules)
- Tree drawing (parse trees and manually)
- Transliteration (for exchange with other linguists)
- Logical structure definition (RRG theory)
- Semantic categories or ontologies of verbs (assists in translation and glossing)
- Lookup in other corpora (for glossing)

There is explicit support for data security, user access control, data separation (multiple databases) and controlled publication (granting public access to defined subsets of the data).

Lex could also be seen as a graphical user interface to a corpus database tool called Emdros. It adds many features which are not in Emdros, such as a lexicon, but fundamentally relies on Emdros for much of its work.

Lex is the result of a collaboration of two project partners over several years:

- Associate Professor, Teol. Dr. Nicolai Winther-Nielsen, teaches Hebrew Bible at the CLST and Persuasive Design at Aalborg University, and specialises in computational linguistics. He was the initiator of the project, has developed

the Role Lexical Module (RLM) for Biblical Hebrew, on which the Lex software is based, and published the first version of the linguistic use of the Lex software.

- Chris Wilson, MA, computer scientist and software developer, programmed the Lex software, designed and implemented the active chart parser and generic database-driven transliterator, and designed the rule-based morphological analyser and parse tree restructuring system.

We would like first and foremost to acknowledge the special role of Assistant Professor, Ph.D, Ulrik Sandborg-Petersen, who invented the Emdros database and who has provided great assistance in the implementation of the software. He has also implemented new features in Emdros to help us with our work.

Many others have also assisted and contributed ideas, particularly in the Role and Reference Grammar community, where we are especially indebted to Robert van Valin, Brian Nolan, Ricardo Mairal, Elizabeth Guest and Chris Butler. We would also like to thank Claus Tøndering for his help with the transliteration and machine translation ideas.

We are grateful to the license committee of the German Bible Society for giving us permission to use the WIVU database for our research, free of charge, subject to restrictions.

Requirements

The design of Lex was based on the Role Lexical Module (RLM) by Nicolai Winther-Nielsen. Lex was used to test and improve the design of the RLM. In order to do so, the RLM had to be applied to a particular language, in this case Biblical Hebrew (BH).

We used the BH corpus database from the *Werkgroep Informatica*, the Department of Biblical Studies and Computer Science at the Faculty of Theology, *Vrije Universiteit*, Amsterdam (WIVU). Lex needed to support the following features of this specific database:

- WIVU BH corpus organised by book, chapter, verse and clause;
- Hebrew surface consonant form, and the original WIT (ASCII) machine representation of Hebrew characters and inflection;

- complete encoded syntactic tagging of clauses, parts of speech, nucleus, arguments and subject and object (not PSA).

Lex provided a user interface to the corpus, and was used for linguistic analysis of this corpus by Nicolai Winther-Nielsen.

We would like to further develop the ideas used in Lex to generalise them for other languages, and to collaborate with linguists working in those languages to integrate corpora and develop rule sets and lexica that can be used for fully automated machine translation. We intend to implement machine translation based on the RRG theory for conversion from syntax to semantics and back.

The processes of conversion from syntax to semantics has been broken down as follows. This list may not be complete, because Lex is not yet finished, however we are nearly there. Lex should support the following:

- multiple languages, by definition for a machine translation project;
- storage of corpus text in a database, and enable easy searching and retrieval of corpus text, to help linguists to develop and test lexica and parser rule sets;
- multiple corpora in each language, to help linguists to correlate across different corpora and check their work;
- transliteration of local languages, to help users to prepare papers and reports on their work;
- sharing of corpora between different users, while isolating their changes to enable them to work independently;
- construction and testing of morphological rules, to enable first-level morphology transformations and to generalise the lexicon from surface forms to semantic primitives;
- construction and testing of parsing rules (templates), to enable automated parsing of clause structure, for machine and machine-aided translation;
- easy experimentation with new parser rules and examination of their effects on the automated parsing of the corpora, to enable efficient development and refinement of rule sets;
- construction and maintenance of the lexicon, which is crucial for converting the syntax tree into logical structures and back;

- connection of the lexicon to the corpora, to assist with development of the lexicon;
- evaluation of the actual semantics (logical structures) in the corpora;

In addition, to enable conversion of semantics back to syntax in another language, we posit that the following steps will be required:

- parsing of logical structures back into syntax in another language;
- rendering syntax trees as surface text, including reverse morphology.

Some features of the project are not linguistically motivated, but rather by a desire to encourage sharing and collaboration in the community, and to build useful tools for the community. Therefore, we decided that Lex must also support:

- assistance for collaboration between researchers working at remote locations;
- low cost availability and easy installation of the software;
- flexibility for future development through open source code and well-structured code;
- customisation of the software through source code availability;
- reliability through unit and integration testing.

Design Decisions

Possible means to satisfy each requirement are analysed, one is chosen and the reasons for the choice are explained below.

WIVU BH Corpus

The WIVU BH corpus is available in two forms: as multiple ASCII table files in fixed column width format, designed for use by WIVU’s original Pascal software; and as an Emdros database which allows efficient searching and retrieval of complex data structures.

We could have used either format, but the Emdros data was significantly easier to access and use. We did have to overcome some hurdles in accessing the data from Java, and developed software libraries to make this easier. We have worked on Emdros to make this process easier, to resolve some bugs and to add missing features.

Chris Wilson. 2009. “Lex: a Software Project for Linguists.” *Technical Report SEE-J Hiphil* 6 [http://www.see-j.net/index.php/hiphil] (2009)

The corpus is licensed from WIVU, and unfortunately not fully available to the public, but we have permission to display certain parts of it as examples to the public, and full access may be granted by permission from the German Bible Society.

Complete Syntactic Tagging

The contribution of WIVU in producing their database was not just in entering the Hebrew characters onto a computer. The Workgroup has spent many years conducting important research on the corpus and machine analysis of BH, resulting in a machine-generated and manually checked and corrected database of syntactic structure of the BH corpus. It was used in the dissertation of Winther-Nielsen (1995).

Access to this database gave us a unique advantage, as for a long time we did not need to develop a parser and we could rely just on the syntactic analysis. However, their analysis was not based on a functional-structuralist grammar like Role and Reference Grammar. It uses several concepts which do not exist in RRG, so their work was not directly usable for RRG analysis.

Here is an example of their structural analysis of Genesis 1,1 from BH:

type / monad	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
book	1 (Genesis)										
chapter	1										
verse	1 (in beginning he created God (et) the heavens and (et) the earth)										
clause	28737										
phrase	PP/Time		VP/Pred	NP/Subj	PP/ObjC						
subphrase					mother				daughter		
word	prep	noun	verb	noun	prep	article	noun	conj	prep	article	noun
surface	*bə-*	*rēʔšît*	*bārāʔ ?*	*ĕlōhîm*	*ʔēt*	*ha-*	*ššāmayim*	*wə-*	*ʔēt*	*hā-*	*ʔārec*

Table 1: Syntactic structures from WIVU database for Genesis 1,1 (key components)

If we compare the tree structure of the WIVU data with the canonical RRG tree for the same clause, we can see that a lot of work is still required to convert the one into the other!

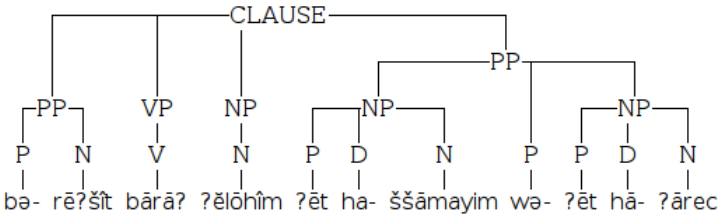

Figure 1: Syntax tree from WIVU database (key components)

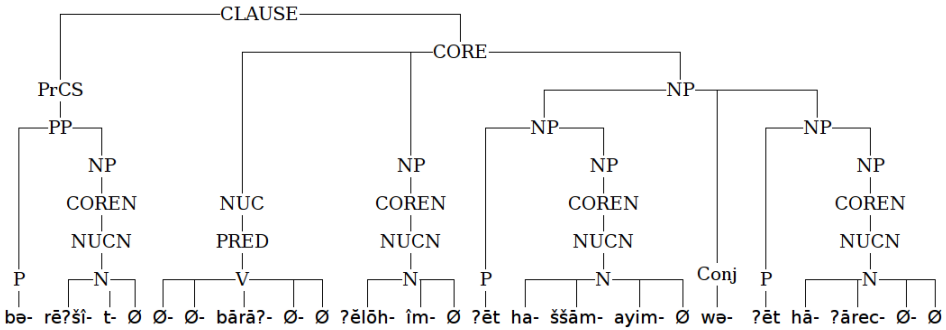

Figure 2: Syntax tree in RRG form

Until we had a parser, we fed the *phrase* level units to the logical structure generator. Now we have replaced the WIVU syntactic structure with RRG from the word level up. Analysis at the sub-word level, i.e. morphology, currently uses the data about *word* structures in the WIVU database. Alternatives for other languages will be discussed further under *Morphological Rules* below.

Multiple Languages

Storage of characters from any language on a computer requires the definition of a character set, which maps characters (*glyphs* or shapes) to binary codes for storage.

In the past, many disparate character sets were used, some overlapping or conflicting, sometimes with incompatible alternatives for the same language. This collection of characters sets could best be described as an absolute mess, and prevented and complicated the exchange of information between computers with different language settings (character sets) for many years.

Recently, a group of experts has developed a single character set called Unicode, which includes almost every known language and character in the world. The aim was to include every language and character, but some may have been omitted due to lack of knowledge or standardisation. We can expect the omissions to be corrected in subsequent versions of the standard.

The advantages of using Unicode over separate character sets to encode a corpus are many:

- the character set for a corpus need not be decided in advance;
- it is possible to mix multiple languages in a corpus;
- it is easy to exchange information in different languages between computers, and between software programs on the same computer;
- the binary encoding is reasonably efficient and not difficult to implement.

Availability of Unicode fonts used to be a major problem, as several traditional or dead languages had non-Unicode fonts available but no Unicode fonts. With the growing popularity of Unicode, it seems likely that most living languages already have Unicode fonts, and any new fonts developed for new languages are more likely to be Unicode than not.

Unicode encoding may be less efficient, using more disk space, memory and CPU power, than language-specific character sets. This is inevitable due to the flexibility of the Unicode character set, and the choices made by the designers of Unicode. In the light of the low and falling cost of memory and disk space, and the quantity of textual information that are likely to be processed by the software package, the authors do not consider this a significant problem.

No alternative universal character set is known to the author. The decision therefore comes down to Unicode versus individual language-specific character sets. The advantages of Unicode over those character sets are considered to outweigh the disadvantages.

The WIVU corpus provides several encodings of the Biblical Hebrew (BH) text. As it pre-dates the Unicode standard, it did not originally contain Unicode characters. Instead, it uses an encoding called WIT, which is a reversible transliteration of Hebrew characters into ASCII.

Chris Wilson. 2009. “Lex: a Software Project for Linguists.” *Technical Report SEE-J Hiphil* 6 [http://www.see-j.net/index.php/hiphil] (2009)

We developed Lex using this encoding, as it was easy to work with for programming. Ulrik Sandborg-Petersen used the conversion algorithm developed by James Tauber, and data tables by Eli Evans, to convert this encoding into Unicode, and we are about to switch to using the Unicode encoding throughout Lex, as this will offer better compatibility with other corpora.

Luckily, the WIVU database now includes Unicode (UTF-8) encoded versions of the Hebrew surface text, so we were able to display and work with actual Hebrew characters without any extra work in conversion between character sets. In addition, the standard Windows fonts have reasonable Hebrew glyphs, and better results can be obtained by the user installing the free SIL Hebrew fonts.

Storage of Corpus Text

Linguists will often work with a corpus of text in the languages that they study, since it provides a useful subject for research and for testing theories about the language. The corpus is traditionally stored as a flat text file, with characters in sequence, separated by spaces into words. This approach is simple and works well for small corpora, but it is not efficient to search larger corpora in this way, nor to store linguistic structural data alongside the text, nor to exchange revisions to the corpus with other researchers.

Ulrik Sandborg-Petersen built on the work of Crist-Jan Doedens to develop a database specifically for storing corpora and structural data about their contents. This database is called Emdros, and it offers a number of advantages over traditional relational databases. Some of these will concern us in the next section, but for now it suffices to say that Emdros makes it significantly easier to navigate, browse and modify structured text than either flat files or relational databases.

The extent of this is shown by the fact that the German Bible Society[1] and the Logos Bible Software company[2] have both purchased licenses to use Emdros. Emdros is the only database software known by the author to be available free of charge (as open source) and designed specifically for storage of corpora and linguistic structural data in any language.

The main disadvantage of Emdros from a user's point of view is that it does not have a graphical user interface, that is a simple graphical tool to allow users to enter and query their data. Thus, working with Emdros is currently significantly harder than

1 http://www.sesb-online.de
2 http://www.logos.com/

working with a flat text file. In order to choose Emdros as the underlying database for this project, it would therefore be necessary to develop a simple user interface to make using Emdros easier to use, while also giving access to the advantages of the powerful structured data storage and query tools that it provides.

We expect that at least some linguists will want to work with reasonably large quantities of text, and therefore that the advantages of using Emdros outweigh the disadvantages. We also hope that the user interface to Emdros that will be developed as part of this project will be a useful tool in its own right.

Navigation of the corpus is currently completely tied to the WIVU database structure of Biblical Hebrew, and therefore needs an overhaul before other corpora can be used successfully and navigated easily. A more generic navigation system would allow the administrator to select some object types that would be used for navigation, and the Emdros *feature* that would be displayed in the drop-down box for each object type. It might be useful to apply processing to these features, such as transliteration. The current system is a special case of this generic one.

Below is a screen shot of the current navigation interface, and the object types and attributes used to populate it:

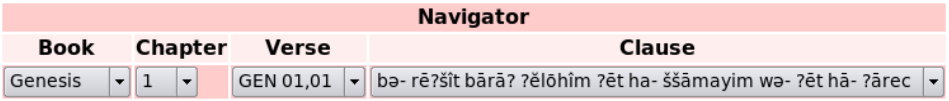

Figure 3: Lex navigation interface for Biblical Hebrew

Order	Emdros Object	Emdros Feature	Processing
1	book	book (name)	none
2	chapter	chapter (name)	none
3	verse	verse_label	none
4	clause	[word GET many]	concatenate and transliterate

Table 2: Lex navigation interface generalised to Emdros objects and features

Searching of the corpus is currently possible in two ways:

- using a simple word search for Hebrew consonants against a single attribute of the [word] object type in Emdros;
- powerful but complex MQL search which allows any [clause] object to be selected by its features, or the objects or features that it contains.

It would be useful to introduce an intermediate type of search that would allow users to quickly select the object type, feature and value that they were looking for, which would be built into an MQL query. Claus Tøndering, a Danish IT consultant and software developer, has developed an Emdros search engine like this in Java as part of his Hebrew teaching software 3ET[3], and we may be allowed to use some of the ideas or code in Lex.

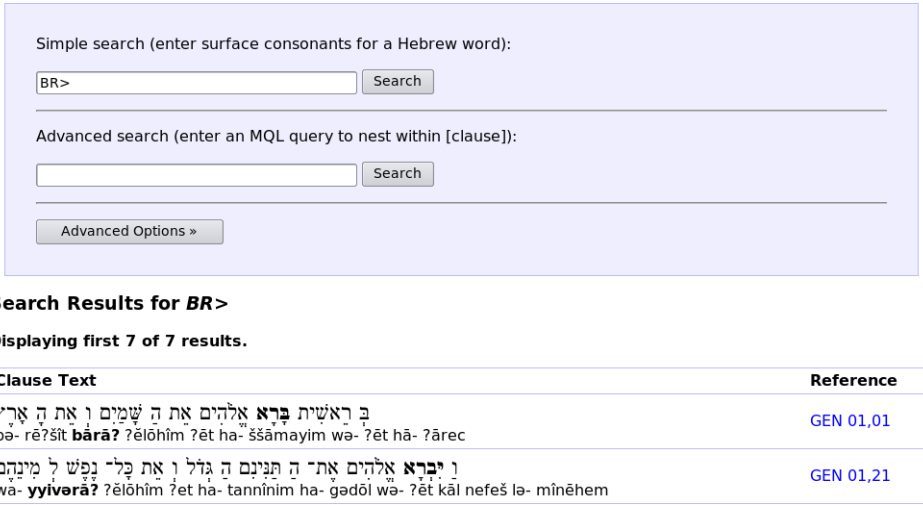

Figure 4: Search interface for Biblical Hebrew, showing two results

As before, it would be useful here for the administrator to be able to specify the type of object returned (to control the amount of text, e.g. one clause or one paragraph), and the features and processing that would be displayed for each result. Figure 4 shows an example of the current search interface, displaying [clause] objects with the Hebrew and transliterated text, and a location indicator, which is a hyperlink to a more detailed display for that clause.

3 Information about 3ET can be found at: http://3bm.dk/index.php?p=82

Multiple Corpora

Although Lex was originally developed using a single corpus and language, it must now be extended to support many of both, otherwise it will hold little interest for linguists outside of Biblical Hebrew.

This requirement is not difficult to achieve using any corpus database, whether flat files, Emdros or a relational database. Emdros has the notion of separate, independent databases within the same engine, which will be used to separate the corpora.

However, Lex also uses a database of its own, which stores the lexicon, parser rules and access permissions for each language, and this database must also be generalised to multiple instances.

The work of actually supporting multiple databases of either type in Lex has not yet been done.

Transliteration

Until recently, Lex was using a custom transliterator written in Java code, based on a specification for Biblical Hebrew by Winther-Nielsen as part of the RLM. Of course, this was completely specific to BH. In addition, its input was the old WIVU transliteration, which made it easier for non-Hebrew readers to understand the code, but at the cost of portability or relevance to other languages.

A new transliteration system has been designed and built and is under testing. The rules were developed by Winther-Nielsen and Tøndering [NWN], and consist of an ordered list of contextual rewrite rules. An excerpt of the list of rules is included below for illustration.

In database	Unicode	Name	Preceded by	Character	Followed by	Transliterate as
,	\u05bf	Rafe		,		Ignore
.	\u05bc	Dagesh		.		Ignore
/		Nominal indicator		/		Ignore
:	\u05b0	Sheva	^ cons	:		ə
				:	$	Ignore
			cons	:	cons : $	Ignore
			longvow cons	:		ə
			@ accent cons	:		ə
			consND .	:		ə
			consND	:		Ignore
:@	\u05b3	Hataph Qamets		:@		o

Table 3: New Biblical Hebrew transliteration rules (extract)

In the table above, the “in database” column is given in the archaic WIVU transliteration, which has been converted to Unicode (using the second column) for implementation of the transliterator.

Each line is a context-sensitive rule. It matches the wherever it sees “preceded by + character + followed by” (the concatenation of these three columns), and replaces the character in the “character” column with the character in the “transliterate as” column. *Ignore* means that the matched characters should be deleted (replaced by the empty string).

The special characters “^” and “$” match the start and the end of a word, respectively, and *cons*, *consND*, *longvow* and *accent* are regular expressions. They would be simple character classes, but sometimes they represent multiple characters in the input. For example, the value of *cons* is the following regular expression:

cons = ([אעבדגהיכלמנפקרסתטוחצז] | שׁ | שׂ)

For efficient implementation, each rule is converted into two compiled regular expressions, one matching against “preceded by” and the other matching against “character + followed by”.

To transliterate a string, we iterate a pointer over the gaps between characters, starting before the first. At each position, we then apply each rule in turn at that position until one matches. When that happens, we consume the characters in the “character” column and output the ones in the “transliterate as” column, and move the pointer forward by the number of characters consumed (at least one position). If no rule matches at a position, that is a transliteration failure, and the original character is output as a debugging aid.

Unlike the previous solution, these rules are perfectly suited to being stored in a database table, and so they are. This makes it possible to completely change the transliteration without any programming, so it is ideal for developing new transliterations.

In addition, a large number of test cases were developed for the rules. There is at least one test case for each rule, sometimes more. These test cases can also be stored in a table in the database. It will be possible to explore the effects of a transliteration rule change by showing the input, expected output and actual output for each test case, highlighting the ones where the actual output does not match the expected one, before making the change. This would facilitate testing new rules and improving the transliteration. It should also be possible to add any number of new test cases.

Finally, in some cases it may be difficult to see why the transliteration did not produce the expected result. In these cases, it would be useful to analyse the transliteration process step by step, to see which rule matched in each case and what the output was. This could be linked from the test case display, and also from other places where transliterations are shown.

Linguists working on complex transliterations could borrow an idea from software engineering called test-driven development [KB 02]. The principle is that tests (test cases with expected results) are used to verify correct functioning of the system at every stage. If the linguist wants to make a change, to correct a mis-transliteration, they would start by adding a new test case, with the expected result being the correct output. This would presumably fail (not match) under the old rules. They could then

edit the rules to make their test case pass (match), while making sure that the other test cases continue to pass as well.

Sharing of Linguistic Data

Related to the concept of multiple databases is the idea that several users might be working independently on a given corpus or language. They might then wish to exchange data with each other, such as the following:

- new or modified parser rules;
- new or modified lexicon entries;
- modified structural data in the Emdros database.

Lex should support the import and export of such data. It should be possible to export a subset of the entire corpus, parser rules or lexicon, and to import it on another computer or database, provided that it makes sense to do so (for example, the underlying language or corpus is the same). It should be possible to undo such operations and review changes. It should be possible to apply changes to existing objects rather than creating duplicates.

In addition, where multiple linguists collaborate on a single database, they should each be able to see the changes made by each other, as well as their own change history, and to undo changes if they are discovered to cause problems later.

Lex provides a foundation for import and export as well as change tracking and reversibility through its Database Change Tracking (DCT) layer. This layer sits between Lex itself and the underlying databases of Emdros and SQL. Although Emdros stores its data in a SQL database as well, Lex is not able to see this database or track changes to it directly. Instead, Lex treats the Emdros database as an opaque interface, and tracks changes to Emdros objects rather than the underlying SQL tables that Emdros uses for data storage.

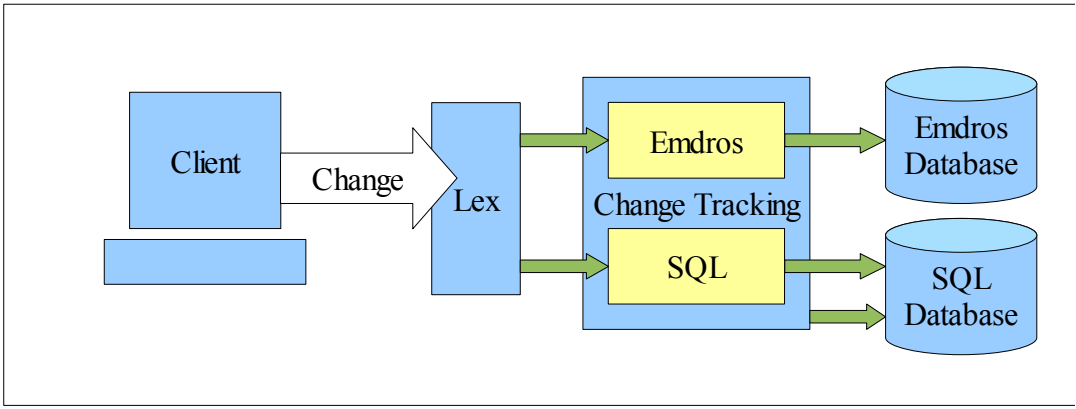

Figure 5: The Database Change Tracking layer (DCT) records changes to the Emdros and SQL databases in the SQL database

The DCT stores the following information about any change to a SQL database table or Emdros object:

- Logged in user name and IP address (that made the change)
- Current date and time
- The database type (either Emdros or SQL)
- The database name (for future expansion)
- The table name (SQL) or object type (Emdros)
- The command type (create object/row, update object/row, delete object/row)
- The unique ID (SQL primary key or Emdros object ID) of each object affected
- The old and new values of each field (SQL) or attribute (Emdros) changed

This data is stored in the same SQL database that Lex uses for the lexicon and for access control. Of course, the changes to the DCT tables are not themselves tracked. Also, changes to table and object structures (e.g. adding and removing columns or attributes) are not tracked, as these are made by administrators and not by end users of the software.

The storage of this data provides sufficient information to:

- view any changes made by any user to the databases;

- undo any change to the databases;
- export any change to be redone on a different set of databases;
- merge local and remote changes on import, and prompt for conflict resolution.

However, these features are not implemented in the user interface yet.

Access Control

Related to the concept of sharing data on the Internet is the need for users to have control over who can access their data. Many corpora are copyrighted, and researchers often wish to protect their work until it can be published.

Lex currently provides the following levels of access control:

- overall control using a password to prevent anonymous access to the system, if desired (currently using Apache Tomcat's built-in password protection)
- restricted access to corpus by user (whole database or sections of database)
- restricted access to corpus for anonymous users (whole database or sections of database)
- user interface to publish (grant anonymous read-only access) to individual clauses

The following additional controls are thought to be necessary or useful:

- user management interface and display of user rights (under development)
- administrative permissions to access this management interface
- separate control of read and write access to lexicon
- separate control of read and write access to parser rules

Access controls are easy to add, and will be added as required by users.

Morphological Rules

The importance of morphology to machine parsing and translation cannot be understated. Understanding and processing of morphology is required to:

- identify the primary syntactic argument (PSA);

- in the case of head marking structures, create the primary syntactic argument (PSA);
- provide semantic information such as tense that cannot be inferred otherwise;
- generalise the lexicon from surface to underlying semantic forms;
- produce grammatically correct output from machine translation in most languages.

Lex currently uses hard-coded morphology rules for Biblical Hebrew that use the information provided by the WIVU database to assist with morphological analysis. The reason for this is simply that the availability of this data makes the task of morphological analysis significantly easier and more reliable. However, unless it can be generalised, it will greatly increase the amount of work needed to build and maintain a lexicon for other languages in Lex.

One idea for generalisation of the lexicon is using expansion rules. Normal parser rules have at least one *non-terminal* symbol on the right, and just one *terminal* symbol on the left. The terminal symbol is placed above the non-terminals on the parse chart, and subsumes and consumes them, so the tree always becomes narrower as we approach the top, where there is a single root node that subsumes all of the others.

However, in morphological analysis, the opposite occurs. A single word, the normal unit of syntactic analysis, is broken up into multiple syntactic units with different functions. This causes an expansion which, if left unchecked, could carry on forever and result in parsing taking infinite time. Therefore, this is a potentially dangerous strategy. However, it may also be very productive, and provided that the same rule is never applied twice to the same word, infinite loops should be impossible as eventually all rules would be exhausted.

Let us take as an example the Hebrew word וַיַּהַרְגֵֽהוּ (way-ya-harəgē-hū[4]), “and he killed him”. This single word is a complete sentence in Hebrew. Here is an example of how the correct set of morphemes may be formed from the surface word, using morphological expansion rules.

Let us formulate some rules that allow us to split the word *wayyaharəgēhū* into six parts:

- the initial conjunction, *way*, which links this clause to the previous one (CONJ)

4 Genesis 04,08

- the tense aspect marker, *ya* (V_{TAM})
- the verbal stem, empty in this case, written as Ø (V_{STM})
- the verbal nucleus, *harəgē* (V_{NUC})
- the agreement clitic for the primary syntactic argument (PSA), empty in this case (AG_{PSA})
- the direct core argument head marker for the direct object pronoun, *hū* ($PRON_{DCA}$)

This diagram shows the morphology tree for the word *wayyaharəgēhū*. Above this tree, attached to the six top nodes, are the nodes of the usual syntactic parse tree, concentrating up to a CLAUSE at the top. Unlike that other tree, this one has its root at the bottom and expands upwards.

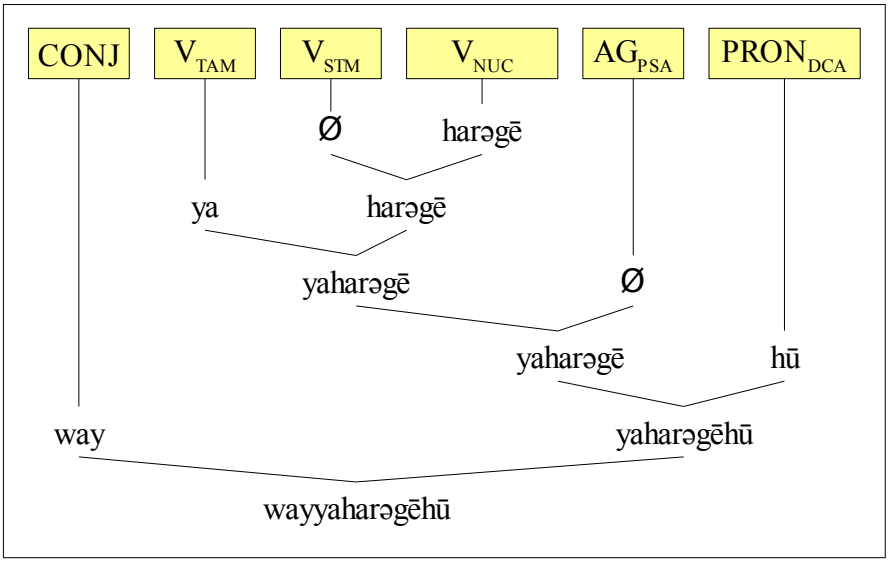

Figure 6: Morphology of the Hebrew word wayyaharəgēhū, "and he killed him"

Let us write our morphology rules like this, in the opposite shape of traditional parser rules, but in the same sense, as they represent expansion and not concentration of the single OBJECT on the right.

OBJECT1 [surface1] OBJECT2 [surface2] → OBJECT [surface]

This rule means that OBJECT is split into two objects (OBJECT1 and OBJECT2) whose surfaces are *surface1* and *surface2* respectively.

Let us shorten and generalise our rules by making use of the asterisk (*) as a wildcard character, that matches any number of surface characters (on the right-hand side), and [1] as a reference to the characters thus matched (on the left-hand side).

The following five rules work to expand *wayyaharəgēhū* into the six objects shown in the morphology tree:

- CONJ ["way"] V_{TSNAP}[5] [1] → WORD ["way*"]
- V_{TSNA} [1] $PRON_{DCA}$ ["hū"] → V_{TSNAP} ["*hū"]
- V_{TSN} [1] AG_{PSA} [Ø] → V_{TSNA} ["*"]
- V_{TAM} ["ya"] V_{SN} [1] → V_{TSN} ["ya*"]
- V_{STM} [Ø] V_{NUC} [1] → V_{SN}["*"]

These rules are not specific to the verb *harəgē*, and may apply successfully to other compound verbs. They may even apply in cases where they should not, especially the rules on empty AG_{PSA} and V_{STM} objects, which could equally well apply to objects that do have an agreement suffix or a verbal stem, as nothing prevents them from doing so. In such cases, it's important that the resulting misapplication of the rule can be identified and discarded by the end of the parse. One way to do this is to ensure that the lexicon contains verbal nuclei only. Then, false matches of such rules will produce a "verbal nucleus" that does not exist in the lexicon, and will be discarded as impossible to parse.

It is possible to include specific exceptions for whole words in the language which follow different rules, for example *broke → to break* and *saw → to see* in English. In decomposition and especially in composition, we may have many alternative matches: the system may generate the past tense of *to break* as **breaked* using an automatic general rule, that does not apply in this case, as well as generating broken using a specific rule for this exception. When we have multiple matches, we prefer the most specific, which is the one that directly matches the most characters, without counting wildcards.

Morphological rules as described above are not implemented yet. Only the hard-coded Hebrew morphological analysis is currently available.

5 I use V_{TSNAP} as a shorthand for V_{TAM} + V_{STM} + V_{NUC} + AG_{PSA} + $PRON_{DCA}$, but the name is arbitrary and only used within the context of the decomposition.

Parser Rules

Canonical RRG theory describes a “linking algorithm” that fits surface text into a complete tree (called a *template*) in a single step. However, the linking algorithm suffers from some problems in regard to computational linguistics:

- it requires significant intelligence and knowledge from the linguist, making it difficult to automate;
- it does not provide a way to specify or control how reordering of words can occur within a template;
- it is unproductive, in the sense that many similar templates sharing many common features will be needed for any given language;
- it does not facilitate the transfer of structure between languages.

Other authors have had some success with machine parsing in RRG theory, particularly Salem et al [YS] and Guest et al [EG 03, EG 04]. However their work was not easily accessible to be used in Lex at the time.

The best-known grammars which are easy to process computationally are Noam Chomsky’s *formal grammars*, which are widely used in computer programming. Chomsky later applied the same rules of formal languages to his controversial Universal Grammar, an attempt to model human languages in formal terms. RRG has rejected the Universal Grammar approach, and by extension formal grammars, as they are incapable of parsing languages with free word order without the use of transformations. They also tend to over-generate when used as generative grammars, potentially leading to infinite numbers of output templates, and structures which are never used by human speakers.

However, it is possible to solve both of these problems, and all of the limitations of template-based approaches, using a modified rule-based grammar which adds attributes, unification, and two types of permutations to traditional formal grammars:

- **attributes** are properties of symbols (terminal and non-terminal) which convey additional information such as the *tense* and *illocutionary force* of a verb or morpheme, without introducing new symbols that would break the generality of the grammar.
- **unification** allows attributes to propagate up the parse tree that would otherwise stay buried inside it. For example, the *tense* and *illocutionary force* of a verb or

morpheme can be propagated up to the CLAUSE, as required by the RRG operator projection. Unification also allows rules to place arbitrary restrictions on how they link to other rules, and hence allows any number of templates of any size and shape to be constructed, without loss of generality.

- **permuting rules**, the first kind of permutation, can find their terminals in any order or permutation, but they must adjoin each other. If the rule CORE → NUC ARG ARG is a permuting rule, then the NUC and ARGs may occur in any order.
- **searching rules**, the second kind of permutation, may find their terminals anywhere within the input, and in any order, without the requirement that they adjoin. Searching rules are required to parse Dyirbal, as in the rule NP → DET N, the matching determiner and noun can appear anywhere at all in the input.

We have developed and tested a rule-based active chart parser that implements all of these features, and can successfully parse Dyirbal free word order examples, and many artificial test cases. The parser is also reasonably fast for an active chart parser. Parsing 9 words against 20 rules takes under 4 milliseconds on a four-year-old laptop.

Lex also has a user interface designed to make it easy to add new rules by simply selecting a set of adjoining nodes. This does not yet support the above features, which must be added manually after the rule is created.

Experimenting with Parser Rules

Changing a single rule can have far-reaching consequences on the entire corpus:

- clauses that previously parsed successfully might no longer do so, or vice versa;
- clauses that previously parsed unambiguously might become ambiguous, or vice versa.

There is a need for Lex to be able to display a list of clauses whose parsing is influenced by a rule change, before that rule change is made permanent. This should help to examine the effects of a rule change and adjust the rule, or rethink the strategy, or recheck the affected clauses if necessary. Unfortunately, this is difficult to implement as it could require applying the parser to hundreds of thousands of clauses, which would be extremely slow.

Lex includes a parser debugger which shows all completed edges generated by the parser for a given input. It might be useful to have the option to display incomplete edges as well, or to drill down to certain types of edges.

As with morphology, the design of parser rules lends itself to test-driven development for quality assurance. The user could mark certain parse trees as “correct” for a given input, or alternatively enter a new parse tree or modify an existing one, to create test cases. The parser could be used to verify the parsing of all test cases on demand, much more quickly and easily than applying the parser to every clause in the database.

This approach could also be used to test theories about topic and focus structure, or other areas of active research in RRG and other grammars, by quickly applying new rules to a set of test cases and checking the outputs for correctness.

Lexicon Editor and Logical Structure Builder

The lexicon editor in Lex allows entries in the lexicon to be created, modified and deleted. Its most important function is to edit the logical structure (LS) of a lexicon entry, which is only relevant to verbs at present. Adverbs may have logical structures in future versions of Lex.

In other work, for example Salem 1, the lexicon is also used to store properties such as parts of speech and person, gender and number of words. So far, Lex has taken that information from the tagged WIVU corpus. Most corpora are not expected to contain this information, so we will need to store it in the lexicon as well.

The logical structures can be hand-written, or can be created using the logical structure builder, which follows the design from the RLM. The LS builder is a Javascript program that runs in the user’s browser, and assembles the logical structure using the answers to various questions which correspond to the grammatical tests on pages 93 and 106 of [VVLP].

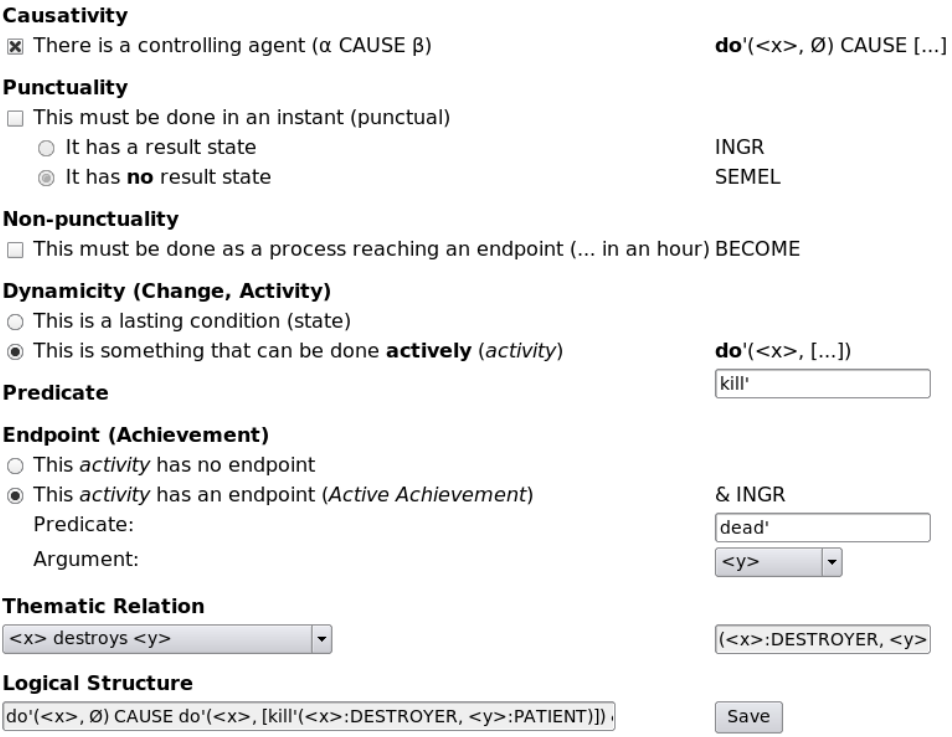

Figure 7: The Logical Structure Builder for the lexicon in Lex

The LS builder helps to avoid mistakes in the entry of the LS, and also to ensure that structures for similar verbs in different languages are represented in similar ways and with properties that can help to match them more easily and accurately across languages.

Not every possible logical structure can be built using the LS builder. For example, [VVLP] defines causative structures as “α CAUSE β, where α, β are LSs of any type” (p. 109). However, using the LS builder above, we cannot build the LS (1) or (2) below:

1. [**do**’(Tom, ø)] CAUSE [BECOME NOT **have**’(prisoner, knife)]

2. [**do**’(man, [**carve**’(man, log)])] CAUSE [BECOME **exist**’(canoe)]

Therefore, the logical structure builder needs more work to make it general enough to represent arbitrary logical structures, or even the more complex examples in VVLP.

Words may be organised into arbitrary tree-structured hierarchies. Any lexicon entry may be specified to have any other as its parent, as long as this does not cause a loop in the tree. Entries with no parent are displayed as roots. We originally imported the FLM verbal ontology described by Faber and Mairal [FM 99].

It could be very helpful for machine translation to use a common ontology across different languages, as this could help to suggest alternative verbs during translation when an exact match was not found. We hope to hear more from Elizabeth Guest and Brian Nolan on their progress on this front.

One point of note above is the Thematic Relation. Introduced at the request of Winther-Nielsen, this allows to assign more specific *actor* and *undergoer* roles to the various participants. This could be used together with coded attributes on nouns to indicate whether or not they could take various roles with various verbs, for example that a knife is more likely to cut than to be cut, and hence to be the actor rather than the undergoer of the verb *cut*. They can also be used to clarify or document the argument positions in a logical structure where they are not obvious.

The lexicon editor has suffered somewhat from an abundance of features. In Lex, the lexicon includes the following items, which appeared at the time to be of similar substance:

- logical structures for Hebrew verbs
- glosses for Hebrew verbs and other words
- categories from the FLM verbal ontology

There is an overlap between the lexicon and parser rules, which are currently held in separate tables. When the lexicon specifies that a particular word exists in the language, and is a verb, this is equivalent to the following parser rule existing:

VERB → Word ["yyaharəgēhū"]

However, if using morphological analysis as above, it may be preferable to store only verbal nuclei in the database, for example:

VERB → V_{NUC} ["harəgē"]

Chris Wilson. 2009. “Lex: a Software Project for Linguists.” *Technical Report SEE-J Hiphil* 6 [http://www.see-j.net/index.php/hiphil] (2009)

Neither of these is currently implemented

Connection of Corpus to Lexicon

The lexicon is only used to look up glosses and logical structures for words in the corpus. While the corpus includes inflected surface forms of words, the lexicon should contain lexemes without inflection, in order to capture linguistic generalisations.

For example, all forms of a particular verb (e.g. *run*, *runs*, *ran*) have the same logical structure, and many or all forms of a word (e.g. *book*, *books*) may have the same part of speech. The inflection of words is language-specific, so machine translation must undo the inflection when converting to semantic metalanguage, and apply appropriate new inflections when converting to syntax in the target language.

Our work so far has benefited from the inclusion in the WIVU corpus of a lexeme for each word, which is simply used directly to link the corpus to the lexicon. We hypothesise that, for any language:

- either rule-based morphological analysis, as described above, will decompose a verb or noun into several morphemes, one of which is the lexeme;
- or alternatively, that the same process can be applied with different rules to derive the lexeme.

It should be possible to specify which text to look up in the lexicon by means of a parser rule attribute, which is passed all the way up the parse tree to the top level. This attribute might originate as the surface attribute of the Word or V_{NUC} below the VERB node, and be copied up the tree by unification. This removes the need for the logical structure resolver to have domain-specific knowledge of RRG or the corpus in question.

Evaluation of Logical Structures

This is an area which currently requires domain-specific knowledge of RRG, as described above. The evaluator looks within the displayed clause to find a single Emdros [word] object, whose part of speech is a verb (according to the part_of_speech feature in Emdros, which comes from the WIVU database). If no verb is found, the empty string is used, as clauses with no verb are grammatical in Hebrew and imply equality or shared attributes of the subject and object, like the verb “to be” in English.

The found string (the name of the verb) is looked up in the lexicon (database table) to retrieve its logical structure (LS). The LS may contain any number of *variables*, which are enclosed in angle brackets, for example <*x*> and <*y*> (the names are arbitrary). These variables must be linked to referents in the syntactic structure, in other words to syntactic arguments.

Lex currently uses Emdros objects of type [phrase], with the phrase_type being one of NP, PP or a few other options, as possible arguments. However, since the development of the parser, the arguments could now be extracted from the core, using attributes and unification in parser rules.

The linkage between variables and arguments is specified manually by the user and saved in the Emdros database. If no values have been specified yet, then the subject (coded by WIVU) links by default to the variable *x*, and the direct object to *y*, assuming that the verb is more likely to be active than passive. Ideally that information would come from the parser, by using unification to pass morphological attributes up the parse tree.

Having resolved the values of the variables, we replace them with the (transliterated) text to produce the final, linked logical structure. For example, the general logical structure (1) of the verb *bārā?* (to create) in Hebrew may be linked in a particular case to create structure (2).

1. do'(<x>, [create'(<x>:CREATOR, <y>:CREATION)]) & INGR exist'(<y>)

2. do'(*?ĕlōhîm*, [create'(*?ĕlōhîm*:CREATOR, *?ēt ha- ššāmayim wə- ?ēt hā- ?ārec*:CREATION)]) & INGR exist'(*?ēt ha- ššāmayim wə- ?ēt hā- ?ārec*)

The values of the arguments are not translated here. In the case of common nouns, they could be looked up from the lexicon as glosses. The layered structure of the noun phrase, particularly conjunctions and adjectives, could be used to generalise further, and avoid the need to have phrases like “*?ēt ha- ššāmayim wə- ?ēt hā- ?ārec*” (the heavens and the earth) in the lexicon.

Parsing of Logical Structures

As noted by several authors, the logical structures, if completely implemented, constitute a semantic metalanguage which is independent of the syntax or expression of any human language. The process of converting syntax to semantics is reversible,

and therefore can be used for machine translation. The first step is to reverse the generation of the logical structure.

When adverbs are not taken into account, the outer form of the logical structure is entirely determined by the verb. Therefore, in the lexicon for the destination language, we must be able to find a verb with the same logical structure. Here it helps enormously to have consistent logical structures between languages, for example by using the LS builder described above.

Errors in the entry of logical structures, both syntactic errors such as missing brackets and semantic errors such as using the wrong predicate or verb classification, would cause a failure to find an exact match in the target language. In such cases, we may be able to prompt the user with approximate matches, derived from an ontology or by examining the settings entered into the LS builder, that may help them to identify and correct the problem.

In some cases, such as the different classification, and hence meaning, of the verb *to die* in Mandarin and English, an exact match may be impossible [VVLP p.106].

All the information from previous steps, including the mapping from variable names to arguments and the tree structure of the entire clause, is known from the previous steps, which took place as part of the same translation process. Therefore we do not need to rely on actually being able to parse all information from the logical structure itself (as a written string).

The arguments may be decomposed to separate nouns from adjectives which have lexicon entries in the target language, and render them according to the rules of the layered structure of the noun phrase (LSNP) in that language.

If the nouns and adjectives have direct matches or glosses in the lexicon of the target language, these can be used to translate them to native words. Otherwise, if there is enough similarity in the transliteration schemes and they are reversible, then a native phonetic writing of the foreign word is possible, at least as a place holder until a lexicon entry is created. Failing that, only a foreign word can be inserted.

Tøndering points out that Russian has no determiners, and therefore the meaning of many noun phrases is guaranteed to be ambiguous with relation to English, which requires determination. The most obvious solution is to allow rendering noun phrases as indeterminate when no determination was explicitly specified. This may result in

some very strange English (1), but does not claim anything that was not in the source language (2) (c.f. Gorbachev):

1. a crisis is caused by a problem with an economic system

2. the crisis is caused by the problem with the economic system

Having completed these steps, we should have a parse tree identical to the original one, but with the nodes below the core, the nucleus and argument nodes, linked to additional information about the choices of translations into the target language.

Rendering Syntax Trees

Before the syntax tree can be written out as text, it may need to be transformed or restructured, to take account of the syntactic differences between languages.

The canonical RRG approach would be to discard the old tree structure entirely and create a new one based on a template. While this works well for intelligent speakers who can see how a template can be adapted to fit their needs, we postulate that mechanically following this route will result in many cases where no suitable template can be found.

As we are using parser rules to generate templates, we may also have the opposite problem. If we are forced to enumerate all possible templates allowed by our rules, we may find that there are an infinite number, or a very large number, and it may take too long to check them all for suitability.

One alternative is to start at the root of the existing syntax tree. If the input parse was ambiguous, then we can use each possible parse tree in turn to generate multiple possible outputs. We start at the root of the tree, comparing the rule used at each level, to generate each node in the parse tree, to the ones in the target language rule set. If there is an exact match, we use it (for now). If not, then we try each possible rule for that node in turn.

For each rule, we try to fill all of its slots using the children of the old node. However, they may not be compatible, due to their type, or due to unification constraints. If we cannot find a place for a child node, we leave it orphaned. If we cannot find a child node to place in a slot, we leave it empty. Then we repeat the process for the child nodes, all the way down to the morphological level of the tree. At this point we proceed to the next filled slot, until all filled slots have been processed.

We may now have a number of orphans or holes left over. We assign as many orphans as possible directly to holes. Then we try adding rules to the left-over holes recursively, stopping and backtracking whenever we have more holes than orphans, or the depth exceeds a fixed limit.

If we still do not succeed in filling all the holes, we can try to strip the top node off each orphan in turn, and try again, until there are no nodes left to strip on any orphans. Stripping the top node may increase the number of orphans, and therefore the number of steps that the hole-filling algorithm can try before giving up.

If the holes in the tree ultimately cannot be filled, then it must be discarded for another alternative. However, it is possible to render a tree with some orphans left over, although it will not convey the full sense of the original input and is therefore not an ideal solution, if one can be found.

This algorithm allows, for example, the moving of argument from the core to the periphery, and vice versa, and the generation or removal of extra nodes required to adapt them to their new position.

We now have to calculate the morphology. As the rules of morphology will vary significantly between languages, there seems to be little point in trying to carry over the structure, so we recursively try every possible combination of morphology rules, provided that each word uses each rule no more than once.

If the transformation and regeneration process was successful, then we should have one or more complete tree from the top node (e.g. Sentence) down to morphemes. If we have multiple possible trees, they can be offered to the user as alternatives, or ranked by preference to disfavour those which discard information from the original, or which use infrequent grammatical constructs in the target language.

Note that this approach has not been tested, and requires further work.

Collaboration Tools

The construction of a lexicon and parser rule set for a single language is a significant undertaking, and even more so when multiple languages are required. Therefore it makes sense to envision the development of the tool as either a massive commercial project or a community collaborative effort.

The author favours community collaborative approaches to science. However the commercial possibilities of an improved machine translation tool should not be ignored, and the copyright ownership of the lexicons and parser rules should remain with the linguists who develop them.

Lex was designed as a web-based tool, in order to facilitate remote access and easy maintenance. It is possible to install it on a local computer, for offline use or for security reasons, if desired by the user, but upgrades may be complex and difficult to test remotely.

With the previously mentioned facilities for import and export of data, and sharing of corpora and rules between multiple users with independent databases and access control, Lex as a web service should be a more powerful collaborative tool than any software installed on a user’s own computer or local server.

Low Cost Software

It was important for the author to demonstrate a commitment to openness in science and research, by making the software available free of charge to all. Many academic projects have done the same, both with software, fonts, and the research that they have published. Cost should not be an obstacle in taking part in the project, as the contribution of each participant benefits all.

Open Source Code

In addition to being free of charge, the Lex software is released under an open source license, the GPL. This allows it to be modified by the user, and to be incorporated into other open-source projects. This license also applies to the Emdros database which we use. For those who would like to use the source code in commercial, closed source projects, alternative licensing for Lex and Emdros software can be discussed with the respective authors. This license does not apply to the databases, corpora, lexica and parser rules developed by users of the software, who are free to use them as they wish.

The open source license allows users to contribute to development of Lex, by customising it to meet their own needs and submitting patches for new functionality that can be shared with other users. It also means that Lex is not controlled by any one entity and the right to use and modify it cannot be taken away.

Chris Wilson. 2009. "Lex: a Software Project for Linguists." *Technical Report SEE-J Hiphil* 6 [http://www.see-j.net/index.php/hiphil] (2009)

Lex has extensive unit tests for most features, to help ensure reliability across new versions, and these will continue to be developed and improved alongside the code of Lex itself.

Lex uses and depends on other open source software, such as the MySQL database, the Java programming language, and the Apache Tomcat web server. Lex does not require any proprietary software to run.

Summary

We have described the software tool called Lex, its various features that may be of interest to linguists, especially in the RRG community, and our plans for future work based on Lex, including machine translation.

Bibliography

[NWN 08] Winther-Nielsen, Nicolai. A Role-Lexical Module (RLM) for Biblical Hebrew: A mapping tool for RRG and WordNet, 2008

[VVLP] Van Valin and LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge University Press

[USP 08] Sandborg-Petersen, Ulrik. 2008. *Annotated text databases in the context of the Kaj Munk Archive: One database model, one query language, and several applications*. Ph.D dissertation defended at the University of Aalborg June 27 2008. Available at http://www.hum.aau.dk/~ulrikp/PhD/.

[KB 02] Beck, Kent. 2002. *Test Driven Development: By Example*, Kent Beck, Addison-Wesley Longman. ISBN 0321146530, ISBN-13 978-0321146533.

[NWN] Winther-Nielsen, Nicolai. Forthcoming. *Biblical Hebrew parsing on display: The Role-Lexical Module (RLM) as a tool for Role and Reference Grammar*. Submitted to Hihpil (http://see-j.net/hiphil)

[YS] Salem Y, Hensman A, Nolan B. Implementing Arabic-to-English Machine Translation using the Role and Reference Grammar Linguistic Model. Private communications, manuscript forthcoming.

[EG 03] Guest E, Moburg L, Etchells J, Kailuweit R, Bender T, Hartung M, Staudinger E, Valet A. Parsing English, Swedish, and French using the

RRG paradigm. Presented at International Conference on Role and Reference Grammar, Brazil, 2003.

[EG 04] Guest E. Natural Language Parsing and RRG. Presented at International Conference on Role and Reference Grammar, Dublin, 2004.

[FM 99] Faber P, Mairal R. 1999. Constructing a Lexicon of English Verbs. Berlin: Mouton.