What have I done?

M. Randall Holmes

January 27, 2016

1 Introduction

Since 2013, I have been working on the project of developing the ability to parse Loglan from the level of letters upward. I wanted a parser whose function at every level was transparent, and which was easy to maintain. LIP did not meet either of these specifications.

Since I was working from the level of letters upward, and since our Sources are not altogether clear about matters of phonetics and lexicography (being usually though not quite always clear about grammar: the preparser is a problem for grammar as well), I had to make design decisions where matters were foggy, and in some cases I had to make changes.

My intentions have been conservative. I have not been interested in installing Novel Features. I have made material changes only when I felt that the state of things handed down to us was entirely unsatisfactory.

I now have quite strong evidence that I am basically parsing the same language that we inherited, mod a definite list of visible changes which I will describe. I have parsed all the words in the dictionary. The words in the dictionary parsed successfully with my algorithm with a very limited set of exceptions, most of which were incorrect in official 1989 Loglan terms.

It is not my intention in this document to describe all changes that I have made, but to describe those which have an effect on the parsing of existing texts. I have made technical changes, for example in the handling of the logical connectives on sentences of different precedences (ACI versus A versus AGE) and in the parsing of predicates with shared final arguments, which I believe have no effect on the parsing of any existing examples of Loglan text. I stand by these changes but I do not need to justify them here.

I parsed the NB3 corpus some time ago, but I made extensive changes in it, because it was already old in official 1989 Loglan terms. More instructively, I have recently parsed most of the text of Alex Leith's novel A First Visit to Loglandia, and I now have substantial evidence that my provisional parser parses the language as it was around the turn of the century after a limited set of characteristic modifications.

I am going to talk about these experiences and in the process explain what the problems were with 1989 Loglan as I found it and what the modifications are that I have made and hope that the Loglan community will accept. In particular, I am hoping to convince the Academy to accept the provisional parser and the remarks on it in the Fall 2015 Report [each providing security against bugs in the other] as the official definition of the relevant aspects of the language.

Because my intentions are conservative, I have not changed things that some speakers (including me!) may find Annoying about the language, as long as they worked. Things that I changed were real problems, in the sense that they compromised the intention that the language be unambiguous and speakable. I think that the language has various delightful eccentricities which we need to live with.

A laundry list of problems that I inherited:

the false name marker problem: There was a besetting problem of how to deal with occurrences of name marker words like **la** in names, having to do with how to recognize the left boundary of a name. This problem is solved.

serial names: There was a separate problem having to do with odd decisions of the Academy late in the last century to do with serial names like John Brown or Ivan the Terrible. These are solved.

acronyms: There was a serious problem with resolving acronyms. Here I made a major change in the language, moving acronyms wholesale from being predicates to being names, and imposing an additional marker on dimensions in dimensioned quantities. The problem was major enough and the acronyms a minor enough feature of the language that this was justified.

structure word breaks: There was a general problem with cmapua (little words): it seems unlovely to have pauses in a stream of little word

syllables seriously affect the parse (as in **le po sucmi ditca** versus **le, po sucmi ditca**). One major instance of this problem survives, because it is massively attested in the Visit and other Loglan texts, but a style fix is implemented for it (the issue of APA and IPA words). In general we stand by the view that one cannot pause in the middle of a word, and there are multisyllable cmapua in Loglan, but problems of this kind are contained. The **le, po** problem is solved.

opaque lexicography: The lexicography of cmapua in LIP is completely opaque and sometimes buggy, and there are demonstrably ambiguities in the language not detected in the course of disambiguating the grammar, because the lexicography was not part of the grammar. There are similar issues with the preparser (a major one was discovered in the course of parsing the Visit: inverse vocatives are demonstrably horribly ambiguous in the trial.85 grammar in a way hidden by the preparser). I have given formal definitions of all the word classes, and I believe I have detected and corrected ambiguities due to the lexer (for example, the existence of the **age** and **ige** series of connectives means that certain grammatical constructions cannot start with the word **ge**).

strong quotation: The strong quotation as defined in L1 is not parsable even with a BNF grammar. I came up with a different solution, which as it turns out is very similar to the solution already proposed to a similar problem with foreign names ("Linnaeans"). This solution works really well in the Visit. There are other small issues with quotation constructions and with any construction that brings alien text into the language.

foreign letters: The letters **qwx** were clearly already being phased out of the language at the end of the previous Academy's activity, and have now been eliminated from native Loglan text.

The list is not exhaustive. Other points may be brought up in the detailed discussion.

2 Orthography and Phonetics

2.1 Capitalization and punctuation

Capitalization and punctuation have been issues which I have actually addressed very recently, because the text of the Visit embodies some conventions that my parser did not support up to that time.

The letters **qwx** are forbidden except in alien text constructions (strong quotation, foreign names with **lao**, **sao** and **sue** predicate constructions).

Capitalization follows the usual conventions, except that names of letters may be capitalized when embedded in a word (this covers internal capitalizations of letters in acronyms and of letteral pronouns) and letters may be capitalized after a hyphen **-** (or a stress marker) representing a syllable break (as in the name **la Beibi-Djein** in the Visit).

Periods, question marks, exclamation points, colons and semicolons are terminal punctuation: they can appear at the ends of certain kinds of utterances (including utterances embedded in other utterances). They cannot be preceded by spaces, and the parser requires that if they are followed by anything it be spaces followed by a letter. Terminal punctuation may have a following inverse vocative attached (so that we can say things like **Tu he speni? hue mi.**).

Ellipses `...` and dashes `--` occurred in the Visit often enough that I added them as freemods, so they can occur in a wide range of contexts and be ignored.

Hyphens `-` are used for syllable breaks (on which more below). This means that other uses of hyphens suggested in L1 are abandoned. Hyphens supersede close commas to indicate unexpected breaks between vowels, as in **Lo-is**. The apostrophe `'` and the asterisk `*` are stress markers (the asterisk represents emphatic stress) – they can appear at the end of the syllable they stress (in place of the hyphen, not in addition to it) so they also mark syllable breaks, but they can also occur in final position to indicate a final stress.

The comma has a very important role in Loglan orthography, indicating a pause in speech, and in many places Loglan grammar requires that the comma be explicitly written. One particular issue which I emphasize as it occurs repeatedly in editing the Visit is that my parser **requires** that the comma be expressed at the end of a name word under most circumstances. Steve Rice and apparently Leith were lax about this. I don't require that all pauses be expressed, and noticeably I do not require that the equally

mandatory pause at the end of a foreign name be expressed by more than a space. The three major cases in which a pause in speech must occur according to L1 and NB3 are:

at the ends of names: just reviewed

before I and A words and related constructions: still required, of course.

between a finally stressed cmapua and a predicate: This can actually be implemented, since I provide explicit stresses, and this rule is enforced.

In each of these cases, a comma is required. In other cases where pauses are required, a comma is always permitted and at least a space must be shown.

There is an additional mandatory comma in my grammar: in certain descriptions ending with names, such as **la bilti, Djin** the name word must be preceded by a comma (Loglan pronunciation already required the pause) unless the name marker word **ci** is used. The explicit comma is needed as part of the solution to the false name marker problem. This is a change which I have had to make in a number of places in the Visit, so something that my reader should be aware of. These descriptions are intellectually perhaps related to serial names, but a serial name always begins with a name word. The description **le Sadji, Djan ci Blanu** combines this species of description with an actual serial name.

2.2 Phonetics

A major phonetic change in my provisional dialect is the elimination of **qwx** from native Loglan text. This has created a gap in the vocabulary, which I propose to fill by adding **kaiu, keiu, vaiu, veiu, haiu, heiu** as names for the letters **QqWwXx**.

Another change related to phonetics is the deprecation of the names for vowels. The vowel names of the forms **Vfi, Vma** are still supported, but we strongly suggest use of the new forms **ziV, ziVma**. Occurrences of the old vowel names in acronyms create interesting phonetic exceptions, but my parser does handle them, I believe. I am not inclined to banish the old vowel names, though I do prefer the new ones. I actually had to make parser corrections in relation to the old vowel names in parsing the Visit. The names

for Greek vowels are not supported at all; the Greek lower case consonants are still supported.

Most of my work on phonetics was not a matter of change so much as a matter of imposing precise definitions. In order to parse borrowed predicates, I needed a precise definition of the Loglan syllable (the units of complex predicates (djifoa) do not coincide with syllables). This does not appear anywhere in the materials, though statements supporting most of my specific decisions can be found in the Sources, and all the words in the dictionary do indeed parse. Indirect reasoning about the position of the stressed syllable in a Loglan predicate merely from the positions of spaces was very entertaining to implement in a PEG!

A change that I made which is almost always immaterial is that Loglan names must resolve into syllables just as predicates must. The only way in which this routinely causes one to need to rewrite names is that it caused me to require that constants **mnlr** used syllabically ("vocalically") must be doubled (it must be noted that this is already suggested in L1). This often causes spelling changes in names in the corpus and in the Visit. It also means that we expect names to be pronounced as written: names with foreign spellings should use **lao**. We should say **la Ainctain** and **lao Einstein**.

A minor point in building complexes which I raise because it came up in revising vocabulary in the Visit is that words like **riyhasgru** were corrected to **rihyhasgru**: a little word is not a djifoa, and all CVh djifoa (hitherto unused) are reserved to represent CV cmapua.

A feature which my parser has which I do not believe that any Loglan or Lojban parser has is that it allows one to write and parse genuine phonetic transcripts. You can write a text in a form in which no spaces occur except comma marked mandatory pauses. Try it (and tell me if you encounter bugs). Main stresses in predicates must be shown explicitly so that the parser can tell where predicates end.

I also note that in ordinary text explicit stress markers can be used to mark rhetorical stress (the parser will cough and die at devices like ALL CAPS).

An important point about explicit stresses is that they are always shown at the end of the syllable: there is existing Loglan text in which apostrophes are attached to the vowel to indicate stress; these have to be moved to the end of the stressed syllable. (the word **Espaniol'** in the dictionary had to be corrected from **Espanio'l**).

3 Lexicography

The definitions of certain classes of words are mysterious in the sources. The worst problems have to do with PA words and the related APA, IPA words and with NI words. The precise extent of these word classes was unclear.

I note a minor point because it is corrected at one place in the Visit. Allowing **noi**-initial logically compounded PA words demonstrably causes ambiguity at word boundaries, because all other uses of **noi** are word final. I changed words like **noipacena** to **panocena**.

Apart from that, my formal definition of the PA class allows all the words that LIP allows, and allows some more words that LIP does not allow.

The NI class is different. The word **mo** for three zeroes had other conflicting uses, and was replaced with **moa**. I added the construction which gives **temoato**, two million, **rimoate**, a few billion. The class has more internal structure. Words of the shape SARA are quantifiers, not predicates, respecting the fact that SA words are prefixes and should be applicable to RA words independently.

The biggest problem with the long word classes (PA, NI and classes with these as components) is that these word classes can be of arbitrary length. It is also the case, as a wise Loglanist pointed out to me, that one might want to pause in the middle of a long numeral without intending to break the word. Originally I had dedicated words for ending PA and NI words without pause, but I decided against this. Basically, pausing in the middle of a PA or NI "word" does not end it (except that the PA component of an APA or IPA word will end at a pause). A grammatical effect of this witnessed in the Visit is that one needs to be able to insert little words between NI words now and then: **ne ge tora**, "one set with two elements", because **ne, tora** would actually be the predicate "is a set with 12 elements". **Mi pa, vi blanu** is grammatically identical to **Mi pavi blanu**, whereas **Mi pa gu vi blanu** is not the same grammatically, though arguably it says the same thing.

In all other cases, I follow JCB's dictum in NB3: to pause in the middle of a word is to break it. Examples of constructions which really are words: **lemi** is a word. I do not allow **lemi blanu** to be written **le mi blanu**. LIP allows the latter form because it is generally ignoring spaces between cmapua. But notice that it does not allow **le, mi blanu**. It is useful to note that **mi** is **not** eligible to replace **la Djan** in the construction **le la Djan, hasfa**: the grammar of **lemi hasfa** is really truly dependent on **lemi** being a word of class LE.

The infamous **le, po** problem is solved in a quite different way. **lepo** is not a word at all. The words **po**, **pu**, **zo** are always long scope in my grammar. Short scope words **poi**, **puu**, **zoo** are provided. The grammar of **po** predicates and **lepo** abstract descriptions is significantly different under my parser than under LIP. The grammar of **po (sentence)** constructions is severely impaired in trial.85: I allow these predicates to be used fluently. To avoid double closure problems, the constructions (PO sentence (guo)) and (LE PO sentence (guo)) are different constructions (the second does not include an instance of the first!) The odd situation (LE ((PO sentence guo) predicate)) is handled by inserting ge: (LE (GE (PO sentence guo) predicate)) is handled correctly. LIP does not allow (PO sentence guo) predicates to occur as proper constituents of other predicates! In practical terms (relevant in the Visit) short scope occurrences of PO and its relatives need to be closed with GUO or the abstractors need to be replaced with the new short-scope words.

There **is** a general problem with existing Loglan text, which I am sure I will find examples of in the Visit, which is that certain constructions are hard to close, **guo** constructions in particular. Another danger is inverse vocatives: I have added the ability to close inverse vocative constructions from terms with **guu**, as otherwise there is a danger of (HUE argument) eating following text and becoming (HUE sentence). It is clear from usage in the Visit that (HUE sentence) is desirable as a construction.

The problem of the left boundaries of names is a lexicography problem. A name word must be preceded by a pause unless it is preceded by a name marker word, such as **la, hoi, hue, ci**. A name word contains a false name marker if it includes an occurrence of one of these name marker words and the tail of the name after the false name marker word is a well-formed name. The key to our solution is that a name word which includes a false name marker will only occur marked with a name marker word. Unmarked occurrences of name words have been avoided in the grammar, by outlawing unmarked vocatives and by requiring that name units in serial names which contain false name markers or which follow predunit components of serial names must be marked with **ci**. Further, name words in the name-final description construction of which **la bilti, Djin** is an example must be comma-marked (to remind us that we must pause) and must be marked with **ci** if the name contains a false name marker. In text, the false name marker problem is seldom an issue, because spaces tell us where the left boundary of the name word must be. Where a name marker word is used and is not actually serving

as a name marker, one should pause after the name marker or anywhere after a vowel before a name actually occurs, to avoid disaster, as in **la la, Djan, hasfa**, where the first pause is not mandatory in writing but must occur in speech to avoid saying "Ladjan is a house" rather than "John's house". Where a name marker word really is a name marker, it is unmistakable. The reason that we believe we have solved the problem is that unmarked occurrences of name words basically do not happen: they happen only in the name-final descriptions and in serial names, and both of these are guarded. We further refine matters so that if a name marker word is uttered and what it marks is not a name, you can secure the correct parse by explicitly pausing after the word: it will then try the parse as a name last, whereas if there is no pause after the name marker and the text following ends in a consonant, it will read it as a name. So **la, ladjan, hasfa** parses as "John's house", while **la ladjan, hasfa** parses as "Ladjan is a house". This does mean that the presumption is that a space after a name marker word is not a pause unless it is explicitly shown as such.

The problem of acronyms is a lexicographic problem. We eliminated acronymic predicates in favor of acronymic names, which makes sense both semantically and syntactically. Acronymic names are left marked by name marker words (they may not occur unmarked). They are right marked by pauses, like any other name. The other place where acronyms occur is as dimensions to quantities. Here we require a left marker **mue** and also a following pause. In this way acronyms cannot be confused with strings of letteral pronouns, which can be pronounced without annoying pauses. Multiletter prounouns are banished. The convenient utterance of a sequence of pronoun arguments without pause is far more important than the rare uses of acronyms or multiletter pronouns.

The problem of strong quotation is a lexicographic problem. It turns out that our solution to this problem has basically the same flavor as the previous solution to **lao** names. Quoted text starts with **lie**. If it has more than one block, these are separated by a little word **y** or **cii** (I used **cii** originally, but I allow use of the same **y** that is used to separate components of foreign names and this is what I use in the Visit.) All blocks are followed by pauses and **y** is flanked by pauses. It works nicely in the Visit.

The serious lexicographic problem which remains is the problem of APA and IPA words. I used to think that these needed to be closed with pauses: this is not the case. What is needed is that where an A or I connective (or even an APA or IPA connective) is followed by a modifier, a pause must

intervene. I have a style fix, which is that I allow any A or I class connective to be terminated with GU; if we use these versions of the connectives, no pauses are needed. Existing text, such as the Visit, will parse if the correct comma pauses are added, and in fact Leith usually puts in the required pauses – but sometimes misses them. The same is true of JCB in older text. The Founders were aware of this issue.

4 Grammar

A global grammatical change which has an effect on how old text will be parsed is the complete abandonment of pause/GU equivalence. I am convinced that the use of pauses to replace closure words cannot be managed consistently. It is the case that when I **did** have pause/GU equivalence, it always worked quite differently from JCB's notions of how it should work. There are multiple places in the Visit where grammatical corrections of closures have to be made which arguably arise from Leith relying on pause/GU equivalence, though they can equally well be construed as errors. There are a couple of places where the **Le mrenu, sadji** "sentences" which pause/GU equivalence allows have to be corrected in the Visit.

The serial name problem is a grammar problem, and it is solved. Components of serial names which are predunits must be marked with **ci** (as was already required for name words in serial names which contained false name markers) [a very subtle point is that a name-final predunit component should be marked with **ci** followed by a pause: **La Alis, ci, cluva je la Djan**; the pause is not mandatory in writing but definitely would be wanted in speech]. This means that **La Djan, Blanu** means "John is blue", while **La Djan ci Blanu** describes "John the Blue". We also require name components following predunit components to be marked with **ci**, eliminating a fertile source of unmarked name words in uncontrolled locations.

Other actual changes in grammar have very little effect on existing text in 1989 Loglan. This is as we should expect. However, I do point out one area in which changes were made.

I require in my provisional parser that in a construction **PA/ga predicate (terms) ga terms** that the initial terms moved to the end include either exactly one argument or all the arguments. The effects of moving two or more initial arguments to the end are perverse, because they cause a massive shift in the apparent meaning of the sentence when the final arguments

are read. A related issue is that a sentence in which the only terms which appear before the (unmarked) predicate are modifiers is parsed as an imperative (which *does* make an actual grammatical difference), and a sentence with a marked predicate with no argument before it is read as a declarative gasent. (**Na crina** says that it is raining, not "Be a raindrop!"). The issues which arise from closure of inverse vocatives strongly tempted me to forbid VOS sentence order (more than one argument before the predicate), as the unintended sentences arising from failures to close inverse vocatives were usually of this form. But I also think that VOS sentence order has actual potential applications, so I resisted such a modification.

5 What actually happens in parsing old text

All words in my working dictionary (code RDC) parse. When I tested the vocabulary, few words had to be corrected, and all were arguably incorrect in 1989 Loglan terms already, except where punctuation conventions have changed.

It is worth noting that at the same time that I corrected the incorrect words in the dictionary, I added the new cmapua which I have defined. There are very few of these: there is only one (the left marker **mue** for acronyms) which has an essential grammatical function.

The changes made in the Visit are of stereotyped kinds.

syllabic consonants: Some Loglan names have to be corrected by doubling consonants used syllabically.

foreign names: Some names containing foreign letters are changed to foreign names marked with **lao**.

close commas: Close commas convert to hyphens.

strong quotations: Strong quotations are changed to the new format.

inverse vocative closures: forms **hue (argument)** have to be closed with **guu** fairly often. The same grammatical issue existed in 1989 Loglan if the trial.85 grammar is taken seriously: the preparser obscures this. 1989 Loglan did not have a device to effect the closure well: what I did is change the rule (HUE terms) to (HUE termset) in my grammar.

omitted commas after Loglan names: Leith sometimes missed these.

fixes to the interiors of serial names and name-final descriptions: Components of serial names marked with **ci** where required; explicit commas inserted before final names in name-final descriptions.

APA/IPA closures: Where an A(PA) or I(PA) connective is followed by a modifier (PA word initial), a comma must be inserted. This happens with varying frequency: Leith usually supplies these commas.

closures with guo or replacement of po with poi: Closures of event and other abstractions have to be watched carefully, especially as the effect is often an unintended parse rather than parse failure (this is also true where inverse vocatives eat following sentences). This is well-known to be an issue in writing Loglan text: it is especially an issue here because the closure rules have been changed due to abandonment of pause/GU equivalence.

sporadic bad predicates: Leith sometimes invented words without proper attention to the rules. This is likely to happen in future text!

NI words: In just a couple of places, things like **ne, tori** had to be corrected to **ne ge tori**. This was an interesting discovery.

noi- initial composite PA words: Words like **noipacena** corrected to **panocena** in a couple of places.

It does not take me terribly long to set up an installment of Leith to parse. I am convinced by my experience doing this that Leith and I are working on the same language.