BRITISH & FOREIGN BIBLE SOCIETY

# Towards a Computable Understanding of Word Formation in Natural Language

*or*
*Lost in the bean field...*

*J D Riding*
*Linguistic Computing*
*at*
*British & Foreign Bible Society*

**Abstract**

This paper proposes the integration of what are currently seen as discreet computer processes for analysing word formation within a single descriptive schema. It attempts to identify fundamental behaviours that are at work in word-formation and which have a generic application across all word-based languages. It does not describe in detail any particular solutions to particular problems of analysis but seeks rather to propose a framework for analysis within which individual systems might be developed in such a way as to contribute to a broad general solution. The emphasis is upon language independent systems which have little or no need of knowledge bases except where those knowledge bases can be demonstrated to have a universal application to the vast majority of natural language. It also seeks to identify processes used in word formation analysis that have a wider application in linguistic analysis beyond word formation itself.

## *Computers and language*

There is a general recognition that computers are becoming more effective at dealing with complex problems, including translating between natural languages. Google translate can give really good results for a handful of commercial *lingua franca* and other knowledge based systems are becoming increasingly competent. These are all good things but in the context of Bible translation they bring fewer benefits, largely because most Bible translation takes place in languages with a much lower global profile.

In the context of Bible translation knowledge bases are less likely to be available, it is certainly unwise to assume that they will exist in a usable form for any given language. This drives those of us working to apply MT within the Bible translation community towards the need for automatic analysis in place of human generated knowledge.

So far so good but the increasing competency of translation systems in some fields can also mask a fundamental fact about the nature of computers. In the end, all computers do is count beans. For beans read: items to be processed (counted). In the context of language these items might be graphemes, characters, phonemes, syllables, morphemes, words, phrases, sentences, paragraphs or an host of other linguistic entities. The great thing about counting beans is that whilst it is not in any way clever, computers can do it blindingly quickly. Enormous datasets can be examined in a few seconds for patterns that might suggest a particular semantic or behaviour. This too is a good thing but the whole edifice of analysis is dependent upon a single fundamental ability. In order to count beans you must first be able to find them.

The computer, being fundamentally a simple soul, really wants the beans presented for counting in a neat row with each bean clearly identified as distinct from its neighbours. Like a row of peas in pod. The snag with natural language is that the beans are more likely to come like a bunch of grapes. Just working out which bean to begin with is often not easy. This problem lies at the heart of many of the difficulties faced by any process

which attempts to analyse the structure of word-formation in natural language.

## Problems posed by word-formation issues

If you are a computer faced with trying to disentangle the components of words there are some particular problems which you must be able to solve. Much of your dataset may well look more like a bunch of grapes than a row of peas in a pod. Finding patterns in a bunch of grapes may well require some sophisticated processing which has the ability to identify the relations between components where those components do not appear to be in linear sets. Often linear relationships can be established where a language generates surface forms largely by component agglutination. Finding patterns representative of such concatenative morphology is relatively straightforward and such processing is now core to the Paratext glossing process. For languages which form words in more complex ways it is far more difficult to identify patterns and the relationships between individual components. Such non-concatenative patterning requires a different approach. Interestingly, even in languages where words are formed by component concatenation, phrase level analysis is likely to require non-concatenative pattern recognition. This suggests the problem is fundamental to automatic linguistic analysis and therefore symptomatic of natural language as a whole. If it is indeed the case that this is a fundamental characteristic of natural language this also offers the prospect that a solution in one context may well be extensible to other contexts of analysis.

The second problem faced by the computer is the confusion generated by the process of encoding what is fundamentally an utterance (stream of sound) into a written orthography. Spoken language forms words by reference to at least two distinct schemata: morphological rules for word construction and phonetics. The former is fundamentally semantic, the latter is generally semantically agnostic and driven by the needs of speech articulation. It is phonetics that requires the Englishman to speak of *an idea* rather than *a idea,* just as phonetics forced the Koine Greek speaker to articulate *οὐκ οἰδαμεν* rather than *οὐ οἰδαμεν*.

## The principle formation schemata

Once the vocabulary for a semantic has been selected the correct surface form for the context must be constructed. In the context of Paratext I identify at least three influences on word formation: morphology, phonology and orthography. The interaction between these schemata results in the surface forms generated and recognised by the speaker. At present the human cognitive process of generating a particular surface form is unclear although studies in language acquisition in small children are beginning to offer a better understanding of how a child forms words. Most of the progress in this context is the result of analysis of template structures that emerge in the first months of language acquisition and from the analysis of common errors in forming words. For the purposes of this discussion, and at the risk of over-simplification, the following hypothesis is offered:

1. The logical construction of a surface form is in the first instance driven by semantics. The speaker first identifies the lemma that most closely represents the required semantic. In the earliest stages of language acquisition the lemma is simply reproduced in what is in effect a citation form which general corresponds to the form of the word most commonly encountered.

2. The function and context of the lemma within the utterance is considered next. Thus an English speaker having selected the pronoun *who* next considers the function or role it has in the phrase. If the role is subjective no modification takes place, if it is objective a prefix *m* is added to generate the surface form *whom*. The fact that more often than not this step is omitted in modern English speech lends weight to the hypothesis that this part of the process of word-formation is

cognitively speaking discreet from lemma selection.[1]
Similar simple transformations take place in negative forms such as *amoral, incompetent, uncaring, impossible* etc... To what degree the mother-tongue speaker generates the negative form as a one or two step process is less clear. It seems likely that familiarity of use will in time result in these two steps becoming one by the inclusion of what is in effect a negated citation form of the lemma the speaker's lexicon. Nevertheless a language speaker's ability to synthesise new forms by applying a generic transformation template to new vocabulary is clear and well-documented.

3. Once a surface form that carries the necessary core semantic has been selected and the semantic modifications required by the context have been applied the third part of the process takes place. In our example above all of the negated forms are created by affixing an alpha-privative before the lemma. The variety of routes by which English has acquired its vocabulary account for the change of vowels (*a-, in-, un-*) but the (*in-, im-*) shift is a consequence of English speech patterns. The form *inpossible* is entirely plausible morphologically but the juxtaposition of the *n* and *p* phoneme feels awkward to most English speakers. The consequent is a morpho-phonological mutation in which *n* shifts to *m*.[2]

4. In some circumstances orthographical mutation takes place. This often happens where a lemma, component or construction has been borrowed from another language. Morphological analysis alone will hypothesise the existence of two Latin lemmata, *dic-* and *dix-*. Semantics tells us that these must be two forms of the same lemma but to the bean counting computer this is not at all clear. In this particular case the lemma is in fact *dic-* (citation form *dico*)[3] and the perfective form of the verb is formed by a technique borrowed from ancient Greek[4] . An *s* is inserted between stem and suffix to make the perfective form, thus: *dics-*. Both Latin and Greek orthography rewrites *cs* as *x* hence the 'new' stem *dix-*.

A computer faced with disentangling even this fairly simple set of transformations may well struggle unless it is able identify the various steps which may have taken place and then disassemble the resultant surface form into its components in orthographical, phonological, morphological and semantic contexts. In some cases it will fail, effectively falling back upon the grammarian's assertion: 'it's an irregular form', which is roughly analogous to the archaeologist's standard explanation for otherwise unexplained evidence: 'its for ritual purposes'. In fact, many of the apparent irregularities encountered in word-formation are in reality the outcome of the interaction of the processes detailed above. To give the computer the best chance of dealing with such irregularities we must provide it with the best tools possible to enable it to identify the transformations required by each schema and then to recognise the complex interactions that take place between them. In other words, we must help it find the beans.

## *Resolving conflicts between formation schemata*

If the computer is to be able to handle the interactions between the different schemata it is important that it begins with a clear understanding of the individual schemata. Once that is in place the business of mapping mutations across schema boundaries can be attempted. The first pre-requisite for accurate word-formation analysis is, therefore, a clear understanding of how the morphology, phonology and orthography of a language

1 It is clear that a child acquiring early linguistic skills would be unable to describe the process of lemma selection and transformation in these terms. Nevertheless, the evidence is seems strong that something of this sort is taking place whether it is described in structural linguistic terms [Harris, 1970] or in terms of cognitive development or socio-linguistics [Tomasello, 2003].
2 For a set of formal axiomata describing phonemic transformations see: [Bloomfield, 1961]. For a general discussion of the fundamentals of voiced utterance see: [Ladefoged, 2005].
3 *dico, dicere,* speak, 3rd conj.
4 first, weak, -σ-, or sigmatic aorist

works. This can be achieved by understanding first how syllables are formed by the language from which the individual phonemes and permitted phoneme clusters can be identified. Traditionally this is attempted by analysis of the sound stream which represents a speech utterance in the language. The sounds identified are then mapped to the preferred orthography. In the context of Paratext this must be done by analysis of the written language. Automatic syllable analysis is not trivial but neither is it impossible. Paratext already has sub-systems in place which can make a reasonable fist of identifying syllable boundaries in words.[5] This information is already used to help set hyphenation points but if the hypothesis above is correct we should be able to apply it to the problem of resolving apparent irregularities in word-formation. A reasonably comprehensive syllable analysis should help us identify valid phoneme clusters and thus predict the potential for morpho-phonological mutations in generated surface-forms.

Combining a syllable construction schema with an analysis of the morphology of a language offers the prospect of a language independent method for predicting or validating surface forms in text. Paratext already has in place an automatic morphology analyser which handles most morphologies sufficiently well to enable inflection paradigms to be constructed for stem lemmata. At present the Paratext morphology analyser is part of the glossing engine but it would be possible to develop a stand alone morphology analyser which was not dependent upon glossing in any way.[6] The addition of a morpho-phonological element of analysis should improve this still further. More work is needed to provide the capability to handle non-concatenative morphologies but the fundamental process of identifying and validating inflection paradigms is known.[7]

## *Possible outcomes for Paratext*

A better understanding of word-formation offers a number of benefits:

- Paratext has already demonstrated the benefits of parsing a word list both syllabically and morphologically. If we are able to put these analyses together into a coherent method for word-formation analysis we have the beginnings of a method for validating and perhaps synthesising surface forms.
- This in turn offers the prospect of intelligent spell-checking, perhaps even as the user types, based upon an analysis of the way words are formed in the text as a whole. Words entered by the user that appear to be in conflict with the general word formation rules for the language may well include typographical errors. Equally, they may be words borrowed from another language in which case the word formation rules for the target language will not necessarily apply, proper-names are the classic example of this.
- At present semantic mapping in Paratext is less effective at sub-word level. Better word component analysis should also enable the semantics of morphemes to be identified more easily.
- Improved hyphenation tables are a real prospect based on an integrated model of word formation including both syllabic and morphological schemata,
- as is a method for proper-name transliteration based not upon character transliteration but on valid phoneme and phoneme cluster transliteration, where the phoneme analysis is derived automatically by the system.

5 We already have to hand systems with the capability to give a useful analysis of syllable structure, see [Riding, 2005] and Brad Olson's syllable parser for Paratext.
6 There are many examples of stand-alone MDL based morphology analysers e.g. [Goldsmith, 2001].
7 MDL analysis cannot be extended to handle non-concatenative morphologies but relational methods do have this capability, e.g. [Riding, 2007].

## Current work in related areas at BFBS

In addition to the huge strides forward made by Paratext in these areas work is also in hand at BFBS into particular phenomena which are fundamental to the kind of analysis described above. The single biggest problem is that of non-concatenative components. Languages which form surface forms largely by agglutination present relatively straightforward problems for the computer. Languages which form surface forms with the use of many infixes, typically but not exclusively vowel mutation, are far more difficult to handle. The same problem applies to phrase structure analysis. Where the components of individual noun or verb phrases appear coincidently in the text stream it is much easier to recognise the relationship between adjacent components. For languages that routinely place components which are closely related to one another in apparently disjoint locations in a phrase it is far more difficult to identify the relationships between those components. In each case the fundamental problem is the need to accurately identify non-concatenative patterns in the text stream.

BFBS has been trying to progress this work for some time now. A prototype machine with this capability exists and is giving encouraging results. More work is needed to demonstrate that the process is indeed a global solution and to broaden the application from word-formation analysis towards phrase structure analysis.

Work is currently in hand on addressing the problems of proper-name transliteration. We now have two UK universities who are interested in working in this area with us and are hopeful that the next year will see useful progress in the form of a working prototype system for character-based transliteration.[8] Further research continues exploring phoneme and syllable component based transliteration models.

## Summary

The capabilities proposed here are unlikely to be ready in anything approaching a usable form in PT 7.x. It seems more likely that we are speculating about what PT8 and beyond may be able to do. Nevertheless, the processes discussed are fundamental building blocks towards language independent solutions. Such solutions are not knowledge base dependent and offer the possibility not only of general application to most natural language but also the prospect of making progress towards a generic and computable model for natural language as a whole.

An integrated system for word formation analysis such as has been proposed may be a little way off but as we develop systems to address individual problems we should do so with at least half an eye fixed on the bigger picture. There are already ways of doing most of the things discussed above but they are in the main specific solutions to particular problems. If we can construct a development framework which would allow individual sub-systems to contribute to a broader analysis and, hopefully, a better understanding of how natural language really works, the benefits in the longer term are enormous.

J D Riding, BFBS LC
20th September 2010

8 BFBS LC Research Proposal – PToleMy, a Proper-name Transliteration Machine, 2010. For more on proper-name transliteration issues see: [Bailey, 2007].

BRITISH & FOREIGN BIBLE SOCIETY

## *References:*


Bailey, N. (2007), 'Proper Names in the Bible: translation and transliteration issues', *Word & Deed*. SIL.

Bloomfield, L.Saporta, S., ed. (1961), *Approaches to the Study of Language*, Holt Reinhart Winston.

Harris, Z. (1970), *Papers in Structural and Transformational Linguistics*, D Reidel, Dordrecht.

Ladefoged, P. (2005), *Vowels and Consonants*, Blackwell Publishing, Oxford.

Raimy, Eric (2007) *Precedence theory, root and template morphology, priming effects and the structure of the lexicon,* University of Wisconsin, Madison, CUNY Phonology Symposium Precedence Conference NYC, NY.

Riding, J. (2007), *A relational method for the automatic analysis of highly-inflectional agglutinative morphologies*, MPhil, Oxford Brookes University.

Riding, J. D. (2005), *First Experiments in Automatic Hyphenation*, Technical report, British & Foreign Bible Society.

Tomasello, M. (2003), *Constructing a Language*, Harvard University Press, Cambridge Mass..