Representing Unish Grammars Based on Tree Adjoining Grammar Formalisms

Purev Jaimai & Hyun Seok Park*
National University of Mongolia & Ewha Womans University

Abstract

To make any sort of optimality argument, or rational engineering decision, one needs a fairly precise understanding of the problem to be tackled. Thus the purpose of this paper is to formalize Unish grammar by developing prototype Unish Tree Adjoining Grammars in the hope that developing and formalizing Unish grammars at this stage will help to direct the way the future version of Unish should be tailored and modified, as an artificial language.

Keywords: Tree Adjoining Grammars, universal language, artificial language

---

1. Introduction

Since Chomsky first proposed formal language theory in 1950's, various researches based on the Chomsky's hierarchy to generate the corresponding languages (both natural and artificial languages) have been exercised. However, the differences between the characteristics of natural and artificial languages are profound. First of all, natural language existed for thousands of years. Nobody yet understands how the language was developed or designed in the first place. But artificial languages are synthesized by logicians and linguists to meet some specific design criteria. The most basic characteristics of the distinction would be the fact that an artificial language can be fully circumscribed and studied in its entirety before it is practically used. During the developing period, the language can be tailored or modified constantly by the linguist's or logician's reasoning to best fit the needs of the new language.

A language design includes many interacting elements such as phonemic inventory, morphology, syntax, semantics, pragmatics and the culture of the society that might use the language. In designing an artificial language such as Unish,[1] many scholars in various fields have made notable contributions including a wide spectrum of studies (Chung 1996, Kim 2001, Lee 2002). Beyond just observing a phenomenon, linguists and logicians should be able to formalize it, or give a reason to prove that their observation is right.

With that in mind, I tried to develop prototype Unish Grammars based on a formalism called Tree Adjoining Grammars. In section 2,

---

[1]   Sejong University has been developing a new universal language called *Unish* (Choo et al. 2000, Diamond 1996) for several years. Unish, which represents a universal language, is an efficient composition of 16 languages. Unish is characterized by its regular grammar and simple pronunciation. To date, Unish has a vocabulary of nearly 10,000 words developed through careful word selection. Its vocabulary is still growing, and the grammar is constantly updated to best suit the purpose of an international language.

I briefly introduce the formalism of Tree Adjoining Grammars, which was first developed by Joshi, Levy, and Takahashi (Joshi et al. 1975). In section 3, I introduce some of the basic Unish Lexicalized Tree Adjoining Grammars. In section 4, a parsing example will be shown for a Unish passive sentence construction, to help readers to understand the operations (substitutions and adjunction) in TAG formalism. In section 5, some exemplary cases of more complex structures will be discussed, including the system and use of relative pronouns and wh-words. Finally in section 6, an example of NP agreement case will be given to show how 'feature' systems have been reduced in Unish due to simplicity of grammar. In many places throughout the paper, evidences have been shown to prove that Unish is simpler, more logical, and more regular in many aspects than natural language.

2. The Formalism of TAGs

 The properties of Tree Adjoining Grammar (TAGs) permit us to encapsulate diverse syntactic phenomena such as unbounded dependencies in a natural way. A Tree-Adjoining Grammar consists of a quintuple (Σ, N, I, A, S), where Σ is a finite set of terminal symbols, NT is a finite set of non-terminal symbols, S is a distinguished non-terminal symbol, I is a finite set of finite trees, called initial trees, and A is a finite set of finite trees, called auxiliary trees. Yves Schabes, Anne Abeille, and Aravind Joshi extended Tree Adjoining Grammars to include lexicalization. Lexicalized grammars systematically associate each elementary structure with a lexical anchor. The grammar consists of a lexicon where each lexical item is associated with a finite number of structures for which that item is the anchor, denoted with the diamond symbol next to the node name (as shown in Figure 1).
    A TAG is a tree-rewriting system and TAGs generate phrase-

structure trees. There are no separate grammar rules, although there are combining rules for combining these structures, i.e., adjunction and substitution (see section 4 for further details).
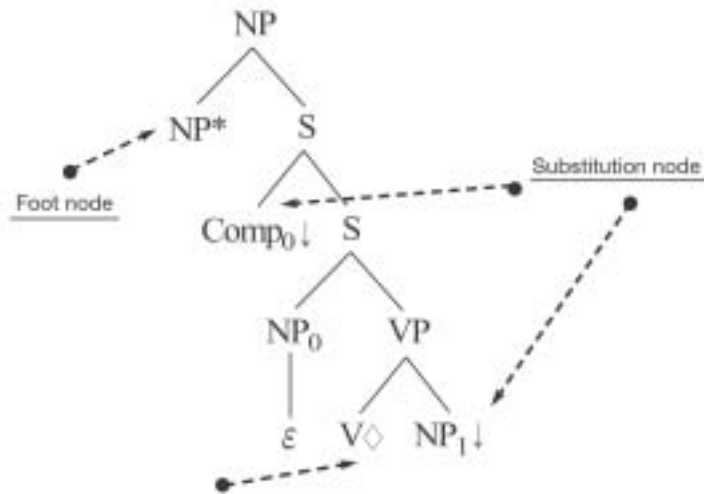

Figure 1. Substitution node, foot node and anchor node

There are two kinds of elementary trees in TAGs: initial trees and auxiliary trees. In describing natural language, initial trees are minimal linguistic structures that contain no recursion. In initial trees, all internal nodes are labeled by non-terminals, and all leaf nodes are labeled by terminals or by non-terminal nodes marked for substitution.

Recursive structures are represented by auxiliary trees, which represent constituents that are adjuncts to basic structures. In auxiliary trees, all internal nodes are labeled by non-terminals and all leaf nodes are labeled by terminals or by non-terminal nodes marked for substitution, except for exactly one non-terminal node, called the foot node. The foot node has the same label as the root node of the tree.

A down arrow ( ) is used with nodes to mark a substitution node, and an asterisk (*) is used with nodes to mark a foot node.

3. Basic Structures

In English, the basic word order of a sentence is a subject (S), a verb (V), and an object (O). In an interrogative sentence a special word appears in the first position of the sentence.[2] In a *wh-question*, which is another type of an interrogative sentence, the word that corresponds to a *wh-word* in English (namely, *who*, *when*, *where*, *what*, *why* or *how*) must appear in the initial position. On the other hand, in Unish the word order of a sentence (SVO) is always kept regardless of a declarative sentence or an interrogative sentence (Lee 2002).

(1) a. De   ver-ed   tori.
      you   see-pst   bird
      'You saw a bird.'
   b. De   ver-ed   tori?
      you   see-pst   bird
      'Did you see a bird?'

The word order of the two sentences in (1) is the same. The only difference between them is that (1a) ends with a period (.), whereas 1b) ends with a question mark (?).

Figure 2(a) shows the TAG tree structure for intransitive verbs. NP_0 is the place where a subject will be combined by substitution. Figure 2(b) shows a transitive verb structure, where NP_0 is the place where a subject will be substituted, and NP_1 will be the place where an object

---

[2] The following abbreviations are used in the gloss:

| acc: accusative | adjl: adjectival | advsuf: adverbial suffix |
| gen: genitive | pres: present | pst: past |

phrase will be substituted. Word order of SVO is reflected in Figure 2(b). Figure 2(c) shows a simple auxiliary tree for an adjective structure.

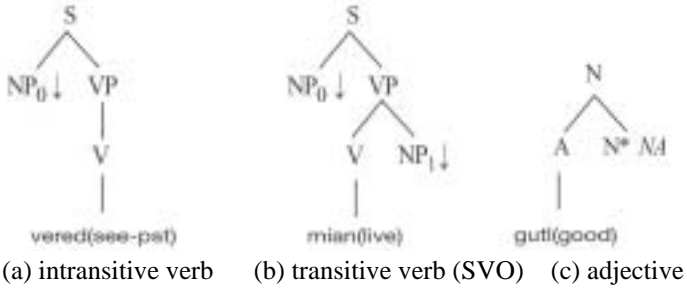(a) intransitive verb   (b) transitive verb (SVO)   (c) adjective
Figure 2. Basis trees for Unish

In Unish, unlike English, only one prefix is related to the form of a main verb in making a passive sentence as illustrated in (2).

(2) a. Me   skrib-ed   buk.
         I   write-pst   book
         'I wrote a book.'
     b. Buk   be-skrib-ed   be   me.
         book   psspref-write-pst   by   I
         'The book was written by me.'

The sentences in (2b) are the passive counterparts of (2a). The only difference between an active sentence and its passive counterpart in the form of a main verb is that the prefix *be* is attached to the main verb in a passive sentence. Therefore, it is much simpler and easier to make a passive sentence that corresponds to its active sentence in Unish.

Elementary trees to express sentence 2(b) are shown in Figure 3. Figure 3(a), 3(b), and 3(d) are initial trees. By convention, initial trees are called alpha (α) trees. Figure 3(c) is an auxiliary tree. By convention, auxiliary trees are sometimes called beta (β) trees. In

the next section, how these trees are combined to form a final tree will be explained.
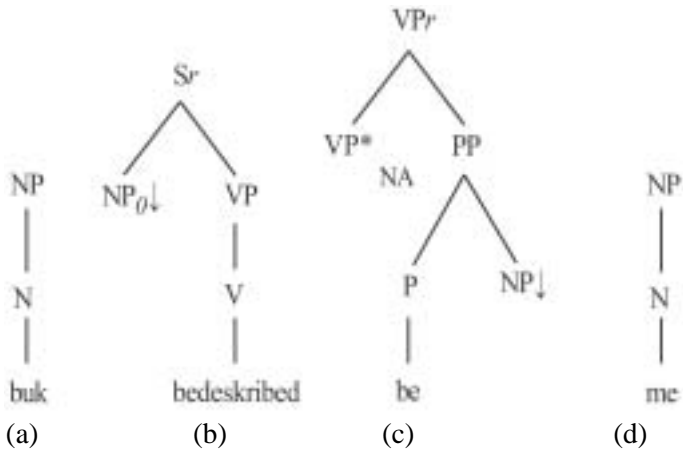

Figure 3. Exemplary trees for Unish
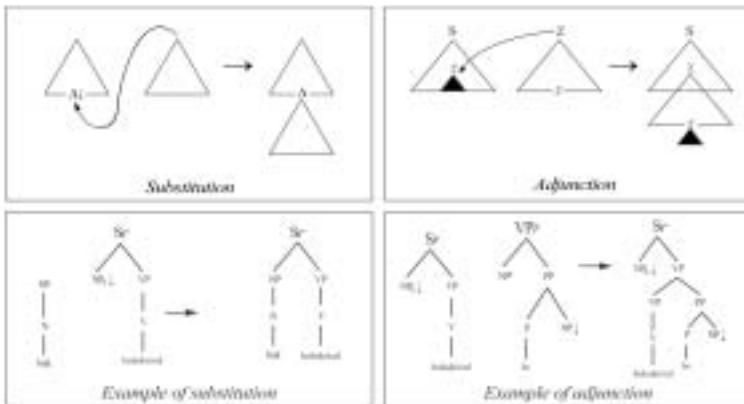
4. Tree Combining Rules

Figire 4. Combining operations

As there are no grammar rules in TAGs, combining operations are needed to combine each lexicalized structure. There are two operations defined in Tree Adjoining Grammars, namely, substitution and adjunction.

Substitution can take place only on non-terminal nodes of the frontier of the tree, and a substitution node is marked by a down arrow (   ). In the substitution operation, a node marked for substitution in an elementary tree is replaced by another elementary tree whose root label is the same as the non-terminal. So, in Figure 4, **A**   is replaced by the tree on the right side, whose root label is **A**.

In an adjunction operation, an auxiliary tree is inserted into an initial tree. The root and foot nodes of the auxiliary tree must match the node label at which the auxiliary tree adjoins. Actually, it is this operation that makes lexicalization possible. The adjunction operation is shown on the right of Figure 4. **VP*** node is used as an adjunction node.
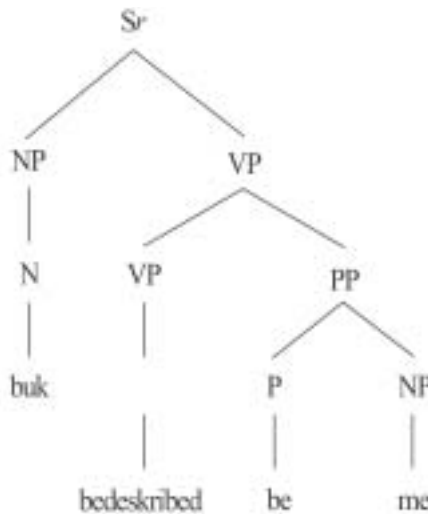

Figure 5. Final derived trees by combining trees in Figure 3

To help understand the tree-combining operations in Figure 4, let us combine all the elementary trees in Figure 3. Figure 5 shows the final derived tree for the sentence 2(b), "Buk bedeskribed be me" built, starting from elementary trees in Figure 2(a), (b), (c), and (d). Figure 2(a) and Figure 2(b) are combined by substitution at NP_0 node. Figure 2(c) and 2(d) can also be combined by substitution at NP_0 node. Finally, an adjunction operation at **VP** node can be applied to combine these trees.

5. Examples of Some Other Structures

Due to the fixed word order (SVO) in a Unish sentence, a *wh-word* appears in situ, in accordance with its function in a sentence, instead of moving to the front of a *wh-question* (Lee 2002).

(2) a. De   ver-ed   wat?
       you   see-pst   what
       'What did you see?'
    b. Wu   mit-ed   de?
       who   meet-pst   you
       'Who met you?'

In sentence (2a) the word *wat* 'what' functions as an object and thus it appears after the verb *vered* 'saw'. The word *wu* 'who' in sentence (2b) functions as a subject and thus it appears before the verb *mited* 'met'. Therefore, the only difference between a declarative sentence and an interrogative sentence in Unish is that the former ends with a period or a falling intonation, but the latter ends with a question mark or a rising intonation.

This greatly reduces the number of elementary TAG trees in Unish. As shown in Figure 6, various trees for wh-words would have been needed as in English, unless they are treated exactly as normal

noun phrases in Unish.

Figure 6(a) and (b) are auxiliary trees corresponding to a relative clause where the subject or object has been relativized. Figure 6(c) and 2(d) are initial trees corresponding to a wh-question on the subject and object. However, these trees are needed for English only where wh-movement exists. In Unish only one relative pronoun (namely, *dat*) is used, irrespective of its function in a sentence.
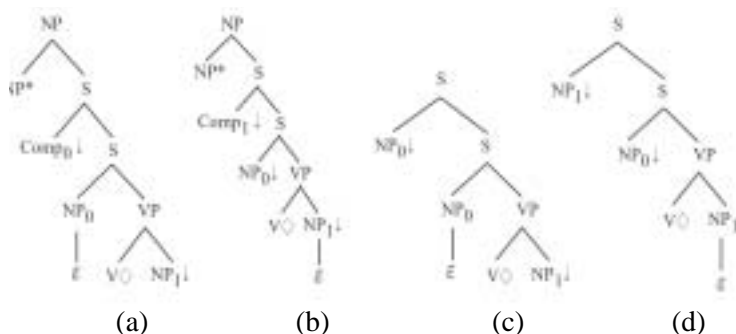

| (a) | (b) | (c) | (d) |

Figure 6. Other tree structures for word movement

(3) a. Les   mit-ed   gens   dat   ver-ed   muze.
        They   meet-pst   people   that   see-pst   museum
        'They met the people who saw the museum.'
     b. Me   ver-ed   buk   dat   de   scrib-ed.
        I   see-pst   book   that   she   write-pst
        'I saw the book that she wrote.'

Regardless of whether the relative pronoun *dat* functions as a pronoun or as an adverb, its form is the same. Except the occasion of *genitive* case the form of a relative pronoun is fixed regardless of the case it takes as illustrated in (3a) and (3b) (for example, the relative pronoun in (3a) takes *nominative* case, whereas the one in (3b) takes *accusative* case). Moreover, even when a special head noun of a relative clause occurs, the form of a relative pronoun does not vary.

Therefore, the total number of relative pronouns in Unish is much smaller than that in English, greatly reducing the number of elementary TAG trees for Unish. The tree in Figure 6(a) will be used to parse sentence 3(a), whereas the tree in Figure 6(b) will be used to parse sentence 3(b). Figure 7 shows a final derived tree for the sentence 3(b). The part of elementary tree in Figure 6(b) is highlighted in Figure 7.
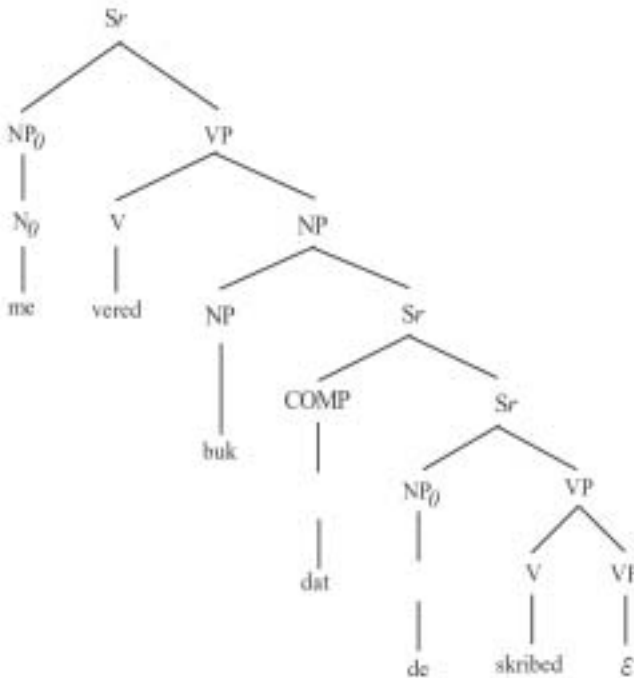

Figure 7. Final derived tree for sentence (3b)

6. Simplified Features

In TAGs, tree structures alone are not enough to represent the

Unish grammar. Thus, the 'feature' system should be introduced. The Feature-Based Lexicalized Tree Adjoining Grammar formalism (FB-LTAG) is based on the Tree Adjoining Grammar, which has been extended to include lexicalization and unification-based feature structures.

Each node of an elementary tree is associated with two feature structures, the top and the bottom. The bottom feature structure contains information relating to the subtree rooted at the node, and the top feature structure contains information relating to the supertree at that node.

Figure 8 shows an auxiliary tree and an elementary tree, and the trees resulting from a substitution operation and an adjunction operation.

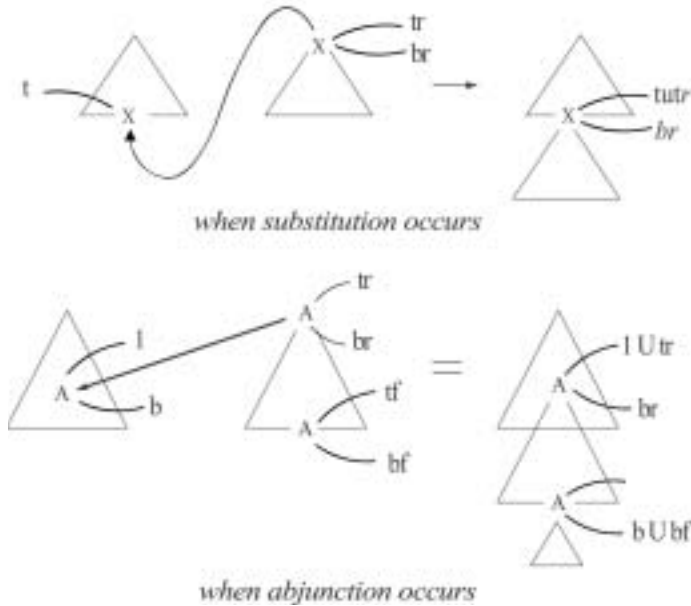when substitution occurs

when abjunction occurs
Figure 8. Feature unification

In the substitution operation, the features of the node at the substitution site are the unified features of the original nodes. The top feature structure of the node is the result of unification of the top features of the two original nodes, while the bottom feature structure of the new node is simply the bottom features of the root node of the substituting tree. So, in Figure 8, the top feature structure, t of **X,** should unify with the top feature structure, tr of the root node **X**.

In the adjunction operation, the top feature structure of nonterminal node, **A,** unifies with the top feature structure of the root node of the auxiliary tree, **A**, while its bottom feature structure unifies with the bottom feature structure of the foot node, **A** of the auxiliary tree, on the right side of Figure 8.

Lexicalized trees allow individual lexical items to instantiate the feature structures in the trees with lexically specific information. This may include, for instance, constraints that verbs place on their complements, or morphological and semantic information associated with an individual word. In lexicalized TAGs, at least one terminal symbol (the anchor) must appear at the frontier of all initial or auxiliary trees.

Nodes of elementary trees may specify constraints on the set of auxiliary trees that can adjoin to them. These constraints enforce obligatory adjunction of any auxiliary tree, selective adjunction of a specified set of auxiliary trees, or no adjunction at all.

Let us see how Unish grammar is influencing the feature system in TAGs by an example of NP (Noun Phrase) agreement. While the agreement in number and case occurs in an English NP, that phenomenon does not occur in an Unish NP.

(4) a. Me's   gut   ami   kom-ed.
      I-gen   good   friend   come-pst
      'My good friend came.'

    b. Me's   gut   ami-s   kom-ed.
       I-gen   good   friend-pl come-pst
       'My good friends came.'

    c. Les   visit-ed   me's   gut   ami.
       They   visit-pst I-gen   good   friend
       'They visited my good friend.'

    d. Les   visit-ed   me's   gut   ami-s.
       they   visit-pst I-gen   good   friend-pl
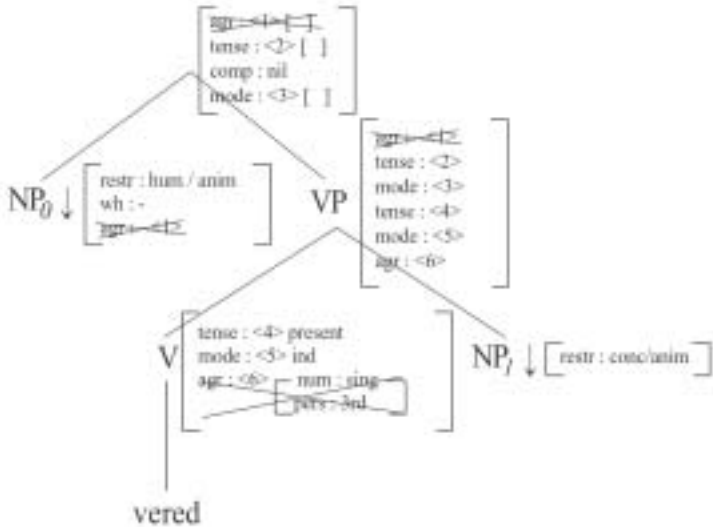   'They visited my good friends.'

Figure 9. "Vered" with NP agreement features missing

As shown in (4a) and (4c) the form of the NP *me's gut ami* is fixed irrespective of whether it is assigned *nominative* case or *accusative* case. The same applies to the form of the NP *me's gut amis* appearing in (4b) and (4d). In addition, regardless of the number of a noun in an NP, the form of a pronoun or an adjective that precedes the noun is fixed as illustrated in (4a-4b) and (4c-4d). Thus in Unish

the agreement between a noun and the words preceding that noun in an NP does not occur. Figure 9 shows how complicated a node would have been if all the agreement features had to be reflected in Unish.

5. Conclusion

I have presented some exemplary Unish TAG grammars. Throughout the paper, some evidences are shown that Unish grammar is substantially tailored, simplifying some complicated features usually existing in natural language. Still, the Unish TAG grammar presented here is preliminary and should be viewed as such; it meets the base requirements of LTAG, namely, encapsulation of predicate argument structures and factoring recursion from the domain of dependencies. Some of the trees in this paper may look arbitrary and indeed may be so, as the grammar is still developing. Further study will help remove this arbitrariness.

References

Choo, M. 1996. The Need for a Universal Language and Methods of its Creation as Suggested by Hangul. *Journal of Universal Language* 1, 5-10.

Choo, M. 2001. The Need for "Unish", a Universal Language and the Principles of its Development. *Journal of Universal Language* 2, 3-14.

Choo, M., E. Kwak, D. Lee, H. Park, Y. Chung, J. Tak, T. Akiko, & K. Bae. 2000. Seykyeye-uy Kaypal Panghyang [Directions for Developing Unish]. *The Second Seminar on Unish in 2000*, Seoul: Sejong University.

Chung, Y. 1996. An International Language for the World to Come. *Journal of Universal Language* 1, 56-70.

Comrie, B. 1996. Natural and Artificial International Languages: A Typologist's Assessment. *Journal of Universal Language* 1, 35-55.

Joshi, A., L. Levy, & M. Takahashi. 1975. Tree Adjunct Grammars. *Journal*

of Computer and System Sciences 10.1, 136-163.

Kim, S. 2001. The Landscape of Languages at the Commencement of the 21st Century. *Journal of Universal Language* 2, 15-23.

Large, A. 1996. The Prospects for an International Language. *Journal of Universal Language* 1, 20-34.

Lee, D. 2002. A Comparison of Unish Grammar with Esperanto. *Journal of Universal Language* 3, 57-74.