

## Tokenization and Morphological Analysis for Malagasy<sup>1</sup>

Mary Dalrymple\*, Maria Liakata\*, and Lisa Mackie\*

### Abstract

The authors present a tokenizer and finite-state morphological analyzer [Beesley and Karttunen 2003] for Malagasy, based primarily on the discussion of Malagasy morphology in Keenan and Polinsky [1998] and Randriamasimanana [1986]. Words in Malagasy are built from roots by means of a variety of morphological operations such as compounding, affixation and reduplication. The authors analyze productive patterns of nominal and verbal morphology, and describe genitive compounding and suffixation for nouns and various derivational processes involving compounding and affixation for verbs. This work offers a computational analysis of Malagasy morphology, and forms the basis of a computational grammar and lexicon of Malagasy within the framework of the PARGRAM project.

**Keywords:** Malagasy, Austronesian, Morphological Analyzer, Finite-State Morphology, PARGRAM

### 1. Malagasy in the PARGRAM Project

Malagasy is an Austronesian language spoken by about six million people on the island of Madagascar [Grimes 1999]. Along with Welsh, it is a focus of the Verb-Initial Grammars subproject (<http://users.ox.ac.uk/~cpgl0015/pargram/>) within the PARGRAM initiative, a collaborative project to develop computational lexicons and grammars within the shared linguistic framework of Lexical Functional Grammar [Butt *et al.* 2002].

The objective of PARGRAM is to develop parallel grammars for a range of different languages<sup>2</sup> using a shared linguistic framework and shared grammar writing techniques and technology. However, each project within the PARGRAM umbrella is driven by a different set of goals. For example, the English, German and Japanese grammars have been under development for a number of years; these grammars aim for very broad and robust coverage

---

<sup>1</sup> The research reported here is supported by a grant from the Economic and Social Research Council, UK (Project RES-000-23-0505).

\* Centre for Linguistics and Philology, University of Oxford, Walton Street, Oxford OX1 2HG UK, telephone: +44 (0)1865 280 403 fax: +44 (0)1865 280 412

E-mail: [mary.dalrymple@ling-phil.ox.ac.uk](mailto:mary.dalrymple@ling-phil.ox.ac.uk); [mal@aber.ac.uk](mailto:mal@aber.ac.uk); [l.mackie@lmh.oxon.org](mailto:l.mackie@lmh.oxon.org)

<sup>2</sup> For details see <http://www2.parc.com/isl/groups/nlft/pargram/>

for industrial applications. In contrast, the focus of attention in the Urdu and Hungarian projects is on theoretical linguistic issues: whether a coherent large-scale grammar and lexicon of these languages can be written in conformance with the linguistic assumptions common to all the PARGRAM grammars. The Welsh and Malagasy grammars fall at this end of the PARGRAM spectrum; the focus is on producing coherent, internally consistent, linguistically well-motivated large-scale grammars and lexicons of these languages, following the common PARGRAM assumptions, and using the common tools. As the grammar development work for Welsh and Malagasy work has progressed, the researchers have found that analyses of these languages as exemplars of the verb-initial type share a good deal of commonality at the phrase structure level of analysis, creating theoretical synergy within the Verb-Initial Grammars subproject. However, the languages also differ in interesting ways: most importantly, Welsh is a VSO language, whereas Malagasy is VOS. Exploring differences between these languages continues to enhance understanding of the range of variation possible within the verb-initial type.

Like all grammars within the PARGRAM project, the development of the Malagasy grammar relies heavily on a computational component for morphological analysis. For most of the other PARGRAM grammar development efforts, the task of building a morphological analyzer does not arise, since large-scale morphological analyzers already exist for many of the PARGRAM languages. For those grammars, the task is instead to incorporate these already existing morphological analyzers, which had often been created for shallow grammatical analysis or information retrieval applications. The challenge for these grammar development projects, then, is to overcome the problems arising from the lack of detailed grammatical information that these transducers made available.

The Malagasy grammar shares with a few of the other PARGRAM grammars (Arabic, Turkish, Urdu, Welsh) the difficulties and opportunities that arise when a morphological analyzer is developed in tandem with a syntactic grammar and lexicon. The advantages are that the morphological analyzer can be tuned to provide exactly the syntactic information that the grammar writer expects, and the division of labour between the morphology and syntax can be made in a well-motivated manner, rather than being imposed on the grammar writer. The disadvantages are that the grammar development effort tends to be delayed if any problems arise in developing the morphological analyzer, and any changes to the architecture of the morphological analyzer can necessitate overhauling the syntactic lexicon and grammar to restore compatibility between the two. Despite these disadvantages, the need for automatic morphological analysis for the Malagasy grammar project is acute, since entering into the lexicon each of the many hundreds of surface forms associated with a single verb, noun, or adjective root would miss important linguistic generalizations and impede progress in grammar development. In related work, Çetinoğlu and Oflazer [2006] explore some issues in

developing a morphological analyzer for Turkish, an agglutinative, morphologically complex language, in the context of the PARGRAM project.

For the Malagasy morphological analyzer, the researchers use Xerox finite-state tools LEXC and XFST [Beesley and Karttunen 2003], which are employed by many of the PARGRAM grammars. As with any finite-state morphological transducer, the Malagasy morphological analyzer is bidirectional: it can be used in grammatical analysis to produce morphologically analyzed input to a parser, or in generation to produce a surface form from a specification of lexical properties [Beesley and Karttunen 2003]. In the following, the authors often describe the tokenizer and morphological component in terms of analysis as opposed to generation of a surface string, but this is only for expository purposes.

## 2. Malagasy Morphology: An Overview

As Keenan and Polinsky [1998] note, there is very little inflectional morphology in Malagasy: there is no verb agreement or nominal inflection for agreement features, for example. Keenan and Polinsky [1998] analyze certain alternations in deictic forms and demonstratives as inflection, but since the forms involved belong to a small closed class, the authors identify them by listing them in the lexicon. The morphological analyzer described here handles many of the productive cases of nominal and verbal derivational morphology, consisting primarily of affixal verbal morphology and genitive compounding.

Besides verbal affixation and genitive compounding, the third productive type of morphological process in Malagasy is reduplication [Keenan and Razafimamonjy 1998], in which a new root is formed by reduplicating all or part of a basic root, giving a diminished, attenuated, or pejorative meaning: for example, reduplication of the root *fotsy* ‘white’ gives *fotsifotsy* ‘whitish’. It is well known that reduplication requires special treatment in a finite-state morphological model, and the COMPILE-REPLACE algorithm described by Beesley and Karttunen [2000; 2003] provides a means of treating these cases. The researchers have implemented and are currently testing a treatment of Malagasy reduplication using the COMPILE-REPLACE algorithm, but, as this has not yet been completely integrated into the full Malagasy grammar, the authors concentrate in the following on describing the treatment of nominal and verbal morphology.

## 3. Lexical Information

Malagasy roots may have one or more syllables. Most roots are regular or ‘strong’, and have penultimate stress if they are multisyllabic. Three-syllable roots take penultimate stress unless they end in one of the ‘weak syllables’ (*na/ny*, *ka*, *tra*) in which case they usually receive antepenultimate stress and are called ‘weak roots’ [Keenan and Polinsky 1998]. Weak and strong roots behave differently in the processes that are treated here, and are listed separately

in the morphological lexicon.

This lexicon currently contains 2,446 roots, including 2,033 roots which form nouns, adjectives or verbs, 288 roots which form adjectives or verbs, and 125 roots which form only verbs. Indeclinable forms, including proper names, adverbs, some prepositions, and free pronouns, are not listed in the morphological lexicon, and so are passed through the morphological analyzer unchanged and treated by the syntax as unanalyzed tokens. Guessed roots are also allowed for and are defined in terms of permissible root patterns; these roots are marked with the tag +Guess, and are permitted, though dispreferred, in syntactic analysis. In treatment of guessed forms, the authors define Syllable (Syll) as in (1); this allows the definition of weak guessed roots as consisting of two syllables followed by one of the weak endings (*na*, *ka*, *tra*). Strong guessed roots are then defined as consisting of one to four syllables, and subtracting the weak root patterns:

$$\begin{aligned} (1)\text{Syll} &= [((\text{Nasal}) ([t|d]) \text{Consonant}) (\text{Vowel} \text{Vowel})]; \\ \text{WeakKTRoot} &= [\text{Syll}^2 [ [ [T|t] [R|r] [K|k] [A|a] ] ] ]; \\ \text{WeakNRoot} &= [\text{Syll}^2 [ [N|n] [A|a] ] ]; \\ \text{StrongRoot} &= [\text{Syll}^{\{1,4\}} - [\text{WeakKTRoot} | \text{WeakNRoot} ] ]; \end{aligned}$$

The definitions in (1) follow standard XFST notation, as defined by Beesley and Karttunen [2003]: square brackets '[' '] indicate grouping, parentheses '(' ')' indicate optionality, '|' indicates union or disjunction, '-' indicates subtraction of the second set of strings from the first set of strings, and '^' followed by a number or range of numbers indicates the amount of times the immediately preceding string is repeated. Note that the use of '^' here is different than in the definition of the continuation classes and orthographic rules. In the latter case, '^' designates a feature to be interpreted by the XFST rules.

#### 4. Genitive Compounding

This analysis of verbal and nominal morphology closely follows the exposition of Keenan and Polinsky [1998]. Nominal morphology consists mainly in the formation of genitive compounds. These are of the form Head+NP<sub>gen</sub>, where the Head can be any of the following: noun (in which case NP<sub>gen</sub> expresses the possessor), passive verb (NP<sub>gen</sub> is the agent), preposition (the NP<sub>gen</sub> is the prepositional object) or adjective (the NP<sub>gen</sub> is an agent or indirect cause). In such expressions, the Head and NP<sub>gen</sub> are concatenated, and the concatenation is regulated by rules referring to properties of the final syllable in the Head and the first syllable in NP<sub>gen</sub>.

#### 4.1 Target Phenomena

The following informal rules follow the treatment of Keenan and Polinsky [1998], and represent alternations in the final and first syllables of Head and NP<sub>gen</sub> respectively. The hyphen, which sometimes alternates with apostrophe, is part of Malagasy orthography. The expressions ‘C’, ‘Co’ stand for consonants and ‘V’ and ‘Vo’ stand for vowels, whereas ‘S’ denotes a stop consonant. The lowercase characters denote the corresponding letters. The expression to the left of the ‘+’ sign stands for the final syllable of the head word, consisting of some strong consonant ‘C’ and some vowel ‘V’ in the case of (1a) and (2) and some vowel ‘V’ followed by one of the weak endings ‘ka’, ‘tra’, ‘na’ in (1b). The expression to the right of the ‘+’ sign stands for the initial character of the NP<sub>gen</sub>.

1. Head is weak, that is, it ends in one of *ka*, *tra*, *na*
  - (a) NP<sub>gen</sub> begins with a vowel Vo: CV + Vo → C-Vo (remove final vowel in Head and concatenate)
  - (b) NP<sub>gen</sub> begins with a consonant C with corresponding stop consonant S:
    - i. Head ends in *na*:  
 Vna + C → Vn-S (S not bilabial), or  
 Vna + C → Vm-S (S bilabial)
    - ii. Head ends in *ka* or *tra*:  
 V{ka|tra} + C → V-S
2. Head is not weak:
  - (a) NP<sub>gen</sub> begins with a vowel Vo:  
 CV + Vo → CVn-Vo (prefix *n*-and concatenate)
  - (b) NP<sub>gen</sub> begins with a consonant Co with corresponding stop consonant S:  
 CV + Co → CVn-S (S not bilabial), or  
 CV + Co → CVm-S (S bilabial)

Similar to noun genitive expressions are pronominal suffixed genitives. If the Head ends in a non-weak syllable or *na*, then the GEN1 suffixes are attached to the Head. Otherwise, the GEN2 suffixes are attached.

(2) person	GEN1 suffix	GEN2 suffix
1sg.	ko	o
2sg.	nao	ao
3sg. or pl.	ny	ny
1pl. incl.	ntsika	tsika
1pl. excl.	nay	ay
2pl.	nareo	areo

## 4.2 Implementation

The rules governing genitive expressions are quite regular and consistent. The morphology of such expressions is modelled by the Xerox finite-state calculus, with a tokenizer written in XFST, a lexicon written in LEXC, and more general orthographic and phonological rules written in XFST [Beesley and Karttunen 2003].

As with the other grammar development projects within the PARGRAM initiative, the grammar is implemented within the XLE grammar development environment ([Crouch *et al.* 2006], see also <http://www2.parc.com/isl/groups/nlitt/xle/>). The XLE requires a tokenizer and morphological analyzer for the language being analyzed, and allows the specification of a sequence of alternative morphological analyzers to be used when analysis with the first alternative fails. The output of the morphological analyzer is the input to syntactic analysis, obviating the need for listing each surface form separately in the syntactic lexicon. Instead, the syntactic component contains information about the syntactic content and behaviour of each root and affix combination as analyzed by the morphological component.

XLE expects a string as input, which is first tokenized according to the rules of the tokenizer for the language being analyzed. Each token is then individually passed to the morphological analyzer for finite-state morphological analysis. Most grammars within the PARGRAM initiative employ at least two morphological analyzers: an analyzer for known forms, and a guesser for forms that fail to be analyzed by the known-form analyzer. Following this paradigm, the known-form transducer and guesser are extracted separately, and applied to the output of the tokenizer in sequence; only forms that fail to obtain an analysis with the known-form analyzer are passed to the guesser.

In most cases, tokenization in Malagasy is straightforward, with tokens usually delimited by whitespace. For the sentence *Hanketo izy*. ‘he will come here’, the tokenizer produces the following result, where TB indicates a token boundary:

(3)hanketo TB izy TB .

The tokenizer (optionally) decapitalizes the first word of the sentence, inserts token boundaries at spaces, and separates punctuation by a token boundary. Each token is then passed separately to the morphological analyzer for analysis.

In the case of nonpronominal genitive compounding, however, the situation is more complex. The compound form *akanjon-olona* ‘a person’s clothes’ consists of two noun roots *akanjo* ‘clothes’ and *olona* ‘person’, and has the following structure:

## (4) akanjon-olona

akanjo (epenthetic N; cf. 2a above) (compound boundary) olona

In the original treatment of nonpronominal genitive compounding, this form was treated as a single token, and was handled by the morphological analyzer [Dalrymple *et al.* 2005]. However, this approach interacted badly with the standard configuration of XLE grammars, where the known-form analyzer and guesser are separate transducers, applied in sequence. If both roots are known, both can be analyzed by the known-form transducer; however, if one root is unknown, the entire compound fails to be recognized by the known-form analyzer. This means that the entire compound must be handled by the guesser, even if one of the roots is known. This undesirable result has led the researchers to revise the treatment of nonpronominal genitive compounds, moving most of the work to the tokenizer.

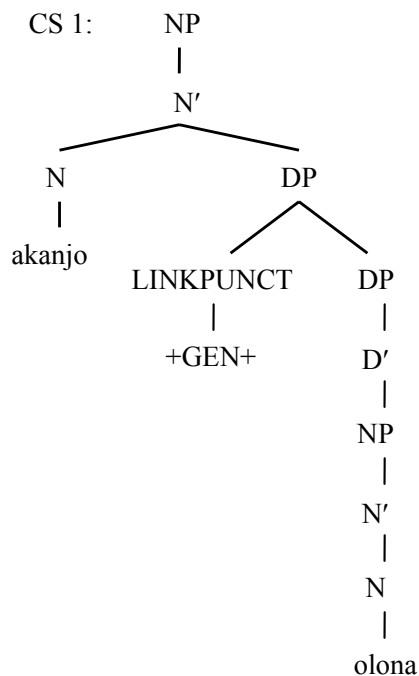
In the current treatment, the tokenizer ‘undoes’ the effects of the compounding rules given above, proposing one or more underlying forms for analysis by the morphological component. For example, the compound form *akanjon-olona* is tokenized as follows:

## (5) akanjo TB +GEN+ TB olona

The root *akanjo* is a three-syllable root; the hypothetical root *akanjona* would have four syllables, which is impossible for a weak root. For this reason, Rule 1 (above), which requires that Head is a weak root, does not apply. Rule 2a covers the case in which the second member of the compound begins with a vowel; “undoing” rule 2a entails removing the epenthetic n inserted after the Head, producing *akanjo*. The compound boundary is treated as a separate token, represented by the special symbol +GEN+, to signal to the syntactic analysis component that genitive compounding has taken place. The phrase structure tree that is produced for *akanjon-olona* is shown in Figure 1, in which the leaves of the tree correspond to the tokens produced by the tokenizer.

Some forms can be tokenized in more than one way. For example, the compound *volon-dRabe* ‘Rabe’s month’/‘Rabe’s money’ is ambiguous [Keenan and Polinsky 1998]:

- (a) Head is weak, reconstructed by the tokenizer as *volana* ‘month’, with Rule 1a requiring removal of the final vowel; or
- (b) Head is strong, reconstructed by the tokenizer as *vola* ‘money’, with Rule 2a requiring insertion of -n



**Figure 1. Phrase structure tree for *akanjon-olona***

After tokenization, the morphological analyzer is presented with all possible forms resulting from ‘undoing’ the compounding rules. Analysis proceeds as in the simple cases, with forms recognized by the known-form analyzer given preference in syntactic analysis over hypothetical forms analyzed by the guesser.

The current treatment of genitive compounding relies on the presence of the hyphen or apostrophe to signal the compound boundary. In a small minority of cases, however, genitive compounding involves only concatenation of roots, and is not signalled by special punctuation. The authors have left the treatment of these forms for future work, since it is unclear how the treatment of such forms will interact with the guesser: almost any simple form can be given a spurious analysis as a compound composed of two hypothetical, guessed roots.

Pronominal genitive compounds, on the other hand, are best treated by the morphological analysis component, which is now described. The LEXC lexicon is a finite-state transducer which specifies a relation between an Upper ‘lexical’ string and a Lower ‘surface’ string for a form [Beesley and Karttunen 2003]. Roots and affixes are organized into sublexicons according to their phonological and prosodic properties, *e.g.* whether the root is weak or strong. The lexicon also specifies possibilities for transitions when a particular form is encountered. For example, the noun root *akanjo* ‘clothes’ is listed in the Noun sublexicon with continuation class Nstrong, meaning that it takes the strong root suffixes listed in the Nstrong sublexicon. The Nstrong sublexicon adds the +Noun tag to the lexical/Upper side of the



transducer, and permits the form to terminate with no suffixation, or alternatively allows genitive suffixation. Thus, the transducer relates the Lower string, the unsuffixed noun *akanjo*, to the morphologically analyzed lexical/Upper string which forms the input to syntactic analysis:

(6)LEXC transducer:

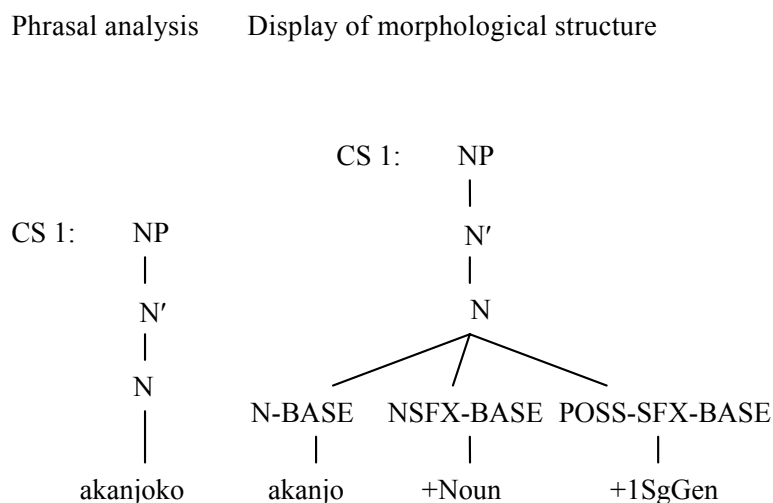
Upper: akanjo +Noun  
 Lower: akanjo  
 ‘clothes’

The related form *akanjoko* ‘my clothes’ involves pronominal genitive suffixation; the NStrong sublexicon relates the first person singular genitive suffix *ko* on the surface/Lower side to the tag +1SgGen on the lexical/Upper side:

(7)LEXC transducer:

Upper: akanjo +Noun +1SgGen  
 Lower: akanjoko  
 ‘my clothes’

Here, the LEXC lexicon on its own is sufficient for analysis of the combination of the root *akanjo* and the pronominal genitive suffix *-ko*. The phrase structure tree that is produced for *akanjoko* is shown on the left side of Figure 2; the right side shows the root and series of tags that is output by the morphological analyzer and analyzed by the syntactic component.



**Figure 2. Analysis of *akanjoko***

## 5. Verbal Morphology

In many cases, however, a set of XFST rules is also needed to cater to phonological and orthographic alternations induced by morphological operations. These rules apply irrespective of the individual entries to be combined, and are controlled by tags introduced by LEXC. These tags are orthographically distinguished from the lexical tags of the LEXC transducer by the use of a carat '^'. The XFST rules define an XFST transducer, which is composed with the LEXC transducer in full morphological analysis. In the following, the authors describe the use of these tags in the analysis of verbal morphology.

### 5.1 Target Phenomena

Malagasy exhibits rich and complex verbal morphology [Randriamasimanana 1986; Keenan and Polinsky 1998]. Verbs are classified according to the case of their arguments: nominative, accusative and genitive. Verbs which take a genitive complement are non-active verbs, a category which includes passive verbs and circumstantial verbs. Passive verbs are formed in three different ways, each corresponding to different semantics. The following discussion follows Keenan and Polinsky [1998], though simplifying somewhat.

First, there are a small number of root passives, that is, roots which are passive verbs. These refer more to the result than the process. The LEXC transducer encodes the schematic relation for passive roots in Figure 3, which is very similar to patterns for noun roots with optional pronominal genitive compounding. ROOT represents the form of the passive root. ROOTTYPE is one of ^StrongRoot, ^WeakKTRoot or ^WeakNRoot; this information is needed by the XFST rules to control certain morphological alternations. (GEN) represents optional genitive compounding with the agent argument of the passive verb, using the affixes listed in (2). An example for the root passive *araka* 'be followed' is:

(8)Lexical: araka +Verb +3Gen

Surface: arany

'be followed by him/them'

The LEXC transducer is composed with the XFST transducer, which performs necessary adjustments as the morphemes are concatenated.

The largest category of passive verbs is suffix passives. These are formed by the suffixation of *ina* or *ana* to a root, which is usually preceded by a root-dependent consonant C epenthesis. They can be prefixed by a tense prefix TENSE, denoting past or future, optionally followed by a causal prefix *amp*. This form can also undergo genitive compounding or imperative suffixation.

**Passive roots**

ROOT +Verb (GEN)  
 ROOT ROOTTYPE ...

**Suffix passives**

TENSE (Caus) ROOT +Verb (C)VnaPass (GEN|IMP)  
 ... amp ROOT ROOTTYPE (C)ina/ana ...

**Prefix passives**

VTPass ROOT +Verb (GEN|IMP)  
 voa/tafa ROOT ROOTTYPE ...

**Circumstantial form**

TENSE (Caus) (ACTIVE) ROOT +Verb (C)VnaPass  
 ... amp i/an ROOT ROOTTYPE (C)ina/ana

**Active Verbs**

(TENSE|NOM) [(Recip)(Caus)][ACTIVE PassROOT|NullPrefROOT] +Verb (IMP)  
 ... if amp i/an PassROOT|NullPrefROOT ROOTTYPE ..

**Figure 3. Verbal patterns**

A third type of passive is prefix passives. These are formed by prefixing a root with any of *a*, *voa*, *tafa*. Passives in *a* refer to the process rather than the result, and usually their subject functions as an instrument. The imperative is formed by prefixing with *a* and adding the corresponding passive imperative suffix. Passives in *voa/tafa* refer to the end result rather than the process and have a perfective meaning. *voa/tafa* passives may not be prefixed by a tense prefix, while a passive does take a tense prefix.

In the circumstantial form of a verb, an oblique argument or adjunct of an active verb is made the subject. The circumstantial is built from roots prefixed by primary active affixes *i*, *an* and, possibly, secondary active affixes *ank(a)*, *amp* by means of the suffixation *-Cana*, where *C* is the root-specific epenthetic consonant mentioned above in the context of suffix passives. Tense is marked in the same way as for suffix passives.

There are a few active verb roots, but the majority of active verbs are derived from roots by means of the active prefixes *i*, *an*. Genitive suffixing is not allowed, but the formation of

imperatives is possible: present tense (*m*) actives take suffix *a*, where consonant mutation and epenthesis *-(C)a* apply. If no epenthetic consonant intervenes, they fuse an imperative with root final *a*. Active verbs can be marked for tense via a tense prefix TENSE (distinguishing past, present, and future). They can also receive a prefix for causality and reciprocity. The active verb roots may be null prefix or they can be prefixed by the active prefixes *ank-/amp*.

## 5.2 Implementation

As discussed above, the LEXC lexicons contain information about subclasses of individual roots as well as more general structural information regarding verb forms. For example, verbs are formed on the basis of a tense prefix sublexicon which contains separate past, present and future prefixes, including a *^TNS^* tag to control morphological alternations with overt tense prefixes. In the following, 0 represents the empty string.

```

LEXICON Tense
PresentTense+:0      Secondary;
PastTense+:no^TNS^  Secondary;
FutureTense+:ho^TNS^ Secondary;
PresentTense+:0      Vroot;
PastTense+:no^TNS^  Vroot;
FutureTense+:ho^TNS^ Vroot;

```

The lexicon VPassRoot represents the passive verbs: inherently passive roots, or guessed passive verbs ending in either a strong or weak syllable.

```

LEXICON VPassRoot
PassiveRoot ;
<StrongRoot %+Guess:0> StrongSuff ; ! guessed Strong root
<WeakKTRoot %+Guess:0> KTWeak ; ! guessed Weak KT root
<WeakNRoot %+Guess:0> NWeak ; ! guessed Weak N root

```

Roots are listed in the lexicon with information about the continuation classes of their suffixes:

```

LEXICON PassiveRoot
araka      WeakSuff;      ! be followed
fantatra   TR2RWeak;     ! be known

```

In this example *araka* is a passive root; its continuation class indicates that it is a member of the class of morphologically weak roots. *fantatra* is a weak passive root with final syllable *tra*, where the TR2RWeak continuation class indicates that the *tra* suffix for this root is replaced with *r* during passive suffixation or the formation of imperatives. Thus, the passive form corresponding to *fantatra* is *fantarina*.

As above, the XFST rules deal with surface phenomena such as syllable deletion and consonant and vowel epenthesis, which take place during affixation. In the previous example, the continuation class TR2RWeak is used with roots where the weak final root syllable *tra* is converted to *r* during passive suffixation or the formation of imperatives. Other weak roots convert *tra* to one of a number of other consonants which must be lexically specified for each root. One way of handling these alternations would be to have a continuation class for each of the possible combinations of suffixes and final syllables of roots. Thus, even though there are only two passive suffixes *ina/ana*, one would need separate continuation classes for the formation of passives for weak roots ending in *tra* where *tra* is transformed to *r, f, t*, or other consonants.

However, this would result in an over-sized, untidy lexicon. Instead, the authors keep a small number of continuation classes corresponding to possible suffixes, and signal the final syllable root transformations by means of tags referenced by rules of the XFST transducer. These tags provide the context for the application of XFST rules for the various cases of epenthesis, deletion and transformation. For instance, the TR2RWeak continuation class is defined in the following way:

```
LEXICON TR2RWeak
+Verb:^WeakKTRoot      WeakKTEnding ;
+Verb:^WeakKTRoot^Ftr2r  Suffixes ;
```

The feature ^Ftr2r is referenced by the XFST rule in (9), which transforms *tra* to *r* if the *tra* syllable is followed by the feature ^Ftr2r.

$$(9)[t \ r \ a] \rightarrow r \ || \ \_ \ ^Ftr2r$$

This XFST rule applies to the underlying grammatical form, the output of LEXC. Formally, it resembles a standard context-sensitive phrase structure rule: the expression to the left of the arrow is replaced by the expression to the right of the arrow in a certain context. The context of application is separated from the rule by double bars, '|'. Here, the sequence '*tra*' is replaced by '*r*' in the context immediately preceding the tag ^Ftr2r. This rule applies after

removal of the tag  $\hat{\text{WeakKTRoot}}$ , which separates the root from the  $\hat{\text{Ftr2r}}$  tag in the Lower string of the LEXC transducer. Directly after the application of this rule, the rule to remove the tag  $\hat{\text{Ftr2r}}$  applies, preventing its appearance in the surface string and its interference with the application of other rules. Similar rules cater to alternations with prefixation, passive and imperative formation.

Features are an efficient way of modelling local morphological dependencies and alternations. However, in the morphology of Malagasy verbs there are long distance dependencies which cannot be modelled by standard FST techniques. For instance, there are roots which can form the passive in either *ana* or *ina* but not both. Thus, we want *fantarina* and not *fantarana* to be recognized as the correct passive form of *fantatra*. This is a problem both in recognition and generation as we do not want our rules to accept or generate incorrect forms. A tag could be added to the root *fantatra* to exclude the passive formation in *ana*, but the tag may be separated from the position of *ina/ana* by other morphemes and tags during passive formation.

For example, if one decided to implement the lexical preference for passive suffixation in *-ina* rather than *-ana* as a feature, the lexical entry for *fantatra* in the lexicon would be accompanied by a feature  $\hat{\text{Fpassi}}$  on the surface level as below. In the following hypothetical lexicon, the root and its associated tags are grouped together by angled brackets ‘<’ and ‘>’ as is standard in LEXC:

```
LEXICON OtherRoot
<{fantatra} 0: $\hat{\text{Fpassi}}$ >      TR2RWeak;
```

However, this means that when the features  $\hat{\text{WeakKTRoot}}\hat{\text{Ftr2r}}$  are added by the continuation class TR2RWeak, they are not immediately next to the root, but rather  $\hat{\text{Fpassi}}$  stands in the way. As a result the rule (9) above for the transformation of the weak syllable *-tra* to *-r*, which precedes passive suffixation, cannot apply.

Fortunately, XFST allows for the treatment of such dependencies by the use of flag diacritics, non-FST handles which can store information that is not compiled into the FST. This information is used at runtime, when a certain phrase is being analyzed or generated. The authors use flag diacritics to store root-specific information, and, therefore, they are entered together with the lexical entry for the root. As they do not take effect until the interpretation phase, they do not interfere with the XFST rules. Thus, the lexical entry for *fantatra* in the previous section becomes:

```
LEXICON OtherRoot
<{fantatra} @U.PASS.I@> TR2RWeak;
```

This information uses a U-type flag diacritic, represented as @U.PASS.I@, to associate the feature PASS with the value I for this root. This feature ensures that the root *fantatra* takes a passive in *ina* and not in *ana*. This is coupled with matching flag diacritics for the passive suffixes:

```
LEXICON PassaSuff
<+Passa:a 0:n 0:a @U.PASS.A@> #;
```

```
LEXICON PassiSuff
<+Passi:i 0:n 0:a @U.PASS.I@> #;
```

The passive suffix *ina* is associated with the flag diacritic @U.PASS.I@, which is defined as above as specifying the value I for the feature PASS, while the suffix *ana* is associated with the flag diacritic @U.PASS.A@ specifying the value A for the same feature. Whenever flag diacritics meet, they must match; therefore the form *fantarana* is not accepted, as the flag diacritics of *ana* do not match the flag diacritics of *fantatra*.

The researchers also make use of R-type flag diacritics to model long distance dependencies between prefixes and suffixes. R-type flag diacritics are similar in structure to U-type diacritics, and are specified in the same way; crucially, however, R-type diacritics check that a certain value of a feature has been previously set by another flag diacritic specification. For example, Malagasy verbs may be formed from roots which are lexically nouns or adjectives. Since this is a very general fact about Malagasy, the lexicon does not list every root form in both the noun lexicon and the verb lexicon; nevertheless, one must ensure that the prefixes that appear with a root are compatible with its suffixes, since, if a root appears with verbal prefixes, it allows verbal but not nominal suffixes.

This is handled by setting the flag POS (part of speech) to VERB or NOUN at the beginning of the word, depending on what prefixes have been encountered. Then one must check that the suffixes that appear with a root are compatible with its prefixes. For example, the strong noun root *halatra* ‘theft’ is specified with the continuation class NStrong, which allows either the +Noun tag and noun suffixes with the continuation class NStrongEnding, or the +Verb tag and verbal suffixes with the continuation class VStrongEnding. NStrongEnding uses an R-type flag diacritic to check that the flag diacritic POS has been set to NOUN, preventing noun suffixes from appearing with verb prefixes; similarly, the VStrongEnding

lexicon checks that the flag POS is set to VERB.

The continuation classes, together with the rules and flag diacritics, give a general model for the construction of different verb forms. In the analysis of Malagasy verbal morphology, there are many exceptions to be taken into account which render the task of modelling verb morphology non-trivial. For instance, a root may not accept a certain affix; this can be handled by more sophisticated flag diacritics which govern the permissible affixes that each root can accept. The current analyzer uses flags, encoding 9 features with various values for the different affixes, which can be negatively specified by particular roots to disallow particular affixes or affix combinations.

As noted by Beesley and Karttunen [2003], more general cases can be ruled out by means of filters, which are sets of rules that apply on the lexical level – that is, on the Upper side of the LEXC transducer. Such filters are used to exclude groups of continuation classes from combining with a certain affix or can merge together morphological information. For instance, the lexical tag +Passa indicates that one has a passive form in *ana*, which can signal either a suffix passive or a circumstantial form. However, if it is preceded by the tag ActiveAN+, it is unambiguously a circumstantial form. The current treatment incorporates 5 filters disallowing certain affix combinations for all roots. Encoding such interactions in the morphological analyzer provides important constraints for syntactic analysis.

## 6. Conclusion

The authors have presented a computational implementation of the derivational morphology of Malagasy, concentrating on the treatment of genitive compounding and affixal verb morphology. This approach closely follows the analysis of Keenan and Polinsky [1998] and realizes the aforementioned morphological processes in terms of LEXC continuation classes, associated with groups of productive roots and general structural information, and general orthographic and phonetic rules implemented as XFST rules.

The morphological analyzer provides a solid basis for continuing work on the syntactic lexicon and grammar of Malagasy. Currently, the syntactic lexicon has been populated with the root and affix forms generated by the morphological analyzer and accepts these forms as input for syntactic analysis. The authors have encoded the syntactic contributions of each affix as well as default syntactic contributions for large classes of verb, noun, and adjective roots, and in current work are refining the lexicon to account for subclasses of roots with exceptional, non-default behaviour. The Malagasy grammar comprises 22 preterminal categories appearing on the left-hand side of phrase-structure rules with regular-expression right-hand sides covering a number of possible expansions. With the morphological analysis component in place, the researchers anticipate now being able to make rapid progress in expanding the coverage of the Malagasy grammar and syntactic lexicon.



### Acknowledgments

The authors are grateful to Charles Randriamasimanana for help in the development of the system.

### References

- Beesley, K. R., and L. Karttunen, "Finite-state non-concatenative morphotactics," in *Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON-2000)*, 2000., pp. 1-12.
- Beesley, K. R., and L. Karttunen, *Finite-State Morphology*. CSLI Publications, Stanford, 2003.
- Butt, M., H. Dyvik, T. H. King, H. Masuichi, and C. Rohrer, "The Parallel Grammar Project," in *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, Taipei, 2002.
- Çetinoğlu, Ö., and K. Oflazer, "Morphology-syntax interface for Turkish LFG," in *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, 2006.
- Crouch, D., M. Dalrymple, R. Kaplan, T. King, J. Maxwell, and P. Newman, XLE documentation. Technical report, Palo Alto Research Center, Palo Alto, CA. [www2.parc.com/istl/groups/nltt/xle/doc/xle\\_toc.html](http://www2.parc.com/istl/groups/nltt/xle/doc/xle_toc.html), 2006.
- Dalrymple, M., M. Liakata, and L. Mackie, "A two-level morphology of Malagasy," in *Proceedings of PACLIC 19, the 19th Asia-Pacific Conference on Language, Information, and Computation*. Taipei: Academia Sinica, 2005.
- Grimes, B. F., *ETHNOLOGUE: Languages of the World*. SIL International, 1999. URL [www.sil.org/ethnologue/](http://www.sil.org/ethnologue/).
- Keenan, E. L., and M. Polinsky, "Malagasy," in Andrew Spencer and Arnold Zwicky (editors), *The Handbook of Morphology*. Blackwell Publishers, Oxford, 1998.
- Keenan, E. L., and J. P. Razafimamonjy, "Reduplication in Malagasy," in *The Structure of Malagasy III: UCLA Working Papers in Syntax and Semantics*. Los Angeles: UCLA Linguistics Department, 1998.
- Randriamasimanana, C., *The Causatives of Malagasy*. Honolulu, 1986.



# Multiply Quantified Internally Headed Relative Clause in Japanese: A Skolem Term Based Approach

Rui Otake\*, and Kei Yoshimoto\*

## Abstract

This paper presents an analysis of Internally Headed Relative Clause (IHRC) construction in Japanese within the framework of Combinatory Categorical Grammar [Steedman 2000]. Shimoyama [1999] argues that when an IHRC appears within the scope of a universal quantifier, the interpretation of the IHRC exemplifies E-type anaphora and that the LF representation of the IHRC should have a variable bound by the quantifier in the matrix clause. To accommodate this argument Shimoyama posits a free variable of a functional type to which the bound variable is applied, and whose denotation is determined by the context-dependent assignment function. However, since there is in principle no limit to the number of quantifiers in the matrix clause (and accordingly that of bound variables in the IHRC), the semantic type of the free variable would be highly ambiguous if the IHRC occurs within the scope of multiple quantifiers. The current analysis assumes that the interpretation of IHRCs exhibits an instance of generalized Skolem term [Steedman 2005], a term whose denotation varies with the value of bound variables introduced by scope-taking operators, but which is interpreted as a constant in the absence of such operators. This paper provides a straightforward account for the semantics of the construction without invoking the complexities of the type ambiguity of free variables.

**Keywords:** Combinatory Categorical Grammar, Generalized Skolem Term, Internally Headed Relative Clause, Japanese, Quantification

## 1. Introduction

This paper presents an analysis of Internally Headed Relative Clause (IHRC) construction in Japanese paying particular attention to the effect of quantification on its interpretation. (1)

---

\* Graduate School of International Cultural Studies, Tohoku University, 41 Kawauchi, Aoba-ku, Sendai 980-8576, Japan. Tel: +81-22-795-7550, Fax: +81-22-795-7850  
E-Mail: otake@linguist.jp; kei@insc.tohoku.ac.jp

illustrates the basic form of the construction:<sup>1</sup>

- Taroo-ga [Hanako-ga ringo-o muita] no-o tabeta.  
 Taro-NOM Hanako-NOM apple-ACC peeled NML-ACC ate (1)  
 ‘Taro ate the apple that Hanako peeled.’

The bracketed clause *Hanako-ga ringo-o muita* ‘Hanako peeled an apple’ is followed by the nominalizer *no* and the accusative particle *-o*, thereby construed as the object of the matrix verb *tabeta* ‘ate’. Since the verb requires as its semantic restriction that the accusative argument be an edible thing, it anaphorically picks up the referent of *ringo* ‘apple’ from the embedded clause. This kind of construction is often contrasted with the Externally Headed Relative Clauses (EHRC), which is illustrated in (2).

- Taroo-ga [Hanako-ga muita] ringo-o tabeta.  
 Taro-NOM Hanako-NOM peeled apple-ACC ate (2)  
 ‘Taro ate the apple that Hanako peeled.’

As can be seen from the translation, (1) and (2) have almost the same meaning. But we hasten to add that IHRCs are not always paraphrasable to the EHRC version, for the former construction is subject to some pragmatic condition for its felicitous use, which we will not attempt to specify. A terminological note: we use the term the *antecedent* of an IHRC to mean the referent of the IHRC which functions as the argument of the matrix predicate. And we also use the term the *head* of the IHRC to refer to the linguistic element in the IHRC which describes the antecedent. For example, the antecedent of the IHRC in (1) is the apple that Hanako peeled, and the head is the noun *ringo* ‘apple’. It is important to notice that we define IHRC construction in terms of the nominal character of its anaphoric referent, despite the name suggesting that the presence of the head noun inside the relative clause is the defining feature of the construction. In fact, there are cases where the IHRC has no explicit nominal head. Such examples will be dealt with in section 2.2.

This paper is organized as follows. In section 2, we review the observation made by Shimoyama [1999] and her E-type analysis of IHRC. We then address the problem that the E-type analysis would raise focusing on multiply quantified IHRCs. We also discuss the

---

<sup>1</sup> Abbreviations used: ACC = accusative, ALL = allative, CL = classifier, COMP = complementizer, COP = copula, GEN = genitive, LOC = locative, NML = nominalizer, NOM = nominative, TOPIC = topic.

interpretational characteristics of IHRC, drawing on the study of Kikuta [2000]. Section 3 introduces the notion of generalized Skolem term, which provides a straightforward account for multiply quantified IHRC. Finally, section 4 concludes.

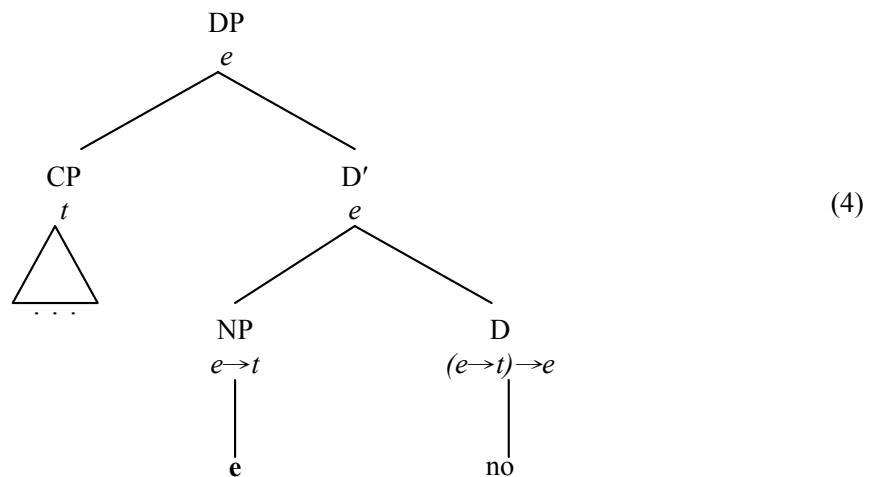
## 2. Previous Analysis and Its Problem

### 2.1 Shimoyama’s [1999] E-type analysis

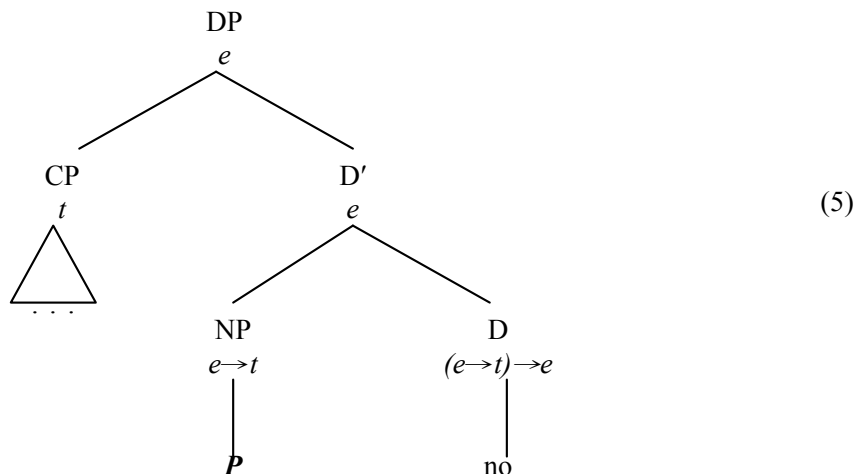
Shimoyama [1999] claims that when IHRC appears within the scope of a universal quantifier, the interpretation of IHRC exemplifies E-type anaphora:

Dono gakusei<sub>i</sub>-mo [soitu<sub>i</sub>-ga kongakki peepaa-o san-bon kaita] no-o  
 Every student he-NOM this:semester paper-ACC three-CL wrote NML-ACC  
 kesa teisyutusita. (3)  
 this:morning turned:in  
 ‘This morning every student turned in the three term papers he or she wrote this semester.’

In (3), the subject of the embedded clause is bound by the universal quantifier in the matrix clause. Note that “[t]he matrix object [...] does not refer to any particular set of term papers [Shimoyama 1999],” and the interpretation of the matrix object can be paraphrased by the definite description *the term papers he or she wrote this semester*. The interpretation of the relative clause is an instance of E-type anaphora [Evans 1980]. Given this observation, Shimoyama [1999] proposes the LF structure of IHRC as schematized in (4). According to this analysis, CP in the Spec of DP corresponds to the relative clause, which moves up to some higher position and is interpreted independently because of type mismatch. The DP is headed by the nominalizer *no*. The complement NP indicated by *e* is an empty pronoun. Basically, the interpretation of this empty pronoun determines the argument of matrix predicate.



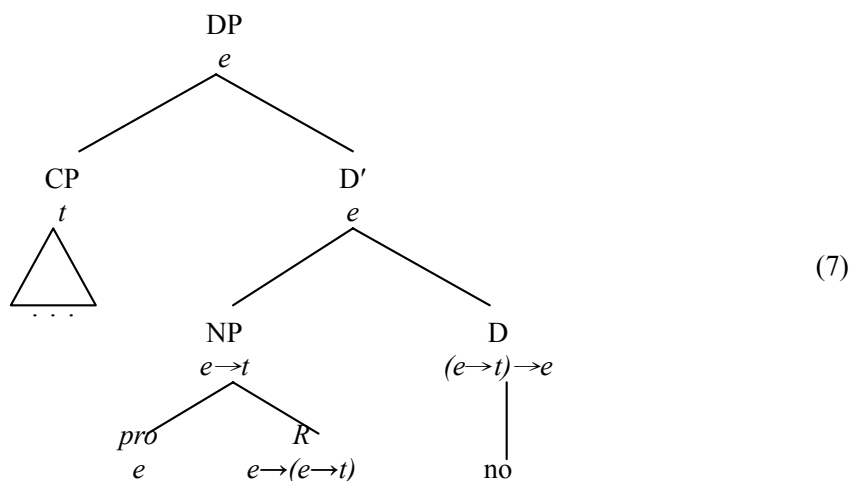
For example, in the LF representation (5) of the IHRC in (1), the empty pronoun is represented as  $P$ , a free variable of type  $e \rightarrow t$ .



Now, it is assumed that free variables in general are assigned a value by the assignment function  $g$  relative to the context  $c$ . In this case,  $P$  is assigned the value in (6): the set of apples which Hanako peeled.

$$\llbracket P \rrbracket_{g_c} = \lambda x. \text{apple}'x \wedge \text{peel}'x \text{ hanako}' \quad (6)$$

Shimoyama assumes that the nominalizer *no* is interpreted as the function from a set to the maximality of the set, adopting Link's [1983] analysis of definite descriptions.<sup>2</sup> As a



<sup>2</sup> Note in passing that our analysis in section 3 assumes that plurals are translated as set individuals.

## A Skolem Term Based Approach

result, the denotation of IHRC is equivalent to the English definite description *the apples that Hanako peeled*. In the case of (3), the LF representation of the IHRC is something like (7), in which the empty pronoun is further divided into *pro* and a free variable *R*.

*pro* is a variable of type *e* bound by the universal quantifier in the matrix clause. On the other hand, *R* is a free variable of type  $e \rightarrow (e \rightarrow t)$ , to which an assignment function  $g_c$  assigns the value in (8) as a salient two-place predicate in the context.

$$\llbracket R \rrbracket_{g_c} = \lambda x. \lambda y. \text{paper}'y \wedge \text{wrote}'yx \quad (8)$$

The latter takes *pro* as its argument to yield the interpretation shown in (9).

$$\llbracket R \rrbracket_{g_c} \left( \llbracket pro \rrbracket_{g_c} \right) = \lambda y. \text{paper}'y \wedge \text{wrote}'yx \quad (9)$$

In words, this is the set of *y* such that *y* is a set of papers that *x* wrote, where *x* is bound by the universal quantifier. Then the result serves as the argument for *no* as before. In the end, (3) is interpreted as ‘for every student *x*, *x* wrote three papers this semester, and this morning *x* turned in the papers *x* wrote this semester.’ We thus get correct semantics. However, this analysis poses two problems. First, as is noted by Shimoyama herself, the assignment function is not properly constrained. It just picks up a property or relation which is ‘salient’ in the context. We will discuss this problem in section 2.2. We will assume that the appropriate constraints can be captured by the predicate *result'* or *abt'* although we will not be determinate on how to decide between these two options. The second problem is the ambiguity in the semantic type of the free variable. As we have just seen, the assignment function  $g_c$  assigns a value to a free variable relative to the context. In order to assign a value, at least the semantic type of the variable needs to be known. However, the semantic type of that free variable can be determined only after the context is available. Then, the problem is that it is unclear as to how the context can be available before the context-dependent interpretation comes in. Furthermore, the semantic type of the free variable can be arbitrarily complex according to the number of universal quantifiers in the matrix clause. This is illustrated in (10-11):

Dono sensei-mo subeteno zyugyoo-de [menomae-de	gakusei-ga	neteiru ]	
every prof	all class-LOC	before:eyes-LOC	student-NOM sleeping
no-o	tatakiokosita		(10)
NML-ACC	woke:up:roughly		
‘Every prof, in all his classes, woke up a student who is sleeping before his eyes.’			

Dono bando-no dono gitarisuto-mo subeteno suteezi-de dono kyoku-demo  
 Every band-GEN every guitarist all stages-LOC every song-LOC  
 [gitaa-no tyuuningu-ga kurutteiru] no-o sonobade tatakikowasita (11)  
 guitar-GEN tuning-NOM wrong NML-ACC on:the:spot broke  
 ‘Every guitarist of every band smashed the guitar which was out of tune in every song  
 on all stages.’

In principle, there is no limit to the number of quantifiers in the matrix clause. Therefore, a mechanism that does not invoke complexity of this sort would be preferred. In section 3, we propose an analysis that can derive the interpretation of such multiply quantified IHRC in the same way as the quantifier-free cases like (1) and singly quantified cases like (2) by introducing the concept of the generalized Skolem term proposed in Steedman [2005].

## 2.2 The Interpretational Characteristics of IHRC

In order to state a proper constraint on the possible antecedent of an IHRC, let us examine the semantics of IHRC in more detail. The anaphoric nature of IHRC is best illustrated by the fact that the antecedent of an IHRC is occasionally not expressed as a linguistic element. (12-14) are such examples of ‘headless’ IHRC from Nomura [2000].<sup>3</sup>

[Nikai-de suisoo-ga ahureta ] no-ga sita-ni  
 second:floor-LOC fish:tank-NOM overflowed NML-NOM downstairs-ALL  
 morete hita (12)  
 leak come  
 ‘The fish tank upstairs overflowed and (the water) leaked to downstairs.’

[Kesa kao-o sotta ] no-ga yuugata-niwa mata nobite kita (13)  
 this:morning face-ACC ahaved NML-NOM evening-TOP again growing came  
 ‘I shaved my face in the morning, and (the beard) started to grow again in the evening.’

[Tuti-o hotta] no-o ue-kara nozokikonda (14)  
 soil-ACC dug NML-NOM up-from looked:into  
 ‘I dug the soil, and looked into (the hole).’

<sup>3</sup> In the translation, antecedents are given in brackets.



## A Skolem Term Based Approach

Determining the antecedent of a headless IHRC obviously requires inference of some sort. Kikuta [2000] addresses the question of exactly how much information is needed for this inference within the framework of Generative Lexicon (GL). To recapitulate Kikuta's argument, if the inference always requires unconstrained pragmatics, the theory would overgenerate IHRCs with an illicit antecedent. Accordingly, there should be a more restricted way for determining the antecedent. And indeed she shows that the antecedent can be identified only by the linguistically specified information. Kikuta's conclusion is that the possible antecedent of the headless IHRC must meet the following conditions: (i) necessary involvement in the event described by the main predicate in the IHRC; (ii) recoverability of the denotation from linguistically pre-specified information; and (iii) presence in the resultant state of the process described by the main predicate in the IHRC. Note that for example (12), the antecedent is the water which is necessarily involved in the overflowing event, lexically specified by the predicate *ahureta* 'overflowed', and exists in the resultant state. Other examples can also be shown to satisfy these conditions. In contrast, violating these conditions leads to unacceptability. Following Kikuta's result (but simplifying the matter somewhat), in our analysis we will use a predicate *result'* which is defined such that *result' px* means that an element *x* is involved in the resultant state of the process or event described by the proposition *p*. However, the three conditions given above cannot straightforwardly be applied to the IHRCs in general. Consider (15), the example taken from a news article (asahi.com, June 22, 2004):

[Denwa-o kawatta betuno otoko-ga nakizyakutteita] no-o syuhu-wa  
 phone-ACC got the:other man-NOM sobbing NML-ACC housewife-TOP  
 kaisyain-no otto da to omoikonda. (15)

office:worker-GEN husband COP COMP believed

'The housewife took the other man sobbing on the phone as her husband, an office worker.'

The main predicate of the IHRC is *nakizyakutteita* 'sobbing', which denotes an activity, apparently lacks a lexical specification of the resultant (or consequent) state. Therefore, there is no way to satisfy condition (iii). And yet the antecedent *betuno otoko* 'the other man' is an entity necessarily involved in the described event, in accordance with the above condition (i). In such cases, we will use a predicate *abt'* for 'aboutness' relation defined informally such that *abt' px* means that an element *x* is necessarily involved in the event described by the proposition *p*. Now we have two distinct relations (*i.e.* *result'* and *abt'*) to describe the semantic constraint on the possible antecedent of IHRCs. Note that *result'* is a subtype of *abt'*

since the former entails the latter. We will define the nominalizer category as initially having the interpretation containing the predicate *abt'*, which is on occasion replaced with *result'*. In the following, we give an informal sketch of how this replacement is done. However, in the analysis in section 3, we will take this replacement for granted, treating *result'* as if it is given lexically. Given the fact that an IHRC is syntactically a complete sentence, we can assume that an IHRC constitutes a separate information unit from the matrix clause. And we adopt the claim of Segmented Discourse Representation Theory [Asher and Lascarides 2003] that every information unit or proposition is connected via some rhetorical relation in order for the entire discourse to be coherent. Then the IHRC and the matrix clause must be connected via some rhetorical relation. By way of illustration, let us consider (12) again. Here, we have two clauses, namely the IHRC and the matrix clause, and the two corresponding propositions: the fish tank upstairs overflowed ( $\pi_1$ ), and *something* leaked to downstairs ( $\pi_2$ ). In the latter proposition, *something* corresponds to the nominative argument IHRC. Being anaphoric, this must be resolved in some way. In addition, the two propositions need to be connected via some rhetorical relation. Suppose that the latter requirement is somehow fulfilled by inferring *Result*( $\pi_1, \pi_2$ ) as the relevant rhetorical relation. Then the semantics of *Result* entails that the event of  $\pi_1$  caused that of  $\pi_2$ . Now, *something* in  $\pi_2$  can plausibly be equated with the water, making the discourse (consisting of two clauses) coherent. We assume that the relation symbol *abt'* originating from the lexical information is replaced with the more specific relation symbol *result'* in the process of such inference. Note, incidentally, that this kind of approach may provide a way to account for the Relevancy Condition on IHRC discussed in Kuroda [1975-6]:

For a p.-i. relative clause [IHRC] to be acceptable, it is necessary that it be interpreted pragmatically in such a way as to be directly relevant to the pragmatic content of its matrix clause.

As the effect of the Relevancy Condition, events described by an IHRC and the matrix clause are typically related in any of the following terms: (i) temporal overlap, (ii) relevance of purpose, (iii) relevance of motivation, or (iv) spatial proximity. Note that Kuroda stated the Relevancy Condition as a mere descriptive generalization, whereas in our approach sketched above, this can be viewed as a consequence of the principle of discourse coherence in general, since if IHRC cannot be connected by some rhetorical relation, there would be no way to attach the information of IHRC to the matrix clause in a coherent manner. The effects of the Relevancy Condition mentioned above can also be regarded as entailments of the inferred rhetorical relation.

### 3. Generalized Skolem Term Analysis

#### 3.1 Generalized Skolem Term

In this section, we will introduce the concept of the generalized Skolem term proposed by Steedman [2005]. We briefly sketch the basic idea here and we will show in section 3.3 how to apply it to the analysis of (multiply quantified) IHRCs. The basic idea is this: a generalized Skolem term is a term whose denotation varies with the value of bound variables introduced by scope-taking operators such as universal quantifiers, but which is interpreted as a constant in the absence of such operators. One of the main motivations for such a mechanism is to give an analysis of the alternation of quantifier scope. By way of illustration, let us consider the sentence *Everybody loves somebody*. The narrow reading of *somebody* can be translated as

$$\forall x [ \textit{person}'x \rightarrow \exists y [ \textit{person}'y \wedge \textit{love}'yx ] ] \quad (16)$$

However, we can entirely eliminate the existential quantifier from Logical Form, by replacing the existentially quantified variables with the Skolem term  $sk'x$ .

$$\forall x [ \textit{person}'x \rightarrow ( \textit{person}'(sk'x) \wedge \textit{love}'(sk'x)x ) ] \quad (17)$$

Note that  $sk'$  here is a Skolem function, and the referent of  $sk'x$  is dependent on who the variable  $x$  refers to. On the other hand, we get the wide scope reading of *somebody* by letting the Skolem term be a constant  $sk'$ .

$$\forall x [ \textit{person}'x \rightarrow ( \textit{person}'sk' \wedge \textit{love}'sk'x ) ] \quad (18)$$

The above transformations of Logical Form illustrate the standard Skolemization. In the current framework, however, indefinite noun phrases are interpreted as Skolem terms right from the start, and the nominal properties such as  $\textit{person}'$  will be directly associated with them. More specifically, when introduced as a Logical Form element, a Skolem term is indicated as  $skolem_p$  where  $p$  designates a nominal property. At this stage, this is *unspecified* as to its arguments. At some later step in the derivation, an operation called *Skolem specification* is applied to this unspecified Skolem term to yield a *generalized Skolem term*, designated as  $sk_p^E$ , where  $E$  is the *environment*, an ordered set consisting of the bound variables of the universal quantifiers that take scope at that point of the derivation. Since  $sk_p$  is a function with  $E$  as its argument, its reference varies depending on the values of bound variables. And if  $E$  is the empty set, the  $sk_p^E$  will be a constant. The notion of environment is incorporated in the grammatical rule of Combinatory Categorical Grammar (CCG) in the following way. First, we have two function application rules:<sup>4</sup>

---

<sup>4</sup> In this paper, we use only the application rules, which (apart from the notational convention) is common to all variants of Categorical Grammars. The reason for this choice is just the simplicity of presentation. For other rules of CCG, we refer the reader to Steedman [2000].

$$XY:f' \quad Y:a' \Rightarrow X:f'a' \quad (>)$$

$$Y:a' \quad XY:f' \Rightarrow X:f'a' \quad (<)$$

The environment is the operator bound variable identifier and is associated with the propositional body of the interpretation. For example, the transitive verb *loves* has the following category with its environment being the empty set:

$$\text{loves} := (S \setminus NP_{3S}) / NP : \lambda x. \lambda y. [\text{loves}'xy]^{\{\}} \quad (19)$$

Application of the rule induces environment passing in the following way: if a function with environment  $F$  is applied to an argument with environment  $A$ , the environment of the argument in the resulting Logical Form is the union of the two ( $F \cup A$ ). We often omit the environment from the notation where it is of little interest. Let us now turn to the way this works. Expressions traditionally analyzed as an existential quantifier such as *somebody* are analyzed here as an unspecified Skolem Term *skolem'person'*. In the sentence *Everybody loves somebody*, for example, if specification applies at a point of derivation in which the Skolem term has not yet been in the scope of the universal quantifier as in (20), the resulting generalized Skolem term will be  $sk_{person'}$ , hence we get the wide scope reading of *somebody*. In the derivation shown below, Skolem specification is indicated by the dotted underline.

$$\begin{array}{ccc}
 \text{Everybody} & \text{loves} & \text{somebody} \\
 \hline
 S / (S \setminus NP_{3S}) & (S \setminus NP_{3S}) / NP & NP \\
 : \lambda f. \forall y [person'y \rightarrow fy]^{\{y\}} & : \lambda x. \lambda y. \text{love}'xy & : \text{skolem}'person' \\
 & & \dots \dots \dots \\
 & & \underline{NP : sk_{person'}} \\
 & & > \\
 & & S \setminus NP_{3S} : \lambda y. \text{love}'sk_{person'y} > \\
 \hline
 S : \forall y [person'y \rightarrow \text{love}'sk_{person'y}]^{\{y\}} & & >
 \end{array} \quad (20)$$

In contrast, if specification applies after it enters within the scope of the universal quantifier, the resulting generalized Skolem term is  $sk_{person'}^{(y)}$ , where  $sk_{person'}$  is a function which takes the variable  $y$  bound by the universal quantifier as its argument, hence the narrow scope reading of *somebody*. The following illustrates the derivation of this reading.

$$\begin{array}{ccc}
 \text{Everybody} & \text{loves} & \text{somebody} \\
 \hline
 S / (S \setminus NP_{3S}) & (S \setminus NP_{3S}) / NP & NP \\
 : \lambda f. \forall y [person'y \rightarrow fy]^{\{y\}} & : \lambda x. \lambda y. \text{love}'xy & : \text{skolem}'person' \\
 & & \hline
 & & S \setminus NP_{3S} : \lambda y. \text{love}'(\text{skolem}'person')y > \\
 \hline
 S : \forall y [person'y \rightarrow \text{love}'(\text{skolem}'person')y]^{\{y\}} & & > \\
 \dots \dots \dots & & \\
 S : \forall y [person'y \rightarrow \text{love}'sk_{person'}^{(y)}y]^{\{y\}} & & 
 \end{array} \quad (21)$$

A Skolem Term Based Approach

The scope alternation is also observed if the universal quantifier noun phrase appears in the object position. This is accounted for in a quite similar way to that shown above. The derivation below illustrates the inverse scope reading.

$$\begin{array}{c}
 \begin{array}{ccc}
 \text{Somebody} & \text{loves} & \text{everybody} \\
 \hline
 NP_{3S} & (S \setminus NP_{3S}) / NP & (S \setminus NP_{agr}) \setminus ((S \setminus NP_{agr}) / NP) \\
 : skolem'person' & : \lambda x. \lambda y. love'xy & : \lambda f. \lambda x. \forall y [person'y \rightarrow fyx]^{y\} \\
 \hline
 & S \setminus NP_{3S} : \lambda x. \forall y [person'y \rightarrow love'yx]^{y\} & < \\
 \hline
 S : \forall y [person'y \rightarrow love'y(skolem'person')]^{y\} & & < \\
 \hline
 \dots\dots\dots \\
 S : \forall y [person'y \rightarrow love'y sk_{person'}^{(y)}]^{y\} & & 
 \end{array}
 \end{array}
 \tag{22}$$

3.2 Distributivizing Verb Category in Japanese

Before going into the Skolem term approach to IHRC, let us consider how the scope alternation is achieved in Japanese. One striking fact about scope alternation in Japanese is that its availability is more restricted than in English. While in English the universal quantifier in the object position can take the inverse scope over the subject indefinite as in *Somebody loves everybody*, the Japanese counterpart does not seem to accept such reading. In (23), we observe that the Skolem term *dareka* ‘somebody’ allows only the narrow scope reading (cf. Nakamura 1993):

$$\begin{array}{l}
 \text{Dareka-ga} \quad \text{daremo-o} \quad \text{aisiteiru.} \\
 \text{Somebody-NOM everybody-ACC love} \\
 \text{‘Somebody loves everybody’}
 \end{array}
 \tag{23}$$

However, if the object NP is scrambled to the sentence initial position, and precedes the subject indefinite *dareka* ‘somebody’, both narrow and wide scope readings are available.

$$\begin{array}{l}
 \text{Daremo-o} \quad \text{dareka-ga} \quad \text{aisiteiru.} \\
 \text{Everybody-ACC somebody-NOM love} \\
 \text{‘Somebody loves everybody’}
 \end{array}
 \tag{24}$$

This situation is quite unlike English examples. To accommodate this fact, we tentatively adopt Steedman’s suggestion (2005:54, fn.51) that languages such as Japanese entirely lack generalized quantifier NPs and that the work of the universal quantifier in such languages are

done by distributivizing verb categories. We assume that *daremo* ‘everybody’ denotes a set of individuals whose members are all the people in the universe of discourse, indicated as *all-people*’. (25) illustrates the derivation of (24), in which *dareka* ‘somebody’ is interpreted as having narrow scope.

$$\begin{array}{c}
 \begin{array}{ccc}
 \text{Daremo-o} & \text{dareka-ga} & \text{aisiteiru} \\
 \hline
 NP_{ACC} & NP_{NOM} & (S_{DIST} \setminus NP_{ACC}) \setminus NP_{NOM} \\
 : all\text{-people}' & : skolem' person' & : \lambda y. \lambda x. \forall w [w \in x \rightarrow love' wy] \{w\} \\
 \hline
 & & S_{DIST} \setminus NP_{ACC} : \lambda x. \forall w [w \in x \rightarrow love' w (skolem' person')] \{w\} \\
 \hline
 & & S_{DIST} : \forall w [w \in all\text{-people}' \rightarrow love' w (skolem' person')] \{w\} \\
 \hline
 & & \dots \\
 & & S_{DIST} : \forall w [w \in all\text{-people}' \rightarrow love' w sk_{person'}^{(w)}] \{w\}
 \end{array} \\
 & & (25)
 \end{array}$$

Here, we defined the category of the verb so that each member of the set *all-people*’ distributes over the Skolem term introduced by the subject NP. If we further assume that distributivizing as well as scrambling is realized by a lexical rule, then the fact that distributive reading is absent in a non-scrambled sentence can be thought of an accidental lack of such lexical rule. The derivation (26) illustrates the only possible reading for (23).

$$\begin{array}{c}
 \begin{array}{ccc}
 \text{Dareka-ga} & \text{daremo-o} & \text{aisiteiru} \\
 \hline
 NP_{NOM} & NP_{ACC} & (S \setminus NP_{NOM}) \setminus NP_{ACC} \\
 : skolem' person' & : all\text{-people}' & : \lambda x. \lambda y. love' xy \\
 \hline
 \dots & & S \setminus NP_{NOM} : \lambda y. love' all\text{-people}' y \\
 NP_{NOM} : sk_{person}' & & \hline
 S : love' all\text{-people}' sk_{person}'
 \end{array} \\
 & & (26)
 \end{array}$$

### 3.3 IHRC as Generalized Skolem Term

The current analysis views the interpretation of IHRC as an instance of generalized Skolem term, and provides a straightforward account for the semantics of the construction. In order to capture the restriction on the interpretation of IHRC, the nominalizer category is defined as in (27), where *R* stands for either *abt*’ or *result*’ (see section 2.2 for the discussion of this treatment):

$$no := NP \setminus S : \lambda p. skolem' (\lambda x. Rpx) \quad (27)$$

The derivation of (1) is shown in (28).

$$\begin{array}{c}
 \begin{array}{ccc}
 \text{Taroo-ga} & \text{Hanako-ga ringo-o muita} & \text{no-o} & \text{tabeta} \\
 \hline
 NP : taro' & S : peel' apple' hanako' & NP \setminus S : \lambda p. skolem' \lambda x. result' px & (S \setminus NP) \setminus NP : \lambda y. \lambda z. eat' yz \\
 \hline
 & & NP : skolem' (\lambda x. result' (peel' apple' hanako') x) & \\
 \hline
 & & S \setminus NP : \lambda z. eat' skolem' (\lambda x. result' (peel' apple' hanako') x) z & \\
 \hline
 & & S : eat' skolem' (\lambda x. result' (peel' apple' hanako') x) taro' & \\
 \hline
 & & \dots & \\
 & & S : eat' sk_{\lambda x. result' (peel' apple' hanako') x} taro' &
 \end{array} \\
 & & (28)
 \end{array}$$

## A Skolem Term Based Approach

The semantics of the embedded clause is fairly obvious. This is then applied to the function defined in (27). The ‘result’ predicate  $result'px$  indicates that the event described by  $p$  yields an individual  $x$ . Thus the IHRC is analyzed as a Skolem term with the property of being involved in the result state of Hanako’s peeling an apple. It is further subject to Skolem specification (indicated by the dotted line as before). Since there is no universal quantifier, the resulting generalized Skolem term is a constant whose referent is the apple that Hanako peeled. It should be obvious that this analysis also accommodates to the examples of headless IHRCs discussed in section 2.2. Interpretation of an IHRC is always determined via the *abt'* or *result'* relation, regardless of whether the antecedent is explicitly expressed or not. Still, we have to admit that the nominal property of the Skolem term  $\lambda x. result'(peel'apple'hanako')x$  is not a sufficient characterization of the antecedent, as it can also be applied to the apple skin, rather than only to the peeled apple fruit. Basically this  $\lambda$ -expression should function as a constraint on the possible referent of the IHRC, and it is subject to the process of anaphora resolution. The resolved Skolem term in this case would be  $skolem \lambda x.(result'(peel'apple'hanako')x \wedge fruit'x)$ . However, we will gloss over this problem here, and the question of how this mechanism works is left open for future research. Now, let us examine the interpretation of the IHRC that occurs within the scope of a universal quantifier. The derivation of (3) is shown in (29):

$$\begin{array}{c}
 \text{Dono gakusei-mo} \quad \text{soitu-ga peepaa-o san-bon kaita no-o} \quad \text{teisyutusita} \\
 \hline
 \text{NP} \quad \text{NP} \quad (SDIST \setminus NP) \setminus NP \\
 : all-student' \quad : skolem'(\lambda x.result'(write'3papers'(pro'u))x) \quad : \lambda z.\lambda w.\forall u[u \in w \rightarrow turn-in'zu]^{u\} \\
 \hline
 SDIST \setminus NP : \lambda w.\forall u[u \in w \rightarrow turn-in'(skolem'(\lambda x.result'(write'3papers'(pro'u))x))u]^{u\} \quad (29) \\
 \hline
 SDIST : \forall u[u \in all-student' \rightarrow turn-in'(skolem'(\lambda x.result'(write'3papers'(pro'u))x))u]^{u\} \\
 \dots\dots\dots \\
 SDIST : \forall u[u \in all-student' \rightarrow turn-in'sk_{\lambda x.result'(write'3papers'(pro'u))x}^{(u)}]^{u\}
 \end{array}$$

Here, *soitu* in the IHRC is interpreted as *pro'u*. As the notation may suggest, it is introduced as an ordinary pronoun. This will later get bound by the universal quantifier.<sup>5</sup> The last line is the result of Skolem specification. Since the Skolem term is in the scope of the universal quantifier, it takes the bound variable  $u$  as its argument. Therefore, the resulting generalized Skolem term refers to different sets of term papers, according to the value of  $u$ . This is the desired result, and this is achieved without invoking any complexity such as the type ambiguity of the empty pronoun. Multiply quantified cases like (10) can be derived in a similar way:

<sup>5</sup> We take this process as anaphora resolution of the usual kind. This is because *soitu* could also be interpreted as deictic pronoun, in which case there is a ghostwriter (referred to by *soitu*) who wrote all the papers that every student turned in (a highly implausible situation, though).

$$\begin{array}{c}
\begin{array}{cccc}
\text{Dono sensei-mo} & \text{subeteno zyugyoo-de} & \text{gakusei-ga neteiru no-o} & \text{tatakiokosita} \\
\hline
NP & VP/VP & NP & (S_{DIST} \setminus NP) \setminus NP \\
: all-prof & : \lambda f. \lambda y. \forall t [t \in all-class' \\
& \quad \rightarrow during't(fy)]^{t\}} & : skolem'(\lambda x.abt'(sleep'student')x) & : \lambda z. \lambda w. \forall u [u \in w \\
& & & \quad \rightarrow wake'zu]^{u\}} \\
\hline
& & & S_{DIST} \setminus NP \\
& & & : \lambda w. \forall u [u \in w \\
& & & \quad \rightarrow wake'(skolem'(\lambda x.abt'(sleep'student')x))u]^{t,u\}} \\
\hline
& & & S_{DIST} \setminus NP \\
& & & : \lambda y. \forall t [t \in all-class' \\
& & & \quad \rightarrow during't(\forall u [u \in y \rightarrow wake'(skolem'(\lambda x.abt'(sleep'student')x))u]^{t,u\})]^{t\}} \\
\hline
& & & S_{DIST} \\
& & & : \forall t [t \in all-class' \rightarrow during't(\forall u [u \in all-prof \rightarrow wake'(skolem'(\lambda x.abt'(sleep'student')x))u]^{t,u\})]^{t\}} \\
\hline
& & & S_{DIST} \\
& & & : \forall t [t \in all-class' \rightarrow during't(\forall u [u \in all-prof \rightarrow wake'(skolem'(\lambda x.abt'(sleep'student')x))u]^{t,u\})]^{t\}} \\
\hline
\end{array}
\end{array}
\tag{30}$$

Since Skolem term automatically takes the bound variables of the environment, the problem of type ambiguity that the E-type analysis suffered does not arise.

#### 4. Conclusion

We developed an analysis of Internally Headed Relative Clause (IHRC) construction in Japanese within the framework of Combinatory Categorical Grammar [Steedman 2000]. In section 2, we first looked at the argument made by Shimoyama [1999] and her E-type analysis of IHRC. We addressed the problem of type ambiguity that her E-type analysis would raise focusing on multiply quantified IHRCs. We then discussed the interpretational characteristics of IHRC, drawing on Kikuta's [2000] study of headless IHRCs. We generalized her idea by partly adopting Asher and Lascarides's [2003] theory, and argued that semantics of rhetorical relation would also help to determine the antecedent. And we also suggested that Kuroda's [1975-6] Relevancy Condition on IHRC can be viewed as a consequence of the principle of discourse coherence in general. In section 3 we introduced the notion of generalized Skolem term, and discussed the problem of scope alternation in Japanese, which is quite different from English, and motivated the distributivizing verb category adopting the Steedman's [2005] suggestion. Finally we integrated the whole discussion into the analysis of IHRCs. The main point is that, if we take the interpretation of an IHRC as a generalized Skolem term, we can attain the uniformed analysis for the several kinds of IHRCs, namely, headless IHRC, simple (quantifier-free) IHRC, and (singly or multiply) quantified IHRC. Our focus here was exclusively on the IHRC, but of course the approach here also applies to other kinds of noun phrases. However, the theoretical implication of this approach in other nominal construction is not entirely clear at this point, and is left for future work.



### **Acknowledgements**

The research reported here was supported in part by the Tohoku University 21st Century Center of Excellence (COE) Program in Humanities, Research and Strategic Center for an Integrated Approach to Language, Brain, and Cognition (<http://www.lbc21.jp/>). We are grateful to two anonymous reviewers for their comments.

### **References**

- Asher, N., and A. Lascarides, *Logics of Conversation*. Cambridge: Cambridge University Press, 2003.
- Evans, G., "Pronouns," *Linguistic Inquiry*, 11, 1980, pp. 337-362.
- Kikuta, C. U., "Qualia Structure and the Accessibility of Arguments: Japanese Internally-headed Relative Clauses with Implicit Target," *The Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, Akira Ikeya and Masahito Kawamori (eds.), 2000, pp. 153-164.
- Kuroda, S.-Y., "Pivot-Independent Relativization in Japanese II," 1975–1976, Reprinted in S.-Y. Kuroda, *Japanese Syntax and Semantics*. Dordrecht: Kluwer, 1992.
- Link, G., "The Logical Analysis of Plurals and Mass Terms: A Lattice-theoretical Approach," in *Meaning, Use, and Interpretation of Language*, R. Bäuerle et al. (eds.), de Gruyter, Berlin, 1983, pp. 302-323.
- Nakamura, M., "Scrambling and Scope in Japanese." In Patricia M. Clancy (ed.), *Japanese/Korean Linguistics*, vol. 2, 1993, pp. 283–298. Stanford Linguistics Association, Stanford CA: CSLI.
- Nomura, M., *The Internally-Headed Relative Clause Construction in Japanese: A Cognitive Grammar Approach*. Ph.D. dissertation, University of California, San Diego, 2000.
- Shimoyama, J., "Internally Headed Relative Clauses in Japanese and E-type Anaphora," *Journal of East Asian Linguistics* 8, 1999, pp. 147-182.
- Steedman, M., *The Syntactic Process*. Cambridge, MA: MIT Press, 2000.
- Steedman, M., *Surface-Compositional Scope-Alternation without Existential Quantifiers*. Draft, 2005. <http://www.iccs.informatics.ed.ac.uk/~steedman/papers.html>



## Data Management in QRLex, an Online Aid System for Volunteer Translators'

Youcef Bey<sup>+</sup>, Kyo Kageura<sup>+</sup>, and Christian Boitet<sup>\*</sup>

### Abstract

This paper proposes a new framework for a system which will help online volunteers to perform translations on their PCs while sharing resources and tools and communicating via websites. The current status of such online volunteer translators and their translation practices and tools are examined, along with related work also being discussed. General requirements are derived from these considerations. The approach taken in this study for dealing with heterogeneous linguistic resources relies on an XML structure maximizing efficiency and enabling all of the desired functionalities. The QRLex environment is under development and implements this new framework.

**Keywords:** Computer-Aided Translation, Web Search for Translation, Memory Translation, Helping Volunteer Translators, Linguistique Ressources.

### 1. Introduction

There have been many misconceptions concerning Machine Translation. In the early days, some researchers promised to "replace translators", while others like Bar-Hillel warned against the impossibility of FAHQMT (Fully Automatic High Quality Machine Translation) *in general*. The famous ALPAC report negatively evaluated the performance of Machine Translation (MT) systems at the end of 1966<sup>1</sup>. It is also known as the "infamous" ALPAC

---

<sup>\*</sup> Laboratoire CLIPS-GETA-IMAG, Université Joseph Fourier, 385, rue de la Bibliothèque, Grenoble, France

E-mail : {youcef.bey; christian.boitet}@imag.fr

<sup>+</sup> Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

E-mail: kyo@p.u-tokyo.ac.jp

<sup>1</sup> As a matter of fact, the ALPAC committee worked on obsolete and incorrect data, and did not even investigate a significant project in the same city (Washington D.C.), the GAT (Georgetown Automatic Translation) project. Further, several members of the committee were themselves heads of labs that received funds to work on MT. These members preferred to work on theoretical linguistics and AI (as it later became known) instead of the necessary engineering.

report because it was biased, as explained in detail in the no less famous counter-report by Zbigniew Pankowicz, the polyglot USAF official who oversaw MT funding at RADC (Rome Air Development Center) from the early days until 1985. The truth, as recognized by Bar-Hillel in 1972 at a seminar on the usability of MT organized at Austin, Texas<sup>2</sup>, is that MT can be quite useful in practice even if the "translational quality" is medium or poor. Later, when MT became widely available (first at the European Community, then on the French Minitel, then on PCs, and finally on the Web), it became clear that commercial MT, *if properly used as a tool to help translators by offering them a kind of "pretranslation"*, can be used efficiently as "MT for translators". According to [Allen 2001] and according to experiments presented on [MT POST-EDITING 2006], the productivity of translators is often multiplied by three, for a variety of tasks.

Nevertheless, it was, and still is, true that MT has three conflicting goals that cannot be achieved together: full automaticity, high quality, and general coverage. However, two of these goals can indeed be achieved together. For example, the METEO system<sup>3</sup> showed decisively that an MT system specialized to an "adequate" sublanguage, in that case the language of weather bulletins (as opposed to weather situations or warnings), can produce better translations than the humans previously employed. Indeed, before 1980—1985, it took about 5—10 minutes to post-edit a bulletin translated by a junior translator, while it has taken only one minute from 1985 on, when METEO reached its top quality. Here, MT is really "MT for revisers" in that MT output can be post-edited without reference to the source text.

By contrast, wide coverage fully automatic MT systems cannot be used in this way at all. Because of unsolved ambiguities, the number of possible valid translations with very different meanings (including nonsensical ones) is extremely high, and it is not feasible to show them all to the post-editor. Exactly the *same* systems that can be very useful to bilingual post-editors are useless to monolingual post-editors. Martin Kay wrote that:

*"...this happens when the attempt is made to mechanize the non-mechanical or something whose mechanistic substructure science has not yet been revealed..."* [Kay 1997]

Yes, but that is not really the point! The point is that the same problem arises with human

---

<sup>2</sup> One of the few research groups really working on MT that continued to be funded after 1966 in the US, such as that of Pr Wang (University of California, Berkeley), and the newly founded Systran and Logos.

<sup>3</sup> That is the name of the operational system, which was further developed, improved and deployed on PCs by John Chandioux and his team, starting from the TAUM-METEO prototype built by the TAUM group at Université de Montréal around 1975-76.

translators: if they are asked to translate texts far out of their domain, they also produce incomprehensible results, impossible or very difficult for monolingual domain specialists to post-edit. Their errors are different, and their translations are more grammatical, but the time required to obtain polished translations based on their draft (or to decide that they are too poor for that and the text must be retranslated by a specialist) is on the same order.

In any case, this perception of MT has promoted research on *computer aided human translation*, which exploits the potential of computers to support translator skills and intelligence [Hutchins 1998]. Many industries have made large investments in developing useful translation-aid tools. These efforts have resulted in commercial Computer-Aided Translation (CAT) systems such as TM-2 (IBM), Trados, Déjà Vu, Transit, and Similis, which usually contain three components:

- *bilingual editors* (often embedded in text or document processors such as Word, WordPerfect, Ichitarou, Interleaf, etc.),
- *on-line terminology banks and dictionaries* (the latter being modifiable by translators and immediately updated), and
- *translation memory systems* (TM), which seek exact or fuzzy matches of the source segments to retrieve their translations as proposed translation [Bowker 2002].

There are two situations in professional translation. In the case of large and repetitive translation jobs such as successive versions of a product documentation, TM is quite useful, and MT is not used, even in the rare cases where its integration is foreseen, as in TM-2: the distance between users and developers is too great, so that MT dictionaries, especially for terminology, are not updated fast enough from the translators' dictionaries, whereas TM grows and becomes increasingly useful [Boitet 2005].

In the case of individual translators working on a variety of jobs, TM is not really useful, because the quantity of past translations of similar texts is too small. Accordingly, cheap commercial MT is used to obtain preliminary translation. (Even if no translations of complete segments are correct, many fragments are correct; hence MT functions as a kind of *dictionary in context*). Relatively few individual translators use commercial CAT tools anyway because of their high price.

What is the situation for online volunteer translators? Here again, there are two cases. In the first case, to which the QRLex system is addressed, translators are online in that they access a website to get documents to translate, retrieve resources such as dictionaries and TMs, deposit finished translations, and communicate with other translators. However, they don't translate online. Rather, they work on their PCs (or PDAs), just like many professional translators. However, they cannot afford to use commercial PC-oriented CAT tools, and, until now, they have not benefited from shared resources as do their professional counterparts.

In the second and more recently encountered case, volunteer translators do translate online, as on Translationwiki [TRANSLATIONWIKI 2006]. The documents to be translated are automatically segmented (paragraphs, sentences) and put up for translation. No CAT functions or resources are available.

In all cases, the CAT tools and resources available, if any, do not provide content and functions that fully satisfy all translators. There is thus a real need to aid online volunteer translators and their communities by providing them with a free environment with a rich set of linguistic resources and tools, and improved workflow and data management.

Recently, the number of volunteer translators has been growing sharply. Volunteers form or join communities, and they translate thousands of documents in different fields, thereby showing the true way to break the language barrier. These developments are mainly due to the Internet's crucial role in allowing translators to take part in such volunteer translation activities.

According to our study, volunteer translator communities are mainly of two types:

- *Mission-oriented translator communities*: strongly-coordinated groups of volunteers involved in translating clearly defined sets of documents (Linux-like communities). These communities translate what can be loosely called technical documentation, such as Linux documentation [TRADUC 2005], W3C specifications, and documentation as well as software (interface, messages, online help) of open source products. For example, in the W3C consortium, 301 volunteer translators are involved in translating thousands of specification documents into approximately 41 languages [W3C 2005]. Documentation in the Mozilla project exists in 70 languages, and is translated by hundreds of volunteer translators located in different countries [MOZILLA 2005].
- Network communities of *subject-oriented translators*: individual translators who translate online documents such as news, analysis, and reports and make translation available on personal or group Web pages [TEANOTWEAR 2005] [PAXHUMANA 2006].

These translators are often involved in non-identified projects. They form translator groups with no *a priori* orientation, but they share similar opinions about events (anti-war humanitarian communities, translation of reports, news translation, humanitarian help, etc.).

In the following, the state of current online volunteer translation is first reviewed and related work intended to develop online computer aided translation tools according to the needs of online translator is presented. Then, several XML standards that are key components in the design of QRLex, and which solve the problem of managing heterogeneous data (such

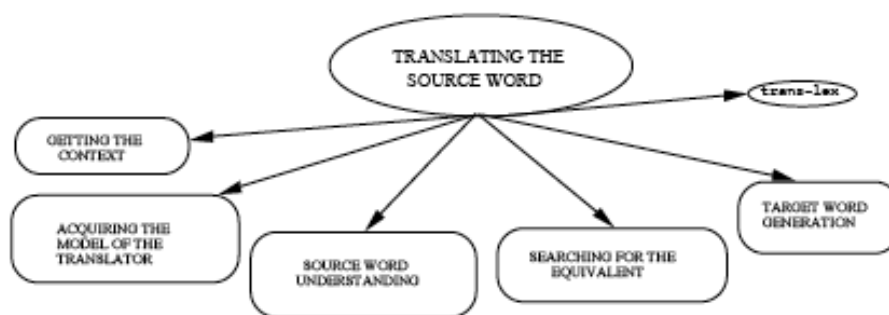
as dictionaries, TMs, documents retrieved from the Web) are introduced. Finally, the results of the first two sections are used to justify the general architecture and the main features of the QRLex system, and the project's current status is presented.

## 2. Current Situation and Related Work

This paper will now review the existing translation environments and online tools designed to help online volunteer translators. In the first two subsections, the Lexical Knowledge Bases (LKB) by [Agirre *et al.* 2000] and an online Translationwiki system [Augar *et al.* 2004] [Schwartz 2004] [TRANSLATIONWIKI 2006] are presented. In the third subsection, a method for implementing a translation workflow demonstrating the usefulness of XML standards for managing the translation of documents and associated linguistic data is outlined.

### 2.1 Translator-Oriented Dictionary Systems

Back in 1994, X. Agirre and his team proposed developing Lexical Knowledge Bases (LKB) based on a model of "dictionary-use" by human translators [Agirre *et al.* 1994] [Agirre *et al.* 2000]. However, the structure of their LKB necessitates a complex transformation from existing dictionaries, implying, in turn, very heavy human labor. By contrast, this project wants to avoid intensive human work and simply give direct access in a uniform way to existing dictionaries, lexicons and term banks (Figure 1).



**Figure 1. First level of the decomposition diagram of the tasks involved in the lexical translation process**

First, they construct a monolingual model from monolingual French and Basque dictionaries. Then they develop a new French-Basque bilingual model. Two levels on top of the monolingual dictionaries have been proposed, which allow the establishment of links between the monolingual entities and facilitating translation from one language into the other.

The LKB is designed as an "active tool", which means that the dictionary tool works autonomously to present translators with potentially useful information and user functionalities during the translation process. The concept was inherited from [Martin 1990]

who stated:

*"...the use of dictionary can be seen as a typical problem-solving activity, and user-orientation should involve both static and dynamic features of the intended user..."*

[Agirre et al. 2000] add:

*"Furthermore, along with the usual information about the meaning of the entries, dictionaries should show how to use words in context. In other words, we advocate that dictionaries should actively co-operate in finding the correct translation."*

They emphasize that such LKBs are useful only if the translators are involved in the design and development of their functionalities. Speaking of their previous work (done in 1994), they say:

*"The LKB provides various access possibilities to data. Even so, limitations are present when trying to exploit this knowledge in a lexical translation context. The cause of this limited usability is that the lexical organization was designed from a general perspective, without taking into consideration functional aspects. Incorporating this functionality means, in our case, transforming such LKB into a user-oriented dictionary system."*

The study of translators' behavior during the translation process prompts one to take several important points into account when designing dictionary systems for translators:

- Expert and occasional translators need distinct and adapted sorts of help.
- Some translators (especially occasional ones) find bilingual dictionaries very useful.
- Multi-word terms are a source of failure when using normal dictionaries.
- Context is important when translating a text.
- Dictionaries for translation must give grammatical and usage information.
- The proximity between languages is helpful, but attention must be paid to "false friends"; dictionaries must prevent translation errors derived from them.

The approach followed for developing the LKB is very pertinent, especially because the model, behavior and needs of translators are taken in consideration. Integrating information related to translators' behavior in parallel with linguistic data is a new and promising direction for the design of future CAT tools. This idea should be extended to the integration in a CAT



environment of document and translation workflow management tools together with language-oriented resources and functionalities.

## 2.2 Online Collaborative Wiki-Based Translation Environment

Translationwiki.net is an online collaborative translation Web service [TRANSLATIONWIKI 2006]. It is based on a Wiki technology which allows translators/users to collaborate and share knowledge on the Web (Figure 2). There are several steps:

- choice of source documents, normalization of format and character encoding;
- automatic segmentation into translation units (TUs), which are paragraphs or if possible; sentences (see);
- translation proper;
- dissemination of translations in various formats.

In the Translationwiki environment, any user can upload a document for translation. The textual content is extracted and segmented automatically to TUs. No quality checking of translation is performed by the site manager (who only manage the environment and presumably has no time and does not know the target language(s)), but if translators or readers notice vandalism, modifications by suspect sources are erased and documents can be protected or semi-protected.

During translation, translators process only one TU at a time and can not see the whole document, as each TU is actually handled as a small Wiki document. Hence, translators need to navigate through the documents to check coherence and avoid translating the same expression differently in different places.

home :: Arabic :: Chinese :: French :: German :: Italian		<a href="#">login</a> / <a href="#">register</a>	
Translation Wiki			
<a href="#">translation</a> > most recently updated (all languages)		<a href="#">help</a> / <a href="#">feedback</a>	
MOST RECENTLY UPDATED (ALL LANGUAGES) ARTICLES:			<a href="#">show by date added</a>
<a href="#">Nejad Calls for National Reconciliation Midst International Expectations</a>	Sep 1 05	Aljazeera.net	<a href="#">Al-Jazeera</a> <div style="width: 17%;"></div> 17%
<a href="#">الرئيس السوري يحضر القمة الأممية لأول مرة</a>	Aug 15 05	-	<a href="#">Al-Jazeera</a> <div style="width: 100%;"></div> 100%
<a href="#">نخص بهاجم صدام أثناء محاكمته في بغداد</a>	Jul 31 05	Aljazeera.net	<a href="#">Al-Jazeera</a> <div style="width: 100%;"></div> 100%
<a href="#">نخص بهاجم صدام أثناء محاكمته في بغداد</a>	Jul 30 05	Wire Sources	<a href="#">Al-Jazeera</a> <div style="width: 0%;"></div> 0%
<a href="#">中海油“世纪收购”揭基</a>	Jul 24 05	Various	<a href="#">财经</a> <div style="width: 79%;"></div> 79%
<a href="#">中国女生遭日本男同事无故殴打 施暴者认罪</a>	Jul 12 05	法制晚报	<a href="#">北京青年报</a> <div style="width: 43%;"></div> 43%

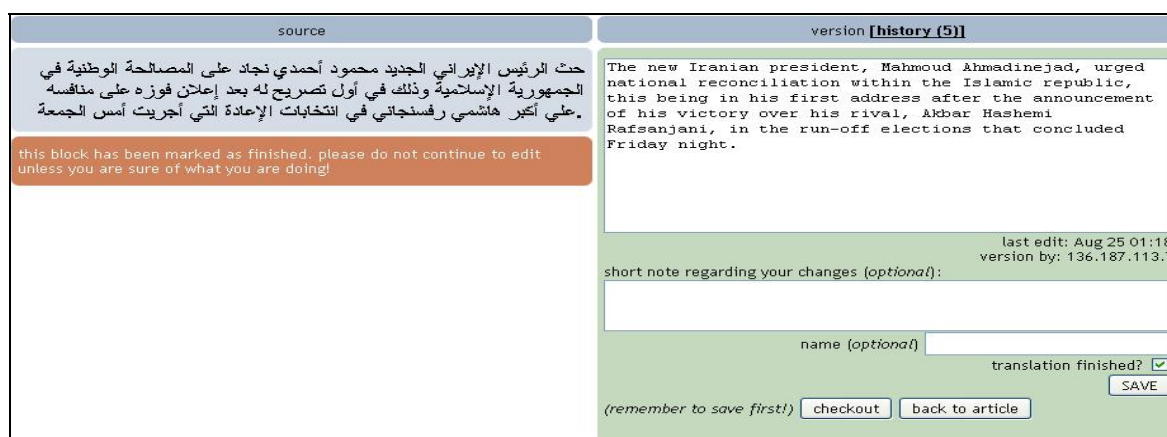
Figure 2. Main user interface of translation in translationwiki.net

Translationwiki is currently limited to five languages (Arabic, Chinese, French, German and Italian). Translators/users can sort and search documents in one language at a time. As for

the direction of translation, an uploaded document may be translated only into one of the supported languages. Translators cannot translate the same document into more than one target language in the same interface, and cannot manage the multilingual content of given document, if any (for instance, they cannot keep original fragments as citations in the translation).

### 2.2.1 Translation Methods and Interface for Editing

Documents are accessible directly from the main list. Volunteer translators are invited to select documents by clicking on their titles. A new screen appears which displays the source of the TU on the left side and an editing area for the target language on the right side. The editor is a simple text area without any formatting functionalities or linguistic aids (Figure 3).



**Figure 3. Translation editor**

### 2.2.2 Translation Units (TU) and Versioning

In this environment, the versioning module keeps the history of the modifications. This allows translators to check the evolution of a translation and avoid losing content. Translators/users can easily restore old translations deleted erroneously or by vandals. When the translation of a TU is finished, the system keeps puts its translation in its repository and allows the translators to check the differences between different translations in a user-friendly interface (In Figure 4, the red terms are the result of comparing two versions of a translation into English from Arabic. All versions are listed and can be compared pairwise.)

Metadata is attached to each modification, so that, for all versions, it is easy to determine the date of the last modification and to identify the users by their profiles. Hence, users can follow the introduction and modification of content, and can distinguish which other translators produce high quality translations and which ones don't.

Aug 25 05 01:18	Sep 1 05 04:47 (current version)
<b>previous edit</b> version by: 136.187.113.*	version by: 136.187.47.*
version	version
The new Iranian president, Mahmoud <b>Ahmadinejad</b> , urged national reconciliation within the Islamic republic, this being in his first address after the announcement of his victory over his rival, Akbar Hashemi Rafsanjani, in the run-off elections that concluded Friday night.	The new Iranian president, Mahmoud <b>Ahmadi Nejad</b> , urged national reconciliation within the Islamic republic, this being in his first address after the announcement of his victory over his rival, Akbar Hashemi Rafsanjani, in the run-off elections that concluded Friday night.
<b>edit</b>	<b>edit</b>
<b>SOURCE/VERSION</b>	
source	current version
حدث الرئيس الإيراني الجديد محمود أحمدني نجاد على المصالحة الوطنية في الجمهورية الإسلامية وذلك في أول تصريح له بعد إعلان فوزه على منافسه علي أكبر هاشمي رفسنجاني في انتخابات الإعادة التي أجريت أمس الجمعة	The new Iranian president, Mahmoud Ahmadi Nejad, urged national reconciliation within the Islamic republic, this being in his first address after the announcement of his victory over his rival, Akbar Hashemi Rafsanjani, in the run-off elections that concluded Friday night.

Figure 4. Wiki-based version management of document.

### 2.3 Translation of Open Software and Associated Documents by Volunteer Translators

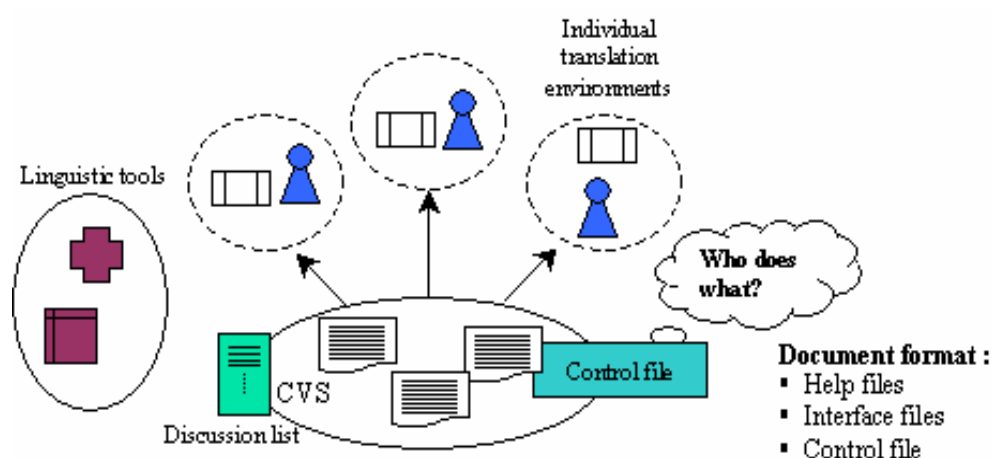
There are many projects aiming at the translation and localization of open software and associated documents. Two quite interesting projects are [MOZILLA 2005] and [TRADUC 2005]. Mozilla is a set of open software tools that includes a web navigator, an HTML page composer, and an e-mail manager. It is available in 70 languages. Translation in this project is a continuous process because each new version has new documentation and a new interface that must be translated. Two main categories of documents have to be translated by volunteer translators:

- *Interface translation*: messages are text stored in various files (Table 1). Volunteer translators download them for translation using CVS (Concurrent Versioning System).
- *Online help documents*: documents are HTML pages. They are translated at the end of the new release because they often contain screenshots which may change up to the last minute.

It is interesting that most online translators show similar behavior. In all the TRADUCT and MOZILLA localization sub-projects (one per language), volunteer translators are invited to translate a list of documents which have been put up on a website (one for each sub-project) in different formats (XML, SGML, HTML, HLP, plain text, etc.). First, they check whether the relevant document has been translated; if not, they make a reservation and announce to other translators, via a discussion list or via e-mail, that they have begun to translate it. To obtain the source document, they download it directly from the CVS (Concurrent Version System) or ask the coordinator to send it via e-mail (Figure 5).

**Table 1. Document types in the Mozilla project**

File type	Extension	Description
Module description files	RDF	XML files containing metadata (version number, language, etc.)
Interface files	DTD	XML files containing a textual part (Text for user interfaces).
Property files	Properties	Files containing messages to be displayed in dialog boxes.
HTML files	HTML	These files contain the online help.

**Figure 5. Translation method in Linux communities**

Once the document is obtained, each translator uses his or her individual translation environment, which often varies from person to person. A typical personal environment consists of a set of tools: textual editor, dictionaries (electronic, paper version, or online), glossaries, terminology, and sometimes Translation Memory (TM). In addition, the importance of the Internet should be noted, as it has become a precious linguistic resource for translators, who use it to recover existing translation segments (such as quotations, collocations, technical terms, etc.).

Documents are translated into their original format. For example, in the TRADUCT project, documents are structured in XML DocBook<sup>4</sup>; translators translate only the text which lies between XML markers. After the translation is finished, they send the whole target document in the same structure to the coordinators who transform it to readable document and disseminate it on the Web.

<sup>4</sup> A rich XML format used to produce readable HTML with OpenJade; for further information refer to <http://www.docbook.org>.

*Translation-aided tools* offered on web servers for aiding translator communities may contain quite poor linguistic resources, but may also contain some useful management-oriented facilities:

- *a set of local free dictionaries*, glossaries, and links to other linguistic web sites.
- *a discussion list* used for exchanging skills and resolving most issues faced during the translation process.
- *control files* for checking "who does what" [MOZILLA 2005]. This is useful for the collaborative translation of a given document by several volunteer translators: before starting to add to a translation, a translator finds the most recent endpoint, and starts translating from there.
- *a server for managing the document versions like the CVS server* (Concurrent Version System).

The linguistic resources are almost never maintained and updated, because of the lack of automatic tools for synchronizing modifications made off-line, and because a process and a team to validate and consolidate updates are lacking.

### **3. Elements Reusable from Current Professional Practice**

#### **3.1 Translation Workflow and XML Standards: the IBM Localization Model**

The localization process generally consists of several steps, from document creation to the final translation [IBMLOCALIZATION 2005]:

- document creation
- preparation of translation
  - normalization of format and character encoding
  - automatic or semi-automatic segmentation into TUs
- translation proper, followed by quality checking
- dissemination of translated document, possibly in various formats (PDF, HTML, etc.).

After segmentation, a document is uploaded to the server and added to the list of documents to be translated, and then is assigned to a translator. (This study's procedure is different mainly in that volunteer translators decide which documents they will translate. Otherwise, the processes are similar.)

The translator downloads the document in a textual format suitable to his or her CAT editor (TM-2 in the case of IBM), together with a "kit" containing a document-specific translation memory and a dictionary, extracted from the resources available on the server.

Each step presents some problems. For example, a good localization (e.g. a user's guide for electrical appliances, websites, slide shows, scripts for advertisements, etc.) should keep the document format intact and produce a high-quality translation. If the localization is a software element, the localization should fit nicely into the interface without causing any trouble.

In the following paragraphs, a translation workflow method exploited by IBM™ called *reverse conversion* is presented, which shows the usefulness of XML standards for managing crucial data during translation. Accordingly, these standards for QRLex data management will be adopted here.

### 3.2 Localization Methods: the Reverse Conversion Workflow

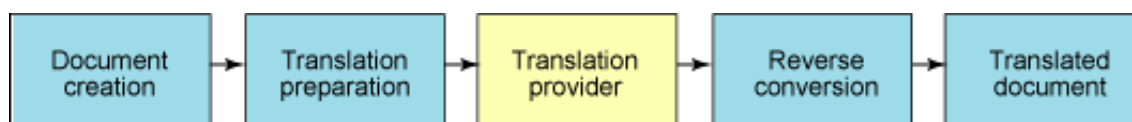
Document creation is a step performed by an individual or by an independent group. The resulting document may be in various formats, such as those of graphical user interfaces or of help and manual files.

The translation of documents containing heterogeneous data in various formats is a heavy task for translators, who must attend closely to these documents both during and after translation (during post-translation). Other difficulties are related to the duplication of content, which increases translation time, and related to the production of the final version.

Another problem is that documents need some adjustment in the post-translation stage, for example, because text length varies from language to language.

To overcome such problems at each step of the translation, the IBM teams have adopted the *reverse conversion* translation workflow.

It consists of extracting only the relevant part of the materials to be translated, and putting it in an XML format for transfer to translation services. Figure 6 illustrates the translation workflow from the creation of a document to the production of the translated documents.



**Figure 6. The reverse conversion localization workflow<sup>5</sup>**

As some critical data, such as confidential data, may have to be kept secret, the producer of the document should not put it in any part of the source code which goes out for translation.

<sup>5</sup> Blue boxes in the graphics represent a process that takes place in house; yellow boxes show tasks done by the translation agency.

Such data is not extracted, but kept in the "skeleton" of the document.

The *reverse conversion* consists in extracting the translated TUs from the XML documents returned by the translators and merging the translation with the remaining data from the original document in order to produce the final translated document [IBMLOCALIZATION 2005].

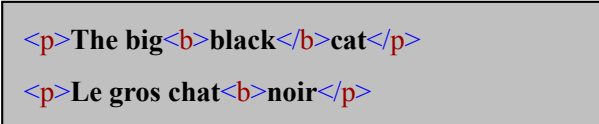
### 3.3 XML Standards for Data Management: TMX (Translation Memory eXchange)

Using XML standard formats for structuring data to be translated makes the management of documents and translation easier. In the first place, such standardization facilitates management of linguistic resources and documents in several CAT (Computer Aided Translation) tools. In addition, using standard formats reduces costs and increases the productivity of translation.

Before terminology standards appeared, terminology specialists and lexicographers exchanged data in various formats. In order to facilitate reuse of linguistic resources and increase communication exchange inside or outside a given organization, a unified format for structuring data seems important. The main aims of international standards are these:

- To facilitate the reuse of existing lexical databases, terminology term bases, translation memories, dictionaries, etc.;
- To increase the data flow between people;
- To facilitate the exchange of data between CAT tools;
- To decrease future programming efforts and avoid the need to define new structures.

TMX is developed by LISA (the Localization Industry Standards Association) for managing multilingual translation memories [LISA 2006]. It is an XML format, where the *seg* element includes translation units in several languages. Figure 7 gives an example of a source/target text within an HTML document. The same content is presented after conversion into TMX in Figure 8.



```
<p>The big<b>black</b>cat</p>
<p>Le gros chat<b>noir</b></p>
```

*Figure 7. Data in tagged HTML documents*

Most CAT tools already offer a text-based input/output format that can be used to transfer data to and from other instances of the same application [SIMILIS 2005] [TRADOS 2005]. Adjusting to the use of the TMX format, as an alternative to managing translated segments, should be relatively straightforward as it is not a complex format and there are

plenty of freely available XML parsers. Here are some advantages of this format:

- *Exchange of memories*: the most immediately obvious benefit of TMX is that it allows translation memory information to be exchanged between existing CAT tools, which permits increased communication between linguists.
- *Choice*: once a standard has been provided and its use has been encouraged, experts are free to change tools, which ensures that they don't become locked in to a particular product.
- *Openness*: given a clearly defined standard, developers of other tools have the opportunity to complement existing translation functionalities with new or proprietary features that can benefit the translation process (Figure 9).

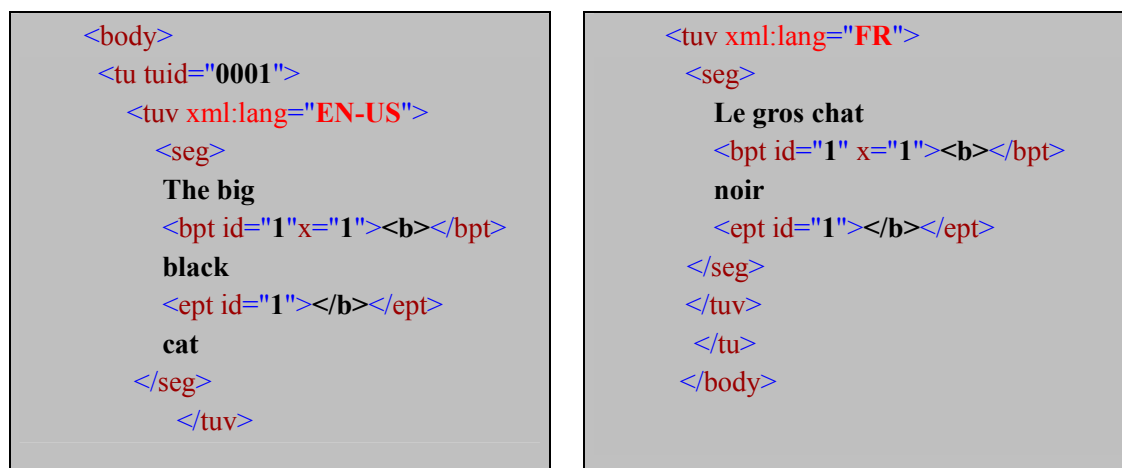


Figure 8. TMX example for translation memory management

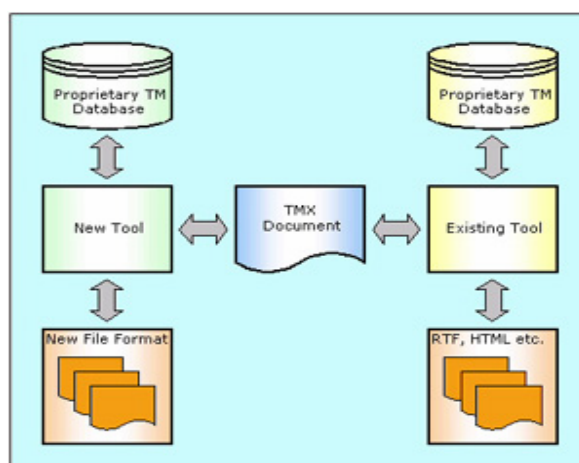


Figure 9. Openness feature of the TMX Standard



## 4. Data Management in QRLEX

In designing the QRLEX environment, the designers tried to take into consideration all relevant strengths of the existing environments described above and combine them with the advantages of the use of XML for managing heterogeneous data. Following LISA, the designers use TMX for handling multilingual content and aligned translatable units. To manage heterogeneous linguistic data, however, the designers have developed and used the new XLD (XML Linguistic Data) format [Bey *et al.* 2005].

### 4.1 Specification

#### 4.1.1 Managing Heterogeneous Linguistic Data: the XLD Format

The following are among existing reference data for which the management structure is defined (Table 2):

- “Eijiro” and “Grand Concise” are two high-quality English-Japanese unidirectional dictionaries widely used by many translators;
- “Nichigai” is specifically used for proper names;
- “Medical Scientific Terms” is included to check the structure of terminological dictionaries;
- “Edict” is a free Japanese-English dictionary, included for checking the directionality of the bilingual dictionaries.

**Table 2. Reference data in the QRLEX framework**

Reference Data	Description	Entries	Format
Eijiro 86	General English/ Japanese dictionary (EDP 2005)	1576138	Textual
Edict	Free Japanese/English Dictionary	112898	Textual
Nichigai	Guide for spelling foreign proper names in Katakana	112679	Textual
Medical Scientific Terms	Medical terms (terminology)	211165	Textual
Grand Concise	Japanese/English Dictionary	360000	XML

For each reference work, there are a few requirements: (a) various levels of recyclable units should be dealt with in a unified framework; (b) existing high-quality content should be properly accommodated; and (c) unnecessary information contained in existing content should be properly excluded, while necessary information appropriated for reference data and useful for translators should be incorporated. To satisfy these requirements, we need an internal XML structure for storing and exchanging content within different QRLEX modules.

Existing high-quality reference data in electronic form takes a variety of formats. After examination of existing XML standard formats for terminologies such as TBX (TermBase eXchange) and MARTIF (Machine-Readable Terminology Interchange Format), the authors found that these formats, unfortunately, did not satisfy the above requirements [LISA 2006]. The authors considered using the XML CDM [Mangeot 2002] format for the unified presentation of many monolingual, bilingual and multilingual usage dictionaries in the Papillon database, but it is somewhat too complex for the needs of this structure.

The authors have thus defined the basic XML structure of the linguistic data by reference to the data elements of various dictionaries and terminological lexicons. Figure 10 illustrates the XLD (XML linguistic data) format that has been developed for managing heterogeneous reference data.

The XLD format consists of three main parts:

- *Source reference data description*: contains the description of the original linguistic resources and their content. This XLD header information includes the creation date, the author profiles, the encoding, the number of entries, the source language, etc.
- *Source element*: contains the source entry and its description, e.g. the language (xml:lang).
- *Target element*: this element is multilingual. Its sub-elements contain the translations of the segment into several target languages.

```

<!-- XLD (XML Linguistic Data) structure definition -->
<!-- Part 1: General description of original version and content -->
<!ELEMENT resource      (res-info, content)>
<!ATTLIST res-info     name CDATA #REQUIRED>
<!ATTLIST res-info     author CDATA #IMPLIED>
<!ATTLIST res-info     version CDATA #IMPLIED>
<!ATTLIST res-info     date-creation CDATA #IMPLIED>
<!ATTLIST res-info     last-modification CDATA #IMPLIED>
<!ATTLIST res-info     original-codage CDATA #IMPLIED>
<!ATTLIST res-info     entries-number CDATA #IMPLIED>
<!ATTLIST res-info     description CDATA #IMPLIED>
<!ELEMENT content      (entry*)>
<!ELEMENT entry        (source, target)>
<!ATTLIST entry        id CDATA #IMPLIED>
<!-- Part 2: Source element definition -->
<!ELEMENT source       (#PCDATA)>
<!ATTLIST source       xml:lang CDATA #REQUIRED>
<!ATTLIST source       additional-info CDATA #IMPLIED>
<!-- Part 3: Target element definition -->
<!ELEMENT target       (expression+)>
<!ATTLIST target       xml:lang CDATA #REQUIRED>
<!ELEMENT expression   (#PCDATA)>
<!ATTLIST expression   add      kata-pronunciation CDATA #IMPLIED>

```

**Figure 10. DTD (Document Type Definition) of XLD format**

Source and target elements contain additional information expressed using a set of attributes:

- *Additional-info*: description of a source element. This will be useful if one transforms the relevant resource direction from, for example, Japanese-English into English-Japanese.
- *Kata-pronunciation*: In the case of Japanese linguistic data this attribute contains the pronunciation in katakana of the foreign words.

Figure 11 shows an entry after the compilation of the "Nichigai" in XLD format. The source element contains the English transliteration of an Arabic proper name and its transliteration in katakana.

```
- <entry id="Nichigai100001280">
  <source xml:lang="en" additional-info="">Abdeslam</source>
  - <target xml:lang="jp" kata-pronunciation="">
    <expression id="Nichigai100001280-1">アブデスラム</expression>
  </target>
</entry>
- <entry id="Nichigai100001290">
  <source xml:lang="en" additional-info="">Abdessadki</source>
  - <target xml:lang="jp" kata-pronunciation="">
    <expression id="Nichigai100001290-1">アブデサドキ</expression>
  </target>
</entry>
```

**Figure 11. Japanese “Nichigai” entries in XLD format**

The developers compiled all of the linguistic resources cited in Table 2 in the XLD structure, in other words, the existing resources have been preprocessed, filtered, and, after that, passed to the structure manager for transforming them in XLD XML documents.

#### 4.1.2 Managing Textual Data: the TMX Format

The document data structure should satisfy two requirements: (a) maximal facilitation of the provision of recyclable units and (b) unified management of translated documents. The first requirement comes from translators, who avidly seek existing translations of linguistic units (especially collocations and quotations) in related translations. The second requirement comes from the mission-oriented community in which translators take part. Although no readily usable reference data format was found, an existing standard framework TMX suitable for the developers' aims was found.

This standard simplifies the storage of textual data extracted from documents that contain formatting information such as HTML tags. It allows one to represent and manage translation memories as well as "multilingual" documents, that is, documents containing source and target translation units in the same file [LISA 2006]. Figure 12 illustrates TMX as used for an

English-French-Italian example translated by volunteer translators of the PAXHUMANA community [PAXHUMANA 2005].

```

<?xml version="1.0" encoding="ISO-8859-1" ?>
- <tmx version="1.4">
- <body>
- <tu tuid="0001">
- <tuv xml:lang="en">
  <seg>I recently caught a glimpse of the effects of torture in action at an
  event honoring Maher Arar. The Syrian-born Canadian is the world's most
  famous victim of "rendition," the process by which US officials outsource
  torture to foreign countries...</seg>
</tuv>
- <tuv xml:lang="fr">
  <seg>J'ai récemment eu un aperçu en action des effets de la torture lors
  d'un événement en l'honneur de Maher Arar. Ce Canadien d'origine
  syrienne est la plus célèbre victime d'un genre d'extradition spécial
  appelé « restitution » [rendition], qui est un procédé par lequel les
  fonctionnaires des États-Unis sous-traitent la torture dans d'autres
  pays...</seg>
</tuv>
- <tuv xml:lang="it">
  <seg>Ho recentemente avuto un compendio in azione degli effetti della
  tortura durante un'avvenimento in onore di Maher Arar. Questo Canadese
  di origine siriana è la vittima più famosa di un genere di estradizione
  speciale chiamato "restituzione" [rendition] un procedimento con il quale
  i funzionari degli Stati Uniti subappaltano la tortura in altri paesi...</seg>
</tuv>
</tu>
</body>
</tmx>

```

**Figure 12. Source and translated document in TMX format**

Structuring linguistic data in an XML format is appropriate according to the needs of online volunteer translators (as already explained) and accords with the overall design of QRLex for managing linguistic data in heterogeneous formats. Such structuring also makes it easy to construct a parser and to develop improved functionalities. At the management level, all imported data (reference data or textual data) and the information flowing between modules will be stored in XLD and TMX format.

## 4.2 Modular Architecture of QRLex Environment

By talking with volunteer translators and coordinators and by examining existing translation aid systems [SIMILIS 2005] [TRADOS 2005], the developers clarified a few essential general requirements:

- Content of language reference tools cannot be separated from system functionality.
- Translators look for information on (i) ordinary words, (ii) idioms and set phrases, (iii) technical terms, (iv) proper names, (v) easy collocations, and (vi) quotations. In general they conceptually distinguish these six classes but want to look them up with unified functionality and interfaces.

With respect to reference contents, therefore, the developers' needs are as follows:

- Use and update good reference material whenever it is available.
- Enhance the material when it is not sufficient.
- Make reusable translation units available from existing relevant translated documents.

Taking these general desiderata into consideration, the developers have defined a system that implements the QRLex framework by means of six functional modules (Figure 14), each of which covers specific tasks and deals with different types of data: a structure manager, a document manager, a database manager, a data control manager, a functionalities manager, and the Akin system.

- **Structure manager:** a module that transforms reference data and textual data into a structured XML format. The authors have thus compiled linguistic data including dictionaries, a Japanese pronunciation guide, technical terms, and proper name resources in XLD format. This module preprocesses and filters original linguistic data in various formats before transforming them to XLD and storing them in the centralized database. In the same manner, source documents and corresponding translated documents are processed in the documents manager module and converted into structured LISA TMX standard (Translation Memory eXchange) format [LISA 2006].
- **Document manager:** This module is based on three functionalities for the detection/extraction of textual content from document. The following paragraphs explain these functions:
  - (i) *Direct document detection:* QRLex gives online users the ability to upload source documents and their translations for internal storage. This functionality needs no access to Web searches; instead, it involves extracting the content from both source and target documents and aligning the uploaded documents.
  - (ii) *External document detection:* volunteers can search the Web to detect translated documents which can help them translate current documents. The search is carried out by crawling the Web in search of documents with bilingual content. For this purpose, the Akin-I<sup>6</sup> system has been developed. Currently, it detects only English-Japanese (in both directions), but enhancement is under way to generalize the system for the detection of other

---

<sup>6</sup> Developed at the Graduate School of Education (the University of Tokyo) by the team of Professor Kyo Kageura.

language pairs. Other functionalities such as the detection of bi-segments would also seem very helpful since translators need to know how segments (*e.g.* expressions, idioms, collocations, etc.) have been previously translated by other volunteers.

- (iii) *Internal documents detection*: this process is based on the result of Akin-I. After the initial detection of source and target documents in a specific community on the Web, the process cyclically returns to the Web, seeking additional translated documents. In the future, Akin-II, developed for crawling internal repositories after the initial identification of translation communities, will also be integrated.

The identified documents are subjected to (i) text extraction, (ii) segmentation and (iii) alignment [Walker *et al.* 2001]. The detection of sentences, or more precisely TUs, is achieved using the LingPipe tool [LingPipe 2006], which carries out sentence-boundary detection (detection of TUs) and linguistic unit detection (*e.g.* named entity detection). LingPipe can be trained to support additional languages (*e.g.* Chinese, Arabic, and French). Finally, a bi-text is constructed for each document, stored in TMX format, and put into the centralized database.

- **Database manager**: this module is the server of data to all modules of QRLex. All data flows are centralized in a relational database, which receives linguistic data in XML format from the structure manager module and serves the functionalities manager module and data control manager module. Structure data is analyzed using the DOM API<sup>7</sup> for the extraction of data from both XLD and TMX formats.
- **Data control manager**: open linguistic resource environments on the Web necessitate the intervention of human experts. In this case, the QRLex environment requires the interaction of linguistic experts or professional translators to increase the accuracy of data content and enhance the control of user interaction. This module is subdivided into two sub-modules:
  - (i) *Data validation and enhancement*: the validation/enhancement process allows the environment to interact with linguists/translators or lexicographers via an interface to temporary data which has been put on the system for revision. Some such users have password permission to revise and update the

---

<sup>7</sup> All the internal dataflow is in XML and the whole environment exploits structured data via Document Object Modeling. For further information, refer to <http://www.w3.org/DOM/>.

data. The content is controlled by active translator communities<sup>8</sup> who continually maintain it and work actively to enhance specialized data for their fields of translation.

(ii) *Linguistic data control and administration*: the administration sub-module helps the administrator to control all access to the environment. He or she has the authority to suspend users (e.g. vandals) or to manage copyrights as appropriate. Furthermore, he or she can give access to information with hierarchical levels of privilege, so that users may have access to all of the data or only to parts of it. The authors emphasize that there is an administrator for each translator community, with the power to control the interactions of users with the environment and data.

- **Functionalities manager**: functionalities are the most important elements of CAT tools. Considering the needs of translators, the authors view functionalities as the most important criteria to be considered during tool development. This module is the main interface for the interaction between translators/users and *QRLex*. It offers the possibility to display, update and use data simultaneously during translation.

The designers have thus developed *Qredit*, a specialized translation editor ([http://hygrocybe.p.u-tokyo.ac.jp:8080/qredit\\_idiom](http://hygrocybe.p.u-tokyo.ac.jp:8080/qredit_idiom)), as a first attempt to allow volunteer translators to do translation with the possibility of exploiting existing dictionaries. Figure 13 gives a snapshot of the editor where the document is downloaded automatically and compiled with dictionary entries for increasing translation speed. Translators do not need to look anything up in a dictionary. In fact, all words in the source document (left side) are linked to their translation entries in the dictionaries. Translators have only to choose words to translate by moving the cursor to them and their translations are directly displayed in a *pop-up* window. When the translation is selected, it is automatically put in the right position in the target text area.

- **Akin System**: the detection of existing translation documents is carried out by the *Akin* system [AKIN 2006] [Tsuji *et al.* 2005], which detects English-Japanese translated documents using keywords (Figure 15). Integrating *Akin* into the *QRLex* framework allows:

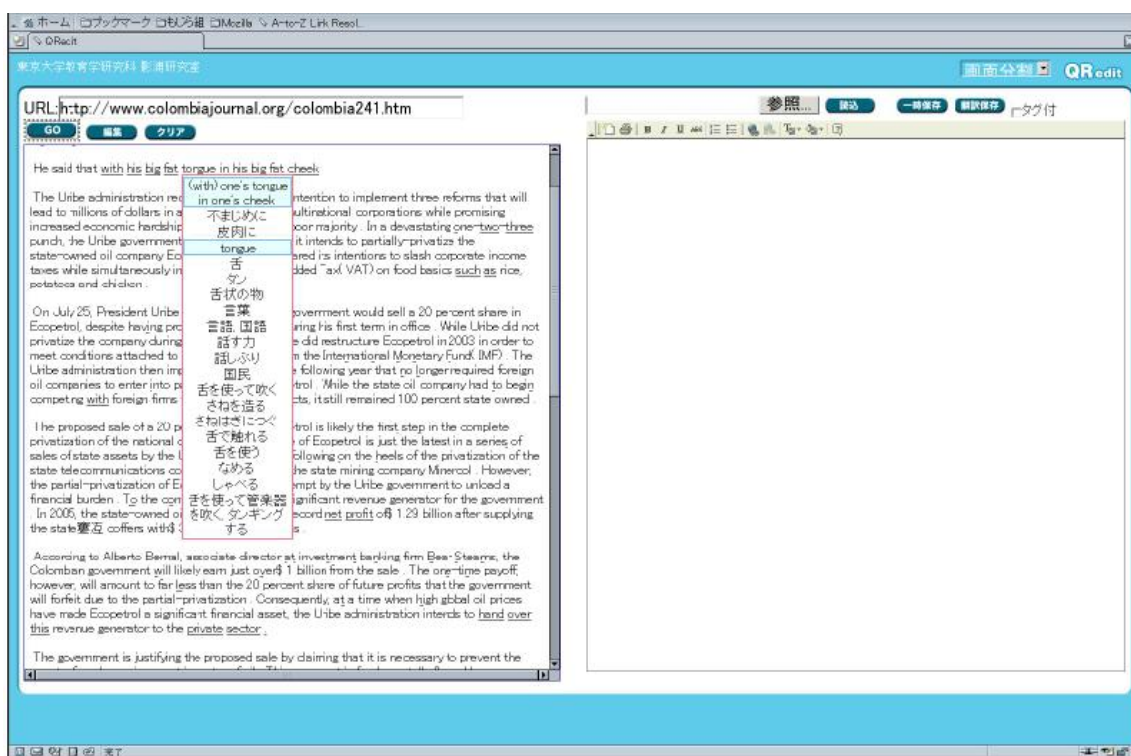
(i) *Avoiding the duplication of translation*: volunteers often check existing translations on the Web before starting the translation of a given document.

---

<sup>8</sup> The translator communities are active when the content is checked daily. This work phenomenon depends on the will of each community but often translators look for another translation and coordinate it together whenever it is possible.

They often search manually, but could make use of more efficient methods. Akin-I is intended to be called by the document manager module to check whether the relevant document has been translated on the Web or not, which avoids the translation of the same document on the Web.

- (ii) *Recycling Web and community repositories*: Akin aims to prepare for the construction of TM by detecting and recycling the existing translation along with crawling the repositories of the translation communities on the Web.
- (iii) *Detection of bi-segments*: Akin can be exploited at several levels. It can detect repositories of translation communities and documents at a high level, but also allows the detection of bi-segments (in source/target language) at a finer grain.



**Figure 13. Screenshot of the QReditor**

Keywords for search are translated using the Eijiro dictionary. They are translated into either English or to Japanese according to the direction of the desired search. The Akin search method thus differs from these of similar systems like STRAND [Resnik *et al.* 2003], which collect parallel corpora even for software or interface components.



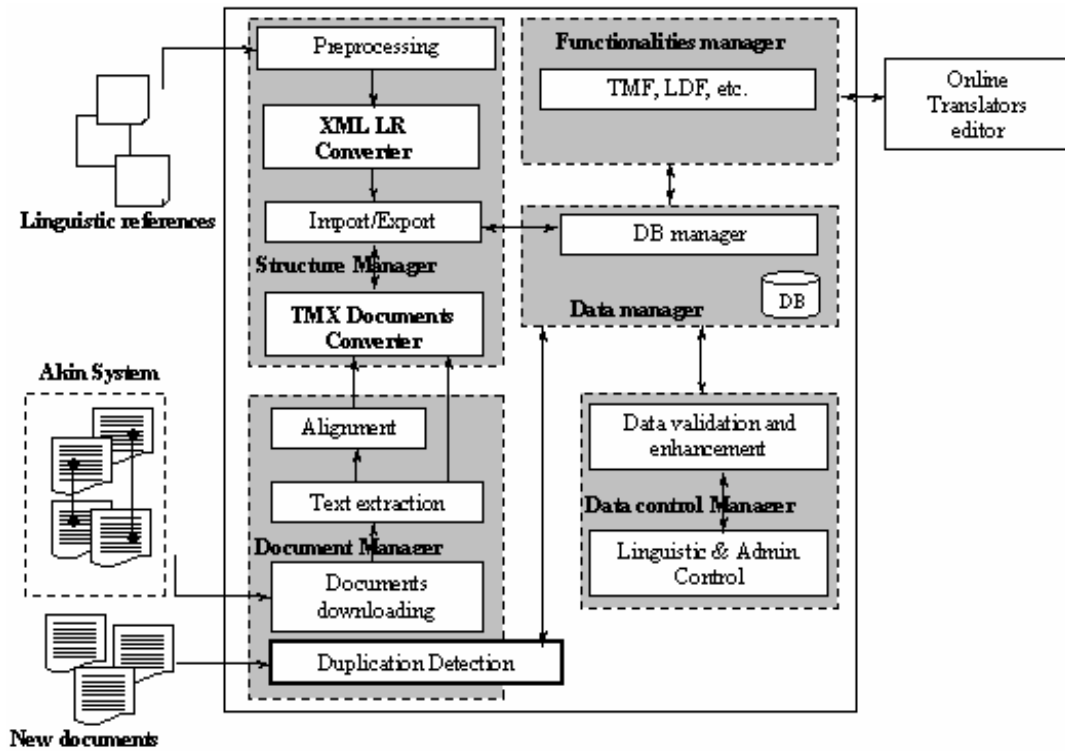
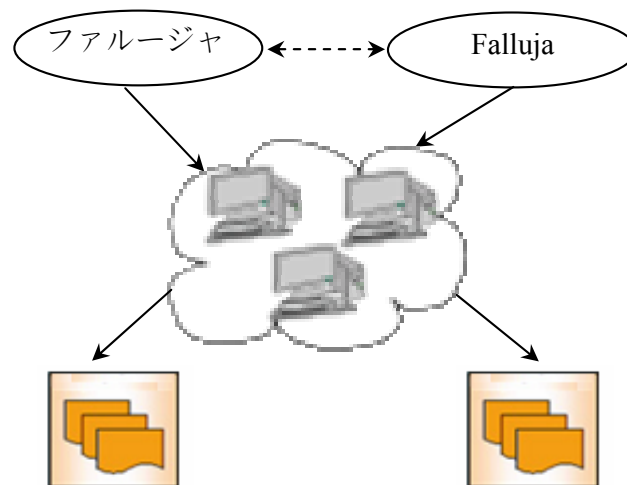


Figure 14. QRLEX architecture for the data management



Source: Japanese documents Target: English documents

Figure 15. Detection of translation on the web using the Eijiro dictionary

Collecting parallel corpora can be helpful for translators, but may not be sufficient as filtering may be needed, and volunteers who seek translations from specific communities or translators may be dissatisfied. Hence, in this framework, collected documents tend to be relevant when they are collected from specific translation communities; if they are found elsewhere, translators may have to create their own translations.

Figure 16 gives an actual snapshot of results produced by Akin-1. The entries are numbered and selected according to the direction of keywords language (e.g. ファルージャ keyword). The URL of the source document in Japanese is displayed with its title and description. The target document is displayed in the same manner as the source document. The score (displayed at the bottom) shows the degree of matching between source and target document. It is calculated according to the number of words from the Japanese document which match translations in the English document.

```
start running AKIN
-----< 1 >-----
JPN_URL = "http://www.jca.apc.org/~kmasuoka/places/iraq0404d.html"
JPN_TEXT = Text
JPN_TITLE = ファルージャの目撃者より:どうか、読んで下さい
JPN_SNIPPET = ただしどの場合でも、「この記事を含む目撃証言が『ファルージャ2004年4月』(現代企画室・1500円)として出版された」と明記して下さい。なお、ファルージャを中心にイラク情報のアップデートをファルージャ2004年4月ブログで行なっています。...
ENG_URL = http://www.onweb.to/palestine/siryo/jo-fallujah-en.html
ENG_TEXT = Text
ENG_TITLE = eyewitness report from Falluja
ENG_HEAD = Please Read - eyewitness report from Falluja by Jo Wilding I'm sorry it's so long, but please, please
SCORE = 0.372241992882562
```

**Figure 16. Detection of existing translation documents on the Web.**

The Akin system, developed separately, aims to identify or detect existing documents which have been translated from English to Japanese. However, most translators disseminate their translations through specific locations on the Web, which serves as the repository of their respective communities. To take advantage of these internal community repositories, Akin has been improved to enable recycling of existing translated documents within each community rather than search of the entire Web.

Several research groups have taken part in the QRLex project, each exploiting its special skills. The server will be set up at the University of Okayama under the direction of Dr. Koishi, who is also working with his team on the automatic compilation of Japanese-French and Japanese-English terminology. The University of the Okayama is contributing help for the construction of bilingual Japanese-French linguistic resources, for alignment of resource, and for other related tasks.

## **5. Conclusion**

The authors have proposed a new framework for a system which will aid online volunteer translators to perform translation on their PCs while sharing resources and tools and communicating via a Web site. The current status and conditions of online volunteer translators and their translation practices and tools have been examined, and related work has been discussed. The researchers examined translators' needs, first by analyzing various translation scenarios within existing online translator communities and existing environments, and subsequently by interviewing online translators. This work has clarified and modified the authors' views regarding the design of a new framework emphasizing two aspects: (i) a rich content and (ii) improved functionalities.

The system's general requirements have been derived from these main points of emphasis. Most translators request rich content in various formats, such as dictionaries, glossaries, and translation memories. The developers have accordingly developed the XLD format for compiling heterogeneous linguistic data, *e.g.* for storing usable free dictionaries, and for allowing importation of new linguistic resources to centralized relational databases within the QRLex system. At the same time, a translation memory (TM) constitutes a precious linguistic resource which most translators need to accelerate translation and improve its quality. TM will be constantly developed by recycling the documents translated on translator community Web sites or documents found on the Web by specialized search utilities like the Akin system.

From a conceptual point of view, volunteer translator communities' principal demands are for (1) storing and accessing rich heterogeneous linguistic data; (2) building large and adequate translation memories; and (3) adding improved functionalities in integrated computer-aided translation environments. The authors have thus proposed a new general architecture for online translation aid systems. They are currently developing QRLex system modules separately and intend to integrate them next year into a first working version, to be used by several online volunteer translator communities.

In parallel, the authors are working on TRANSBey [Bey *et al.* 2006], geared toward fully online translators such as contributors to translationwiki.net. The intent is to enable online translators to collaborate to solve difficult problems during the translation process so as to jointly produce high quality translations. A document would not necessarily be translated once and for all by a unique translator, but could instead be translated by several translators, and certain passages might be translated several times. For this purpose, the developers will have to design and implement another module (again using the Wiki technology again), a Web-oriented translation editor usable through any navigator and allowing the online collaborative edition of documents.

### Acknowledgements

The authors are very grateful to Prof. Akiko Aizawa for her advice and help during their stay at NII (National Institute of Informatics, Tokyo, Japan). Special thanks go to Dr. Christophe Chenon (IBM, France) for his valuable suggestions and orientation and to Dr. Mark Seligman (Spoken Translation, Inc., Berkeley, USA) who kindly improved the English level of this paper.

Last, but not least, this work is partly supported by grant-in-aid (A) 17200018 "construction of online multilingual reference tools for aiding translators" of JSPS (Japan Society for the Promotion of Sciences).

### References

- Agirre, E., X. Arregi, X. Artola, A. Díaz de Ilarraza and K. Sarasola, "Lexical KnowledgeRepresentation in an Intelligent Dictionary Help System," In *Proceedings COLING'94*, 1994, Kyoto, Japan, pp. 544-550.
- Agirre, E., X. Arregi, X. Artola, A. Díaz de Ilarraza., K. Sarasola and A. Soroa, "Constructing and intelligent dictionary help system," *Natural Language Engineering, Cambridge University Press*, 3, 1996, pp. 229-252.
- Agirre, E., X. Arregi, X. Artola, A. Díaz de Ilarraza, K. Sarasola and A. Soroa, "A Methodology for Building Translator-oriented Dictionary Systems," *Machine Translation*, 15, 2000, pp. 295-310.
- AKIN, Detection of Translated Documents, version 1.0.0, <http://apple.cs.nyu.edu/akin/>, 2005.
- Allen, J., "Postediting: an integrated part of a translation software program (Reverso Pro 4)," *Language International Magazine*, 13(2), 2001, pp. 26-29.
- Augar, N., R. Raitman and W. Zhou, "Teaching and Learning Online with Wikis School of Information Technology," In *Proceedings of the 21st Australasian Society of Computers in Learning in Tertiary Education Conference*, 2004, Australia, pp. 95-104.
- Bey, Y., C. Boitet and K. Kageura, "The TRANSBey Prototype: An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators," In *Proceedings of the 3rd International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III)*, E. Yuste, (ed.), *LREC Fifth International Conference on Language Resources and Evaluation*, 2006, Genoa, Italy, pp. 49-54.
- Bey, Y., K. Kageura and C. Boitet, "A Framework for Data Management for the Online Volunteer Translators' Aid System QRLex," In *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation*, 2005, Taipei, Taiwan, pp. 51-60.
- Boitet, C., "New architectures for "democratic" tunable quality MT systems," In *Proceeding of Pacific Association for Computational Linguistics*, H. Sakaki, (ed.), 2005, Meisei daigaku, Hino campus, Tokyo, Japan, pp. 33-57.

- Bowker, L., *Computer-Aided Translation Technology: A Practical Introduction*, Didactics of Translation Series, University of Ottawa Press, 2002.
- HTMLArea, In-browser WYSIWYG editor for XWiki, <http://www.htmlarea.com/>, 2006.
- Hutchins, J., "The Origins of the Translator's Workstation," *Machine Translation*, 13, 1998, pp. 287-307.
- Hutchins, J., Machine Translation and Computer-Based Translation Tools: What's Available and How It's Used, <http://ourworld.compuserve.com/homepages/WJHutchins/>, 2003.
- Kay, M., "The Proper Place of Men and Machines in Language Translation," *Machine Translation*, 12, 1997, pp. 3-23.
- IBM LOCALIZATION, XML in localization: a practical analysis, <http://www-106.ibm.com/developerworks/xml/library/x-localis/#example1>, 2005.
- LingPipe, Linguistic Toolkit: Sentence-Boundary Detection, Named-Entity Extraction, Language Modeling, Multi-Class Classification, <http://alias-i.com/lingpipe/demo.html>, 2006.
- LISA, Localization Industry Standards Association, <http://www.lisa.com/>, 2006.
- Mangeot, M., "An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language," In *Proceeding of International Standards of Terminology and Language Resources Management, LREC 2002 workshop*, 2002, Las Palmas, Islas Canarias, Spain, pp. 37-44.
- Martin, W., "User-orientation in Dictionaries: 9 Propositions," In *Proceedings BudaLEX'88*, 1988, Budapest: akadémiai kiadó, Hungary, pp. 393-399.
- MOZILLA, Open Software Localization, <http://frenchmozilla.online.fr/>, 2005.
- MT POST-EDITING, MT postediting and translators tools reviews. <http://www.geocities.com/mtpostediting/>, 2006.
- PAXHUMANA, Translation of Various Humanitarian Reports in French, English, German and Spanish, <http://paxhumana.info>, 2006.
- Queens, F. and U. Recker-Hamm, "A Net-Based Toolkit for Collaborative Editing and Publishing of Dictionaries," *Literary and Linguistic Computing Advance Access, Oxford Journal, Lit Linguist Computing*, 20(1), 2005, pp. 165-175.
- REDOX, Wiki Engine of XWiki, <http://radeox.org/space/start>, 2005.
- Saha, G.K.A., "Novel 3-Tiers XML Schematic Approach for Web Page Translation," In *ACM IT Magazine and Forum*, 6(43), 2005, [http://www.acm.org/ubiquity/views/v6i43\\_saha.html](http://www.acm.org/ubiquity/views/v6i43_saha.html).
- Schwartz, L., "Educational Wikis: Features and Selection Criteria," In *the International Review of Research in Open and Distance Learning*, technical report R27/0311, Athabasca University, Canada's Open University, 2004.
- SIMILIS, Second Generation of Translation Memory Tool, <http://www.lingua-et-machina.com/>, 2005.

- TEANOTWAR, Human Rights Documents Translation, English to Japanese News Translation, <http://teanotwar.blogtribe.org/>, 2005.
- TRADOS, Translation Memory Tool, <http://www.trados.com/>, 2005.
- TRADUCT, Linux Documentation Translation, <http://wiki.traduc.org/>, 2005.
- Tsuji, K., Sato S. and Kageura K., "Evaluation of the Usefulness of Search Engines in Validating Proper Name Transliterations," In *Proceeding of 11th Conference of Natural Language Processing Society of Japan*, 2005, Japan, pp. 352-355.
- W3C, Specification Translation, <http://www.w3.org/Consortium/Translation>, 2005.
- Walker, D.J., Clements D.E., Darwin M. and Amtrup W., "Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality," In *Proceedings of the 8th Machine Translation Summit*, 2001, Santiago de Compostela, Spain, pp. 369-372.
- WIKITRANSLATION, Collaborative Wiki-Based Translation, <http://www.translationwiki.com>, 2006.
- XWIKI, Open Source Java-Based Wiki, <http://www.xwiki.com/>, 2006.

## Using a Small Corpus to Test Linguistic Hypotheses: Evaluating ‘People’ in the State of the Union Addresses

Kathleen Ahrens\*

### Abstract

This paper argues that small corpora are useful in testing specific linguistic hypotheses, particularly those dealing with rhetoric, stylistics, and sociolinguistics. In particular, we hypothesize that creating a database of U.S. presidential speeches will allow for a diachronic exploration of language use at the highest political level, and enable a contrast to be drawn between legislative advances for minorities in the United States and the integration of those advances into the presidential lexicon. In order to test this hypothesis, we examine the corpora of State of the Union Addresses from 1945 to 2006. We demonstrate that while there was clearly a shift two decades ago to systematically portraying human beings as being made up of two genders, or being subsumed under a gender-neutral term, other aspects of gender, such as parenthood, are still stereotyped by American presidents. In short, analyzing lexical instances related to ‘people’ in the State of the Union address allows us not only to reflect on the values held by U.S. presidents, but also to systematically uncover how they use language to exercise power on the very people they are elected to serve.

**Keywords:** Small Corpora, Politics, Language, Gender, Diachronic Analysis

### 1. Introduction

In the latter half of the twentieth century, American presidents have had the enviable task of shaping the way Americans think about themselves by delivering a State of the Union address near the beginning of each calendar year. This speech is broadcast live across the nation on major television and radio channels. In the address, the president emphasizes his accomplishments to date and sets out a new agenda for the year. Topics touched upon may include both foreign and domestic policy, and run the gamut from justification for war to a

---

\* Graduate Institute of Linguistics & Department of Foreign Languages and Literatures, National Taiwan University, 1 Roosevelt Road, Section 4, Taipei 106, Taiwan  
Phone: +886-2-2366-1381 ext. 307 Fax: +886-2-2363-5358  
E-mail: kathleenahrens@yahoo.com

fervent plea to pass an education bill. The complete text of the address appears the following day in major newspapers and in on-line news resources.

Thus, these addresses constitute a narrow, but influential media genre, since subsequent discourse in the news media often reports on the proposals put forth by the president in his own terminology [Barrett 2004]. This terminology reflects the ideology of the ruling political party, and it is this ideology that is used to exercise power “through the manufacture of consent” [Fairclough 2001]. Moreover, as Van Dijk [1993] notes, “More control over more properties of text and context, involving more people, is thus generally (though not always) associated with more influence, and hence with hegemony” (p. 257). Thus, a linguistic analysis of presidential speeches has the potential to shed light on how the president views (and wants the country to view) economic, political, and social issues of the day.

Recent advances in corpus linguistics have facilitated the collection and analysis of presidential speeches. Kowal *et al.* [1997], for example, created a specially marked-up corpus of Inaugural Addresses by hand in order to look at the interaction between literacy and orality in presidential speeches, while Charteris-Black [2004, 2005], who looked only at the text, was able to examine the use of metaphor as well as rhetorical devices used by U.S. presidents based on the corpora of U.S. presidential speeches found on-line.

Lim [2002] also used the corpora of U.S. inaugural addresses, as well as the annual messages (State of the Union addresses), in order to identify rhetorical change in presidential speeches. He argued that presidential speeches have become more anti-intellectual, as well as more abstract, democratic, assertive, and conversational, according to changes that can be seen in the categories of words that occur in speeches over time. He also found that words related to cognitive processes and states have decreased since Hoover, while interjections have increased, indicating a decline in the difficulty of presidential rhetoric. In addition, words having to do with ‘kinship’, which Lim suggests reflects democratic rhetoric, have increased substantially since Franklin Roosevelt, as have words relating to ‘children’ and ‘youth’ since Carter. While Lim’s paper suggests that frequency of lexical use is a way of judging what is important to the president and to the public, to date there has been no systematic analysis of changes in lexical use within the scope of presidential speeches. Thus, it is the goal of this paper to demonstrate that by combining presidential speeches into a corpus, subtle changes in language use over time can be determined by examining the frequency of occurrence of key words as well as their associated collocations [Stubbs 1996].

In order to examine this issue, we will explore language use pertaining to ‘people’ in all of the State of the Union addresses (SUO corpus) from 1945 to 2006 by analyzing the tokens:



*humankind, mankind, man, men, woman, women, mother, father and parent.*<sup>1</sup> We will demonstrate that while there has clearly been shift away from using *man, men, and mankind* to refer to all human beings, other aspects of gender, such as parenthood, are still stereotyped by American presidents.

## 2. Methodology for Corpus Creation

The State of the Union (SOU) corpus was downloaded one speech at a time from the C-Span website (c-span.org). All State of the Union speeches from 1945-2006 (excluding Nixon's five SOU speeches from 1970-1974) were directly downloaded to text files (Table 1). Nixon's speeches were printed out from Adobe files, manually typed into a document file, and then saved to a text file. These five files were then checked by two additional readers for accuracy and any errors or omissions were corrected. Each file was then imported into Microsoft Word and the president's words in the speech were highlighted and counted with the word-count feature. In most cases, this meant that the heading was omitted (*i.e.* "President Harry S. Truman's Address Before a Joint Session of Congress"). In other cases, information about where and when the speech was given, or who introduced the President also had to be omitted from the word count. In the case of Bush Jr., indications of "(Applause.)" had to be deleted as well.<sup>2</sup> In short, every effort was made to include the words used by the president himself in the word count.<sup>3</sup> In addition, it is important to note that these speeches were given orally from a prepared text. The version that being examined here is the version that was provided for the written, historical record and the content may therefore vary slightly from the actual words that the president spoke.

---

<sup>1</sup> 1945 was chosen as the start date because C-span starts their database with Truman's 1945 address to Congress. 1945 is also significant because the Second World War came to a close in that year. 1947 was also considered as a starting date, since that was the first year the SOU address was broadcast on television. However, since all the other presidents have their complete SOU corpus included in the study (complete to date for Bush Jr.), 1945 was selected as the starting year so that Truman's corpus would also be complete.

<sup>2</sup> "Bush Sr." refers to George H. W. Bush, and "Bush Jr." refers for George W. Bush, following the usage in Charteris-Black [2005].

<sup>3</sup> It is important to note that Microsoft Word counts the dash punctuation mark as one word when it is written as two short hyphens close together with a space on either side or as one short hyphen with a space on either side (*i.e.* as "--" or as "-"). When it is written as a long, unbroken line, as in "—" the program does not count it as a word. In addition, in some speeches, only a short hyphen is used without a space on either side, which the word count program then interprets as a hyphenated word. Ideally, for the most precise word count possible, each speech should be re-edited for uniformity among the various types of dash marks used. However, such editing carries the risk of altering the intent of the original and was not carried out for this study.

The total and average word counts for each president are given in Table 1 for the Democratic and Republican presidents. Since some presidents gave more than one SOU address in a given year (*i.e.* Bush gave two addresses in 2001, one on 2/27 and one on 9/20), there are a total of 65 speeches in this corpora.<sup>4</sup>

**Table 1. State of the Union Speeches included in current corpora**

Name	Year	Number	Political Party	Word Count	Avg. # Words/Speech
<b>Truman</b>	1945-1951	7	Democrat	52,934	7562
<b>Eisenhower</b>	1953-1960	8	Republican	48,185	6023
<b>Kennedy</b>	1961-1963	3	Democrat	18,168	6056
<b>Johnson</b>	1963-1969	8*	Democrat	34,902	4363
<b>Nixon</b>	1970-1974	5	Republican	19,422	3884
<b>Ford</b>	1975-1977	3	Republican	13,801	4600
<b>Carter</b>	1978-1980	3	Democrat	11,250	3750
<b>Reagan</b>	1981-1988	8	Republican	36,664	4583
<b>G.H.W. Bush</b>	1989-1992	5*	Republican	20,477	4095
<b>Clinton</b>	1993-2000	8	Democrat	60,591	7574
<b>G. W. Bush</b>	2001-2006	7*	Republican	32,358	4623
<b>Total</b>		<b>65</b>		<b>348,752</b>	<b>5365</b>

\*Presidents gave more than one SOU in a given year.

Clinton, is, as noted by many pundits, the most prolix speaker in terms of the average number of words per speech (Table 1), although Truman is a close second, and Kennedy and Eisenhower follow with approximately 6,000 words per speech. These are also the only four presidents to have a higher word count than the average of 5365 words per speech; all the other presidents average between 3700 and 4700 words per speech (Table 1 above).

After saving all 65 speeches as text files, a meta-file was created and word searches were run using Wordsmith, Version 3 (<http://www.lexically.net/wordsmith/>).<sup>5</sup> Wordsmith creates a concordance for all instances of the lexical item chosen, with links to the full-text. Data on the number of instances found is then saved into Excel tables for further analysis.

<sup>4</sup> C-span incorrectly lists Carter as giving a SOU speech in 1981. The actual file under Carter's name is in fact, Reagan's first SOU address (as accessed on February 6, 2006 at [http://www.c-span.org/executive/transcript.asp?cat=current\\_event&code=bush\\_admin&year=1981](http://www.c-span.org/executive/transcript.asp?cat=current_event&code=bush_admin&year=1981)). This has not been included in the Carter's corpus herein, but it is, of course, included in Reagan's.

<sup>5</sup> Note that in this version possessives are included when a noun is entered as a search term (*i.e.* when searching for "mankind" the form "mankind's" is also returned).

### 3. Socio-Cultural Background

This paper hypothesizes that the language used by politicians became more inclusive from the middle of the 20<sup>th</sup> century to the beginning of the 21<sup>st</sup> century. In particular, social gains from the Civil Rights Act of 1964, which prohibited employment discrimination on the basis of race, color, religion, national origin, or gender, should become apparent.<sup>6</sup> However, many doors to women were still closed, even after this law went into effect. For example, the state law of Virginia prohibited women from being admitted to the College of Arts and Sciences of the University of Virginia as late as 1970.<sup>7</sup> Due to increasing demand for equal access to education at all levels, Title IX was passed, and signed into law in 1972 by Nixon. Title IX prohibits institutions that receive federal funding from practicing gender discrimination in educational programs or activities. It took two years for regulations to be drawn up for Title IX, and in 1974 they were published, with Ford signing them into law in 1975. Thus, there is a societal and legislative shift during the decade from 1964-1974. Since the corpus under consideration here identifies all uses by date and speaker, it is possible to contrast critical legislative and legal events with the occurrence of relevant terms or changing use of terms in the presidential lexicon and ask the question: How soon after this legislation was discriminatory language use dropped from presidential parlance?

In particular, we hypothesize that there should be a marked decrease in the use of *mankind* to refer to all humankind, as well as a decline in the use of *man* to refer to all people. In addition, references to women should go beyond motherhood and include the contributions that women make to society. Lastly, use of *mother(s)* and *father(s)* should demonstrate the variety of roles that each parent plays in the family and society. These changes would reflect the advances American women have made over the past half a century and would indicate that women's contributions are being recognized at the highest levels of power in the government.

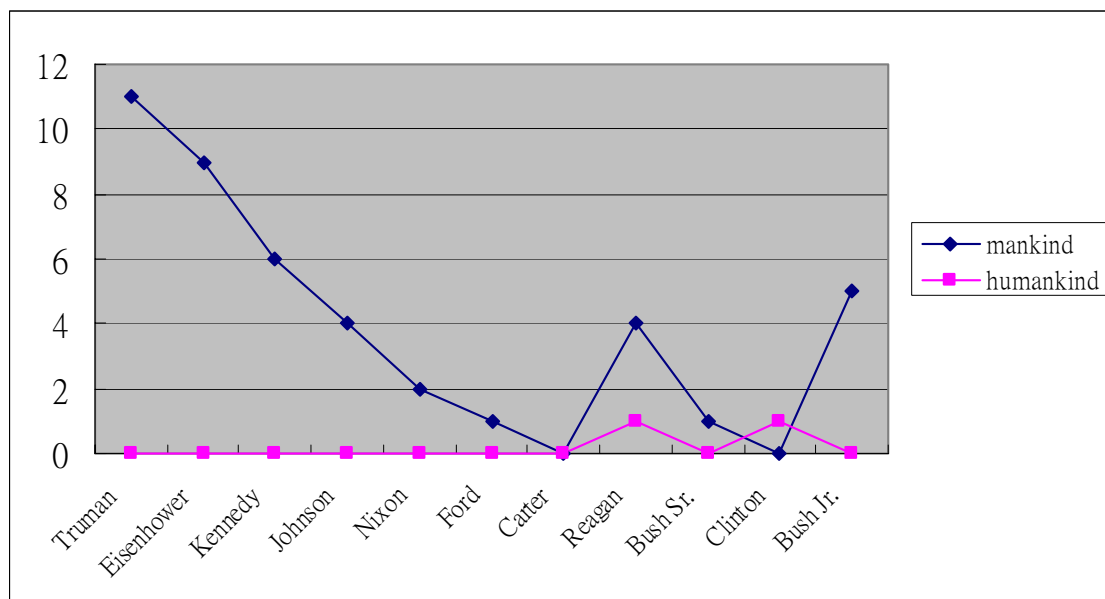
### 4. Data Analysis

We first look at the number of occurrences for *mankind* and *humankind* (Figures 1 and 2). Figure 1 shows the tokens of occurrences of *mankind* and *humankind* in each presidential State of the Union Corpus, while Figure 2 shows the percentage of occurrence of *mankind* and *humankind* when divided by the total number of words in each presidential SOU corpus.

---

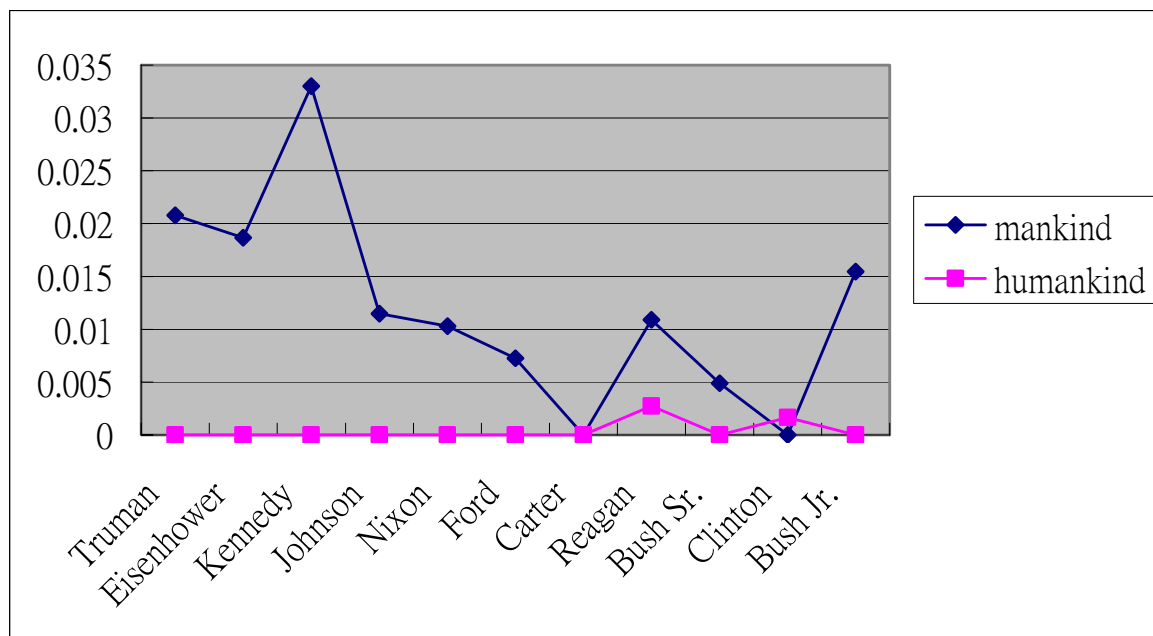
<sup>6</sup> The prohibitions on employment discrimination were codified in Title VII of the 1964 Civil Rights Act.

<sup>7</sup> *Kirstein v. Rector and Visitors of University of Virginia*, 309 F.Supp. 184 (E.D. Va. 1970), accessed from [www.ed.gov/pubs/TitleIX/part3.html#road](http://www.ed.gov/pubs/TitleIX/part3.html#road) on October 20, 2005.



**Figure 1. Tokens of occurrences of mankind versus humankind**

It is clear that there is a steady decrease in the use of *mankind* from 1945 to 1979, although Figure 2 shows that Kennedy, overall, had the highest percentage of occurrences of *mankind*.



**Figure 2. Percentage of occurrences of mankind versus humankind**

What is interesting is that, even though there is a clear decline in the use of *mankind* from Truman through Carter (in terms of number of tokens) and between Johnson and Carter (in terms of percentage of use), Ronald Reagan and both Bushes keep the term alive (cf. examples

1-3).

- (1) *That we would use these gifts for good and generous purposes and would secure them not just for ourselves, and for our children, but for all mankind.* [Reagan 1987]
- (2) *What is at stake is more than one small country; it is a big idea: a new world order, where diverse nations are drawn together in common cause to achieve the universal aspirations of mankind -- peace and security, freedom, and the rule of law.* [Bush Sr. 1991]
- (3) *The cause we serve is right, because it is the cause of all mankind.* [Bush Jr. 2004]

However, Carter and Clinton (both Democrats) clearly shun usage of the term, with Clinton preferring to use the inclusive term *humankind* instead (example 4).

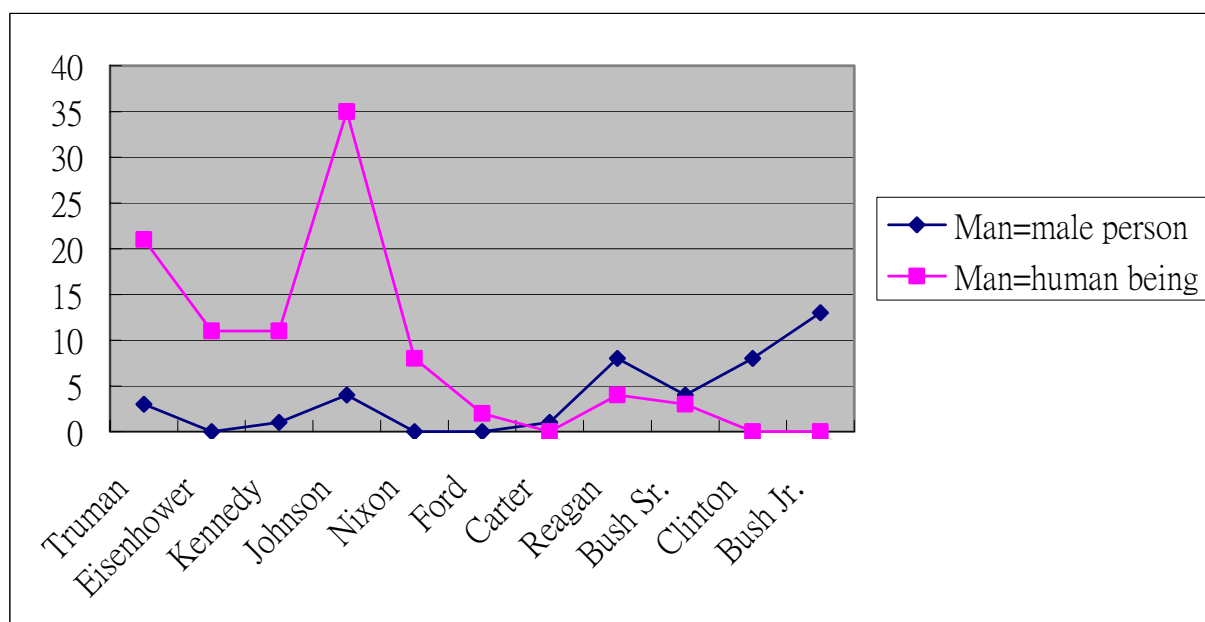
- (4) *Throughout all history, humankind has had only one place to call home, our planet, Earth.* [Clinton 1998]

This indicates insensitivity to language use on the part of these three Republican presidents (Reagan, Bush Sr., and Bush Jr.) in comparison with Carter and Clinton. Reagan's case is especially telling since he uses both *humankind* (example 5) and *mankind* (example 6) in the same speech (albeit paragraphs away from each other).

- (5) *...the belief that the most exciting revolution ever known to humankind began with three simple words: "We the People"--the revolutionary notion that the people grant government its rights...* [Reagan 1988]
- (6) *It reduces the risk of war and the threat of nuclear weapons to all mankind. Strategic defenses that threaten no one could offer the world a safer, more stable basis for deterrence.* [Reagan 1988]

Thus, for Reagan, these two terms are interchangeable (or on an alternative reading of (6), nuclear weapons only threaten men and not women).

Reagan and Bush Sr. also use *man* to stand for ‘human being’ (Figures 3 and 4; examples 7 and 8 respectively), again indicating an insensitivity to the language and gender issues, or else, on an alternative reading, an admission that the problems faced by society are beyond the ken of men, but not of women (example 8).



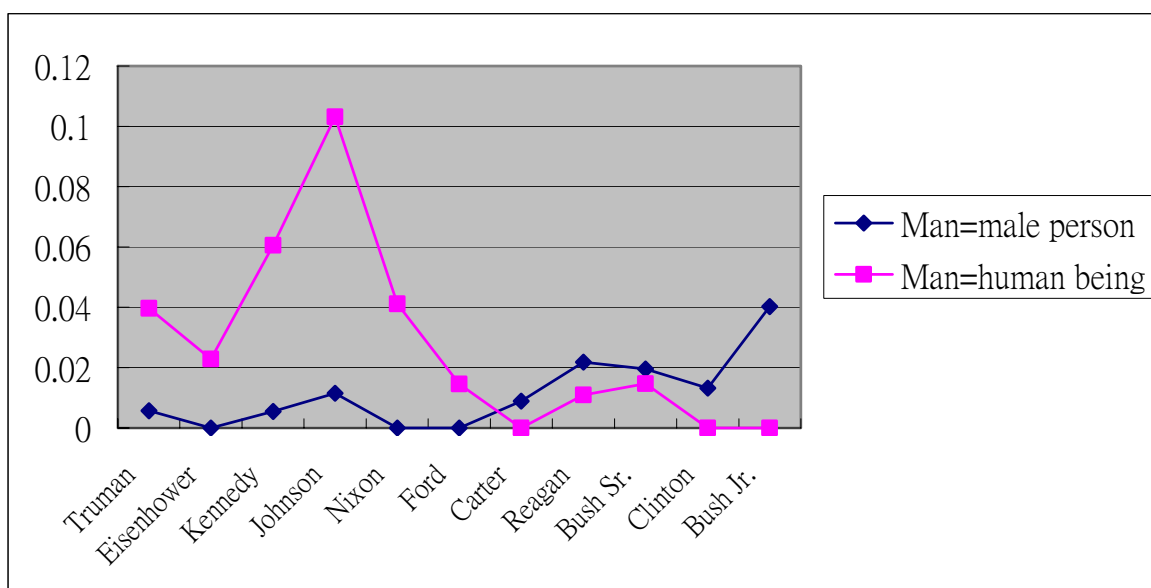
**Figure 3. Tokens of occurrences of man as a male person versus man as standing for all humans**

(7) *How can we not do what is right and needed to preserve this last best hope of man on Earth?* [Reagan 1984]

(8) *Twice before, those hopes proved to be a distant dream, beyond the grasp of man.* [Bush Sr. 3/6/1991]

Yet even so, the downward trend in the use of *man* to stand for ‘human beings,’ in both number of tokens used and overall percentage of usage in each SOU corpus (Figure 4), indicates that this usage decreased after 1970.<sup>8</sup>

<sup>8</sup> We ascertained that ‘man’ referred to ‘human being’ if it could be replaced by ‘human being(s)/person/people’ without changing the truth condition of the sentence.



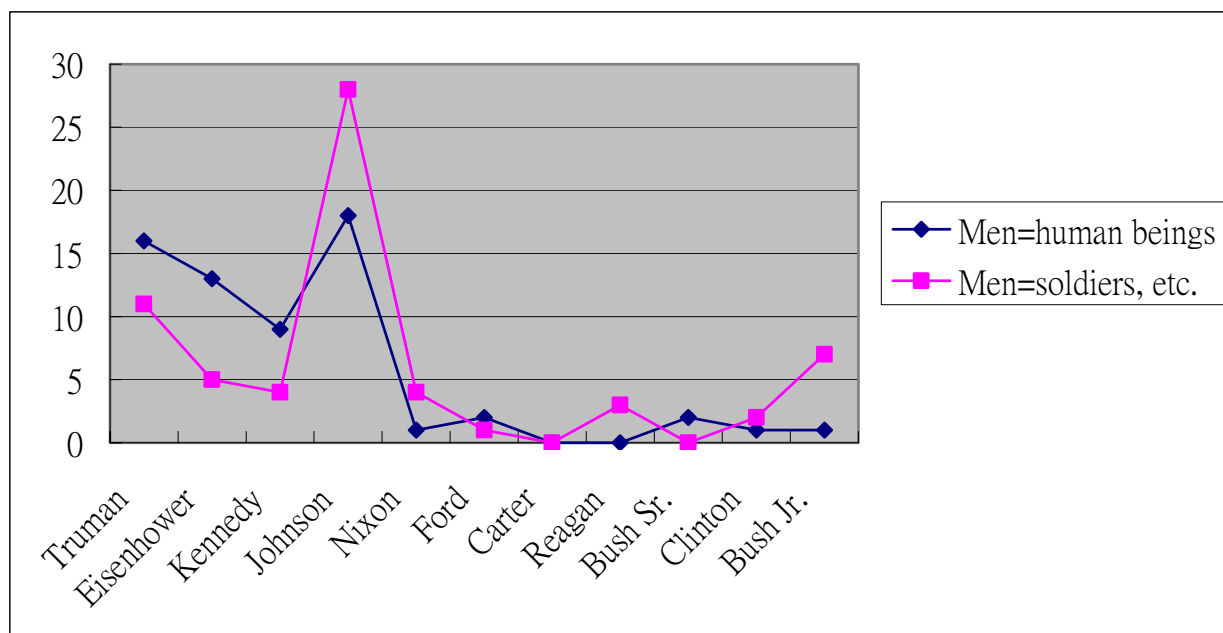
**Figure 4. Percentage of occurrences of man as a male person versus man as standing for all humans**

In fact, the two most recent presidents, one Republican and one Democrat, both clearly avoid the use of *man* to stand for ‘human being.’ Their data (or absence thereof) contrasts sharply with Democratic and Republican presidents fifty years earlier, who regularly used *man* to stand for all human beings, as for example, when Johnson states, “I speak tonight for the dignity of *man* and the destiny of democracy” (January 4, 1965). In addition, the fact that Bush Jr. never uses *man* to refer to ‘human being’ contrasts with his usage of *mankind* as discussed above, and distinguishes himself from the rhetoric of Bush Sr. and Reagan who use both *man* and *mankind* to refer to men and women.

Thus, from the data so far, results are mixed. In terms of *mankind* there is a decreasing trend in usage from 1945 to 1979. However, recent Republican presidents are keeping the term alive. In terms of *man* standing for all people, this usage did not disappear until 1993, much later than hypothesized. However, there was a clear difference between its frequency of use pre-1970 as compared with post-1970.

A similar decline is found in the use of *men* to mean ‘human beings’ as shown in Figures 5 and 6.<sup>9</sup> In addition, the use of *men* to refer to soldiers, doctors, senators showed a clearly precipitous decline as well.

<sup>9</sup> The search was run for ‘men’ excluding the collocation ‘men and women.’ We ascertained that ‘men’ meant ‘human beings’ if it could be replaced with ‘humans/human beings/people’ without changing the truth condition of the sentence.



**Figure 5. Number of tokens of ‘men’ as soldier, etc. versus ‘men’ as standing for all people**

This indicates that these positions were viewed as being filled exclusively (or almost exclusively) prior to 1970 by men, even if this was not necessarily the case. That is, one possibility is that the presidents knew that women served in the armed forces, but yet chose to use the word ‘men’ to refer to members of the armed forces in any case. The other possibility is that the fact that women served (in albeit much smaller numbers) in auxiliary roles in the armed forces was not recognized by presidents.

The use of *men* pre-1970 is used generically, but is also used to refer to soldiers (9), senators (10), and lawyers (11), *i.e.* jobs that were prototypically male. Johnson, who was president during much of the Vietnam war, used *men* to refer to soldiers most frequently (Figure 5), but he also used *men* to refer to people in prototypically male positions (examples 10 & 11), such as senators and law enforcement agents.

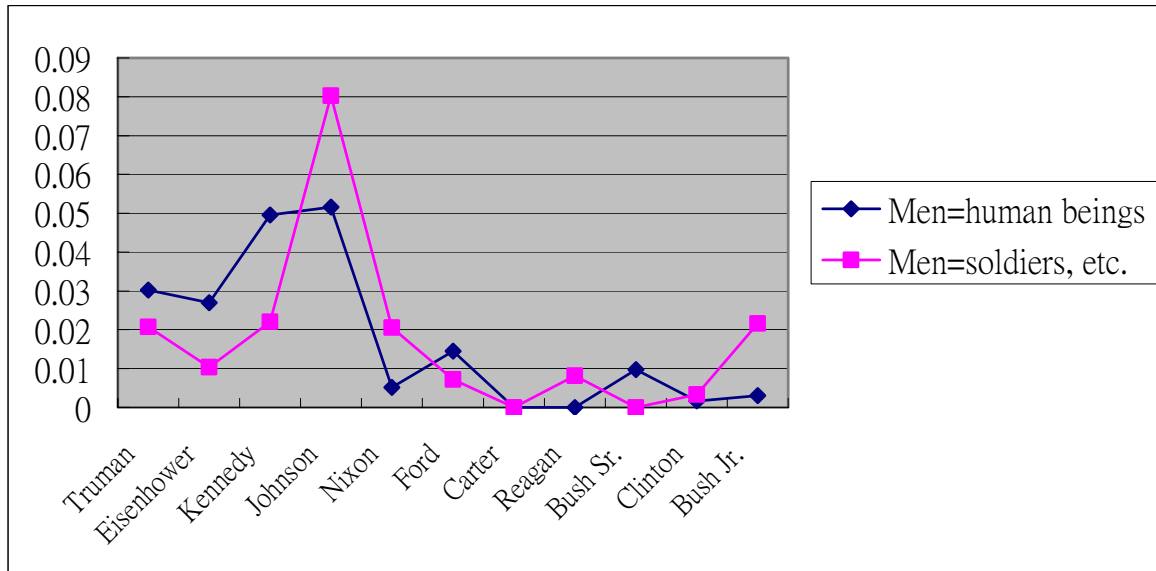
(9) *Our men are fighting, alongside their United Nations allies, because they know...* [Truman 1951]

(10) *You will soon learn that you are among men whose first love is their country, men who try each day to do as best they can what they believe is right.* [Johnson 1965]



## Evaluating 'People' in the State of the Union Addresses

(11) *I ask the Congress for authority to hire 100 more. These young men will give special attention to this drug abuse, too.* [Johnson 1968]



**Figure 6. Percentage of occurrences of 'men' as soldier, etc. versus 'men' as standing for all people**

In fact, it is not until Reagan that women are recognized as being part of the Armed Forces, although many fought and died as nurses and support staff in the Armed Forces prior to 1980.<sup>10</sup> Both Bush Sr. and Clinton use *men* when quoting from the Declaration of Independence (“all *men* are created equal”). Bush Sr. also talks about how the world is growing “stronger, more united, more attractive to *men* on both sides of the Iron Curtain...” These two usages are the only instances of *men* being used to stand for both men and women in Bush Sr.’s corpora; however, he also uses *man* to stand for all humans three times as well. Although this could be considered a continued insensitivity to issues related to language and gender; from another point of view, his percentage of usage of the gender-encompassing reading is in sharp contrast to the percentage of usage to presidents prior to 1970. In fact, the overall percentage of usage of *man* or *men* to stand for all humans showed a large decrease in percentage of usage post-1970 as compared with the period prior to 1970.

Bush Jr.’s use of *men* deserves special note because he uses it to refer to ‘evil-doers,’ as in example (12).

(12) *This conviction leads us into the world to help the afflicted, and defend the peace, and confound the designs of evil men.* [Bush 2003]

<sup>10</sup>Truman did refer to men and women who have suffered in the Second World War in his first SOU address when he said, “Our debt to the heroic men and valiant women in the service of our country...”

Although Bush Jr. is careful to talk about America's servicemen and servicewomen, he terms propagators of acts of terror as 'men' and does not use gender-inclusive language that can be found elsewhere in his speeches. Although we sincerely hope that it is the case that terrorist masterminds will not in the future include women among their ranks, it is interesting to compare Bush's language with that of Eisenhower, Truman, Kennedy, and Johnson, as it is apparent in their speeches that they did not foresee women joining the ranks of lawyers, politicians and soldiers.<sup>11</sup>

The data used to provide Figure 5 also shows that the use of the word *men* occurs alone 128 times in the total corpus (referring to both human beings and soldiers, lawyers, politicians, etc.) *Women* by contrast, occurs alone only 32 times (*i.e.* when it does not co-occur with *men*). Comparisons of *woman* and *man* (23 tokens versus 137 tokens) and *she* and *he* (68 tokens versus 294 tokens) substantiate this approximately five to one ratio between male terms and female terms.

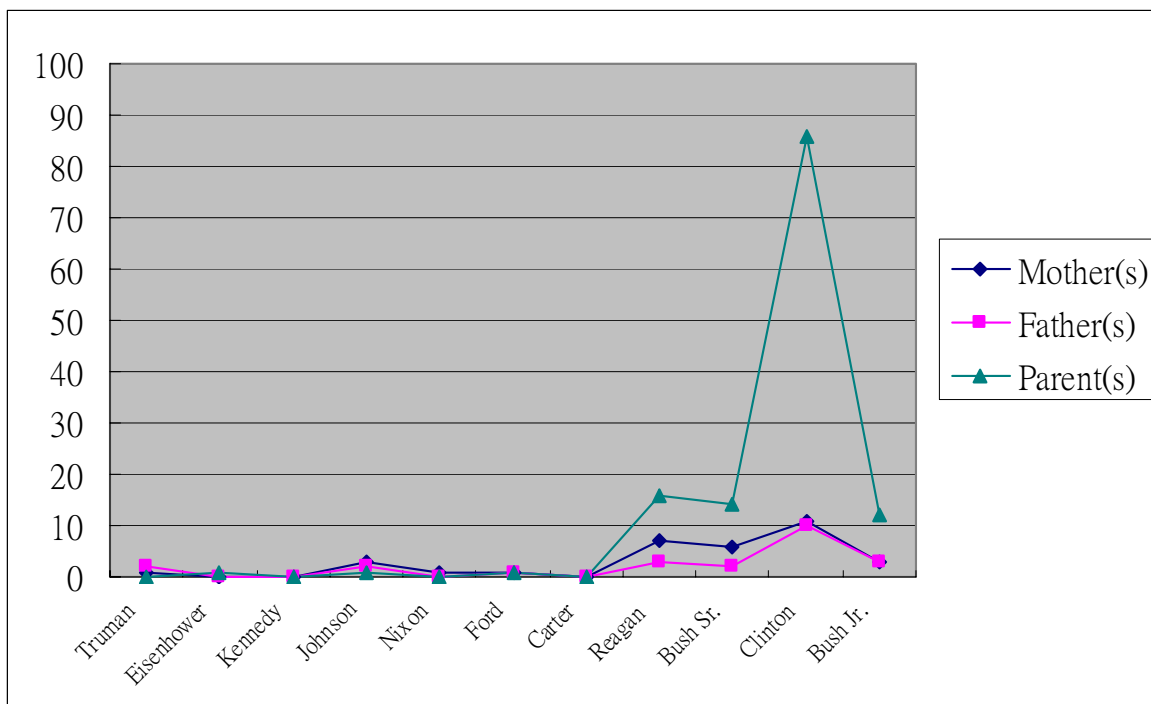
In addition, when women are mentioned, they are not mentioned as politicians or business leaders or doctors. Truman talks about working women (1 time), Carter emphasizes legal rights for women (3 times), Reagan talks about legal and economic equity for women (6 times), and women as mothers (1 time) and workers (1 time). Bush Sr. talks about pregnant women and working women (1 time each). Clinton talks about taking better care of women and children (7 times) and equity for working women (2 times). Bush Jr. talks about rights for women at home and abroad (6 times), and protecting women and children from terrorist acts (2 times) and respecting women (1 time). In short, women are talked about much less frequently than men, even in recent years, and when they are talked about they are not thanked for their contributions to the society as the Founding Fathers, senators, lawyers, and soldiers are (although later presidents (Reagan and on) do thank the men and women serving in the armed forces). Women as a group are not held up for emulation, they are only mentioned in reference to having their economic and political lot improved.

This finding can be corroborated by contrasting modification of specific men and women. Eight specific men are modified by 'good' (twice), 'great' (once), 'good and honest' (once), 'prudent' (once), 'evil' (once, for Saddam Hussein), 'reasonable' (once), 'very wise' (once), and 'brave young' (once). Women, when occurring alone, are modified as follows: Tired, decent cleaning woman (once), young woman (once), battered women (once), elderly women (once), and pregnant women (three times). Thus, it seems while there has been advances in recognizing the fact that women serve alongside men in the Armed Forces, and that 'man' is not a gender-inclusive term, still women face an uphill battle to have their deeds and accomplishments recognized – to be held up as the standard bearer for others to follow.

---

<sup>11</sup> Note, for example, the recent news about female suicide bombers in the fall of 2005.

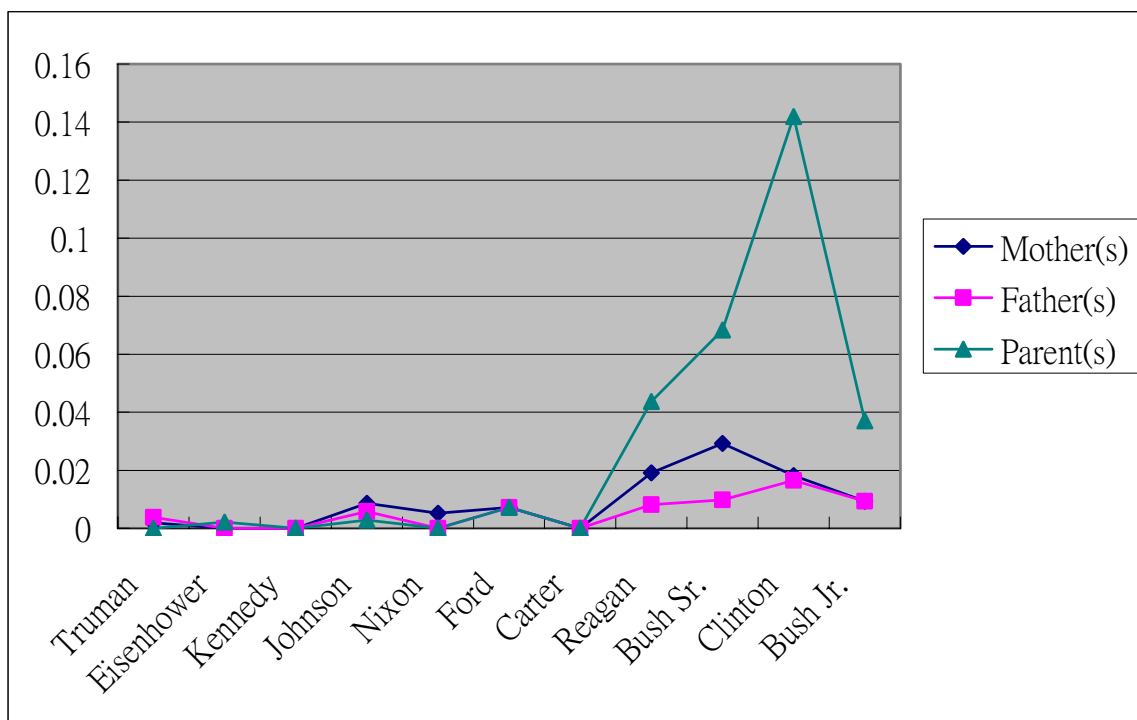
Lastly, the role of motherhood has traditionally been the contribution that women were supposed to play to society. Yet Figures 7 and 8 show that the presidents before Reagan rarely discussed *mother(s)* or *parent(s)*. When presidents talk about *father(s)* it is to mention the Founding Fathers or fathers who went away to war (in terms of the children who are left without a father).



**Figure 7. Tokens of occurrences of father(s), mother(s), and parents**

More recent presidents (Clinton, for example) talk about the responsibilities fathers have to support their families. Mothers, on the other hand, are often mentioned in terms of their age ('young' or 'teen'), work status ('working' or 'poor') or health ('expectant' or as someone who uses drugs). While the sample size is admittedly small, it demonstrates that the fathers are still primarily considered the providers and the mothers are primarily nurturers. Imagine, for example, the use of 'working fathers.' It doesn't appear in the presidential corpus because it is a given. People do not yet prototypically identify fathers as being either working or non-working as they do mothers.

The word *parent*, however, shows a marked increase in usage post-1980, indicating that the issue of child-rearing gained importance in the discussion of domestic policy. Reagan is the first president to use *parent(s)* more than once in a SOU speech, while Clinton uses the word eighty-six times. There is also a clear increase in percentage of use of *parent(s)* between Reagan and Clinton, which drops off sharply with Bush Jr. This most likely has to do with Bush Jr.'s focus on terrorism and foreign policy issues, as opposed to domestic issues.



**Figure 8. Percentage of occurrences of father(s), mother(s), and parents**

## 5. Discussion

In a recent tribute to Betty Friedan, Ellen Goodman [2006] noted that in 1963, Adlai Stevenson told her graduating class at Radcliffe that their education would be important when raising their children. In 1964, the year Friedan's *The Feminist Mystique* was released in paperback, Goodman was working in a gender-segregated research pool at *Newsweek*. And on August 26, 1970, the day women marched down New York's Fifth Avenue to strike for equality, Goodman recalls the front page of the newspaper she worked for showing two pictures of women: "On the left was was the pretty, blond, smiling figurehead of some unknown group of Happy Homemakers. On the right was Betty Friedan, mouth open in mid-shout, face contorted, as unattractive a photo of this woman as was ever chosen by any editor. Under both pictures ran a simple, loaded question: Which one do you choose?" Which one, indeed?

We can see from the data presented here that language use relating to gender changed dramatically in presidential speeches in the time period between 1965 and 1975, with frequent occurrences of biased language use before 1970. If we contrast the language that was used in presidential speeches prior to 1970 with the situation on the ground, we can see that the use of *man* and *mankind* to stand for all human beings reflects a gender-bias in the society and in the workplace that we would nowadays consider completely unacceptable. Yet, at the time, it was

the status-quo.

What, then are the implications for the language use that occurs in current presidential speeches? What implications, for example, can be drawn from the continued use of 'mankind' by recent Republican presidents? Is the use of *mankind* simply a reflection of the interaction between the constraints of stress on multisyllabic words and oratory (*i.e.* 'mankind' is easier to say than 'humankind'), in addition to being an issue of frequency (*i.e.* *mankind* occurs more often than *humankind*)?<sup>12</sup> Or is the continued use of *mankind* by Republican presidents indicative of what Lakoff [1996, 2002] calls "Strict Father Morality" of the conservatives?

## 6. Conclusion

Wodak [2004] notes that "in-depth analyses in empirical data (newspapers, interviews, parliamentary debates) contribute to our theorizing on genre and persuasive language strategies." In addition, we would like to suggest that collecting empirical data, such as presidential speeches, into a diachronic corpus will allow meaning change to be charted over time. This, in turn, will allow corresponding legislative and social events to be contrasted with the linguistic terminology used by politicians.

In this paper, we hypothesized that the language used by politicians would become more inclusive from 1945 to 2006. In some aspects, it certainly did. There was a marked decline in the use of *man* to refer to all people, with a similar decrease in the term *mankind* to refer to all humankind, although its usage still occurs even today. However, references to women did not emphasize the contributions that women have made to society. In the eyes of the presidents, they are still struggling for equal pay and equal rights. Lastly, discussion of issues relating to parenting showed a slight increase in the past twenty years. However, the use of *mother(s)* and *father(s)* did not yet demonstrate the variety of roles that each parent plays in the family and society. Thus, the gains American women have made over the past half a century are not yet reflected in the eyes of the American presidents.

## Acknowledgements

This research was funded by the National Science Council Grants #94-2411-H-002-038 and #95-2918-I-002-007. I would like to thank participants of the 19<sup>th</sup> *Asia-Pacific Conference on Language, Information and Computation*, especially Stephen Bird, Shu-chuan Cheng, Jonathan Evans, Lily I-Wen Su, Siaw-Fong Chung, Louis Wen-lun Lu, and Chu-Ren Huang, as well as two anonymous reviewers, for their comments on an earlier version of this paper. In addition, I would like to thank Director Nicoletta Calzolari and her colleagues at ILC-CNR, Pisa for providing me with research space and computer facilities to write up this paper during

---

<sup>12</sup> Jonathan Evans, personal communication, December 4, 2005.

a portion of my sabbatical from February to May 2006. Thanks also go to Siaw-Fong Chung for assistance in downloading the speeches from C-Span.org and saving them to text files and to Ya-wen Hsieh for typing in Nixon's speeches and aiding in verifying counts in the above analyses. Any errors are my sole responsibility.

## References

- Barrett, A. W., "Gone Public: The Impact of Presidential Rhetoric in Congress," *American Politics Research*, 32(3), 2004, pp. 338-370.
- Charteris-Black, J., *Corpus Approaches to Critical Metaphor Analysis*, London: Macmillan, 2004.
- Charteris-Black, J., *Politicians and Rhetoric: The Persuasive Power of Metaphor*, London: Macmillan, 2005.
- Fairclough, N., *Language and Power*, London: Pearson ESL, 2001.
- Goodman, E., "Betty Friedan, Woman Warrior," *International Herald Tribune*, February 8, 2006.
- Kowal, S., D. O'Connell, K. Forbush, M. Higgins, L. Clarke, and K. D'Anna, "Interplay of Literacy and Orality in Inaugural Rhetoric," *Journal of Psycholinguistic Research*, 26(1), 1997, pp. 1-31.
- Lakoff, G., *Moral Politics: What Conservatives Know that Liberals Don't* [2<sup>nd</sup> edition published as *Moral Politics: How Liberals and Conservatives Think*], Chicago: Chicago University Press, 1996/2002.
- Lim, E., "Five Trends in Presidential Rhetoric: Analysis of Rhetoric from George Washington to Bill Clinton," *Presidential Studies Quarterly*, 2002, pp. 328-366.
- Stubbs, M., *Text and Corpus Analysis*. London: Blackwell, 1996.
- Van Dijk, T. A., "Principles of Critical Discourse Analysis," *Discourse and Society*, 4(2), 1993, pp. 249-283.
- Wodak, R., "New direction in research on political discourse?" *Journal of Language and Politics*, 3(1), 2004, pp. 1-2.

# A Pragmatic Chinese Word Segmentation Approach Based on Mixing Models<sup>1</sup>

Wei Jiang\*, Yi Guan\*, and Xiao-Long Wang\*

## Abstract

A pragmatic Chinese word segmentation approach is presented in this paper based on mixing language models. Chinese word segmentation is composed of several hard sub-tasks, which usually encounter different difficulties. The authors apply the corresponding language model to solve each special sub-task, so as to take advantage of each model. First, a class-based trigram is adopted in basic word segmentation, which applies the Absolute Discount Smoothing algorithm to overcome data sparseness. The Maximum Entropy Model (ME) is also used to identify Named Entities. Second, the authors propose the application of rough sets and average mutual information, etc. to extract special features. Finally, some features are extended through the combination of the word cluster and the thesaurus. The authors' system participated in the Second International Chinese Word Segmentation Bakeoff, and achieved 96.7 and 97.2 in F-measure in the PKU and MSRA open tests, respectively.

**Keywords:** Word Segmentation, N-Gram, Maximum Entropy Model, Rough Sets, Word Cluster, Machine Learning

## 1. Introduction

The word is a logical semantic and syntactic unit in natural language. Unlike English, there is no delimiter to mark word boundaries in Chinese language, so, in most Chinese NLP tasks, word segmentation is the foundational task which transforms the Chinese character string into a word sequence. It is a prerequisite to POS tagging, parser or further applications, such as Information Extraction, and the Question Answer system.

Word segmentation has attracted long-term attention in the research community for more

---

<sup>1</sup> This investigation is supported by the Key Program Projects of National Natural Science Foundation of China (60435020), and the National Natural Foundation of China (60504021).

\* School of Computer Science and Technology, Harbin Institute of Technology, Heilongjiang Province, 150001, P. R. China

E-mail: jiangwei@insun.hit.edu.cn

than two decades. Various methods have been proposed, which fall into two main categories. The first category is made up of rule-based approaches that make use of linguistic knowledge. Cheng [1999] and Liang [1993] described Maximum Forward Match and Maximum Backward Match segmentation. Hockenmaier [1998] and Palmer [1997] used transformation-based error-driven learning. Wu [1998] combined segmentation with a parser and word segmentation became a by-product of the sentence parser. The second category is made up of statistical methods that make use of machine learning algorithms and training on corpus. The typical language model is n-gram [Gao 2002]. Zhang [2003] used the Hierarchical Hidden Markov Model (HMM). In addition, there are some other machine learning methods, such as EM [Peng and Schuurmans 2001], and the channel noise model [Gao 2003]. Sproat [1996] used the WFST method. At present, many state-of-the-art systems use hybrid approaches. Gao [2004] proposed a unified method via the class-based model, and Zhang [2003] presented a unified approach using the Hierarchical Hidden Markov Model. Xue [2003] used Maximum Entropy. Peng [2004] used the Conditional Random Fields model.

Though many methods have been proposed and many improvements have been achieved, as a challenge task, word segmentation is not well-performed. The disambiguation and the out-of-vocabulary (OOV) identification are the main bottlenecks. Due to Zipf's Law, the sparse data problem is rarely avoided, while this problem brings great difficulties in improving the performance of the disambiguation and OOV identification. A meaningful direction for exploration to overcome the sparse data problem is to collect more linguistic knowledge or features and incorporate them into the processing systems.

In this paper, the authors propose to solve the Chinese word segmentation task based on mixing models. The "No Free Lunch Theorem" and "Ugly Duckling Theorem" in Machine Learning theory have indicated that domain knowledge is essential for improving the processing performance. For this reason, different language models will be applied to solve each special sub-task, which is classified according to its linguistic phenomenon and the Natural Language Processing (NLP) technology used in the word segmentation. Another consideration is the pragmatic attribution, *e.g.* some successive processing may require different kinds of balance between precision and efficiency. So, this approach is a pragmatic one, which may incorporate several delicate processing modules, some of which can improve precision by introducing complicated models and utilizing more linguistic knowledge. However, this does result in a decrease in efficiency. Based on the assumption that more delicate linguistic knowledge or some fine linguistic statistical phenomenon can bring information gain to the segmentation task, the authors propose to apply Rough Set Theory and Average Mutual Information, etc. to extract complicated and long distance features. and the authors will also explore combining the word cluster and the thesaurus to extend the features so as to overcome the sparse data problem. This system participated in the Second



International Chinese Word Segmentation Bakeoff (SIGHAN 2005), and a simplified version participated in the SIGHAN 2006.

Section 2 describes the structure of the system. Section 3 describes in detail Named Entity Recognition, which is one of the difficult tasks in word segmentation. Section 4 presents experimental results obtained with the authors' system. Finally, some conclusions will be drawn and direction for future work will be given in Section 5.

## 2. System Description

All words in this system are categorized into five types: Lexicon words (LW), Factoid words (FT), Morphological derived words (MDW), Named entities (NE), and New words (NW). Table 1 shows the tag, description, and some examples for each word type.

**Table 1. The tag, description and examples for each word category**

TAG	Description	Examples
LW	The word in the Lexicon	最近(recent), 博士(doctor), 学位(degree)
FT	Number, Date, Time etc.	2910, 46.12%, 2004年05月12日, 01:06
MDW	Morphological derived words	朋友们(friends), 高高兴兴(happily), 进出口(imports and exports)
NE	Named Entities	孙桂平(Sun Gui-Ping), 哈尔滨(harbin)
NW	The other OOV except FT, MDW, NE	海风牌(sea breeze brand), 古典式(classical), 景观灯(sighting lamp)

To the sentence “同学们下午两点三十分到孙桂平家做客” (Some students visit Sun Gui-Ping in his home at 2:30 p.m.), the segmentation result is “{同学们/[MR\_Suffix]} {下午两点三十分/[TIME]} {到} {孙桂平/[PER]} {家} {做客}”. where the word “同学们/[MR\_Suffix]” is a morphologically derived word, and “下午两点三十分/[TIME]” is a factoid word, all of which can be detected by Segmentation module while “孙桂平/[PER]” is a named entity, and detected in NE Recognition module. Figure 1 shows the structure of this system.

The input character sequence is converted into one or several sentences, which is the basic dealing unit. The internal encoding is UNICODE, and the "Code Convert" module is used to convert the permitted encoding, such as GB2312 and BIG5, into UNICODE. “Basic Segmentation” is used to deal with the LW, FT, MDW words, and “Named Entity Recognition” is used to detect NW words. The authors adopt the New Word Detection algorithm to detect suffix-based new words. The “Disambiguation” module is performed to classify complicated ambiguous words, and all the above results are connected to the final result, namely “word sequence”, which is denoted by XML format. The sequence of each

applied component is decided by the performance of the system. In the following part of this section, the authors will detail the basic theory and the implementation of the system.

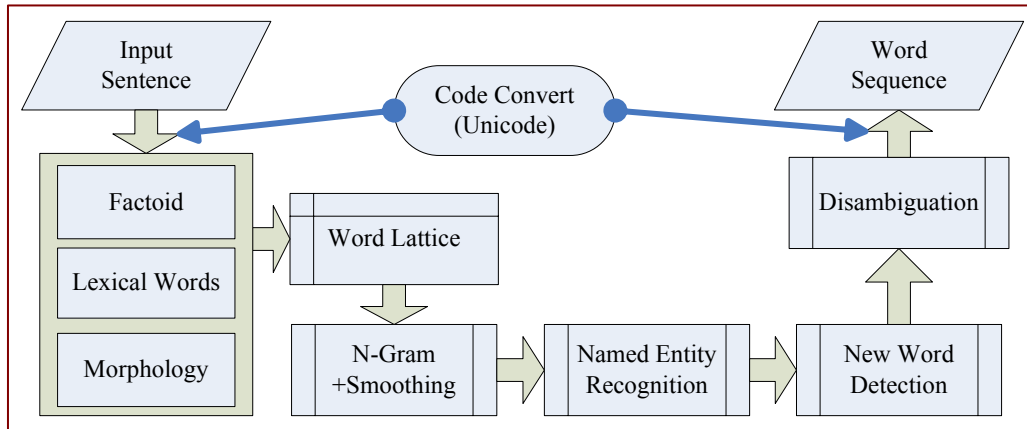


Figure 1. The structure of the proposed system

### 2.1 Trigram and Smoothing Algorithm

The authors apply the Trigram model to the word segmentation task [Jiang 2005; Jiang 2007], and make use of the Absolute Discount Smoothing algorithm to overcome the sparse data problem.

The Trigram model is used to convert the sentence into a word sequence. Let  $\mathbf{w} = w_1 w_2 \dots w_n$  be a word sequence, then the most likely word sequence  $\mathbf{w}^*$  in Trigram is:

$$\mathbf{w}^* = \arg \max_{w_1 w_2 \dots w_n} \prod_{i=1}^n P(w_i | w_{i-2} w_{i-1}), \tag{1}$$

where let  $P(w_0 | w_{-2} w_{-1})$  be  $P(w_0)$  and let  $P(w_1 | w_{-1} w_0)$  be  $P(w_1 | w_0)$ , and  $w_i$  represents LW or a type of FT or MDW. In order to search for the best segmentation way, all the word candidates are filled into the word lattice [Jiang 2006B], as shown in Figure 2, and the Viterbi algorithm is used to search for the best word segmentation path over the built word lattice.

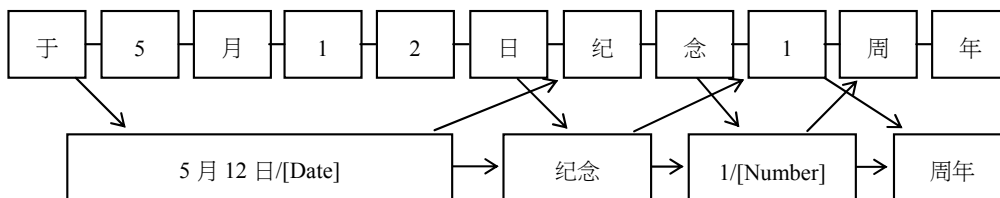


Figure 2. The class-based word lattice in the segmentation task

FT and MDW need to be detected when constructing a word lattice (detailed in section 2.2). The data structure of the lexicon can affect the efficiency of word segmentation, so

lexicon words are represented as a set of TRIEs, which are tree-like structures. Words starting with the same character are represented as a TRIE, where the root represents the first Chinese character, and the children of the root represent the second character, and so on (detailed in section 2.3).

When searching a word lattice, there is a zero-probability phenomenon due to the sparse data problem. For instance, if there is no co-occurrence pair “我们/吃/香蕉”(we eat bananas) in the training corpus, then  $P(\text{香蕉}|\text{我们}, \text{吃}) = 0$ . According to formula (1), the probability of the whole candidate path, which contains “我们/吃/香蕉”, is zero as a result of the local zero probability. In order to overcome the sparse data problem, this system has applied the Absolute Discounting Smoothing algorithm [Chen 1999].

$$N_{1+}(w_{i-n+1}^{i-1} \bullet) = |\{w_i : c(w_{i-n+1}^{i-1} w_i) > 0\}| \quad (2)$$

The notation  $N_{1+}$  is meant to evoke the number of words that have one or more counts, and the  $\bullet$  is meant to evoke a free variable that is summed over. The function  $c()$  represents the count of one word or the co-occurrence count of multi-words. In this case, the smoothing probability can be calculated by the Equation 3.

$$p(w_i | w_{i-n+1}^{i-1}) = \frac{\max\{c(w_{i-n+1}^i) - D, 0\}}{\sum_{w_i} c(w_{i-n+1}^i)} + (1 - \lambda)p(w_i | w_{i-n+2}^{i-1}) \quad (3)$$

where

$$1 - \lambda = \left( \frac{D}{\sum_{w_i} c(w_{i-n+1}^i)} N_{1+}(w_{i-n+1}^{i-1} \bullet) \right) \quad (4)$$

In this trigram model, the maximum  $n$  may be 3. A fixed discount  $D$  ( $0 \leq D \leq 1$ ) can be set through the deleted estimation on the training data. They arrive at the estimate

$$D = \frac{n_1}{n_1 + 2n_2} \quad (5)$$

where  $n_1$  and  $n_2$  are the total number of  $n$ -grams with exactly one and two counts, respectively [Jiang 2006B; Jiang 2007].

After basic segmentation, some complicated ambiguous segmentation can be further disambiguated. In the Trigram model, only the previous two words are considered as context features, while in disambiguation processing (detailed in section 2.4), one can use the Maximum Entropy Model-fused features [Jiang 2006A] or a rule-based method.

## 2.2 Factoid and Morphological Words

As all of the Factoid words can be represented as regular expressions, the detection of factoid words can be achieved by Finite State Automaton (FSA). The categories of factoid words, which can be detected [Jiang 2006B; Jiang 2006D] by this system, are shown in Table 2.

**Table 2. Factoid word categories**

FT type	Factoid word description	Examples
Number	Integer, percent, real etc.	2203, 25.78%, 零点五, 20.542
Date	Date	2004年5月12日, 2004-06-06
Time	Time	8:00, 十点二十分, 晚上6点
English	English word,	Hello, How, are, you
www	Website, IP address	http://www.hit.edu.cn; 192.168.140.133
email	Email	jiangwei@insun.hit.edu.cn
phone	Phone, fax	+86-451-86413322; (0451)86413322

Deterministic FSA (DFA) is efficient because a unique “next state” is determined when given an input symbol and the current state. However, it is common for a linguist to write rules, which can be represented directly as a non-deterministic FSA (NFA), *i.e.* which allow several “next states” to follow a given input and state.

Since every NFA has an equivalent DFA, an FT rule compiler was build to convert all the FT generative rules into a DFA [Jiang 2007]. The rule description is in Table 3.

**Table 3. The demonstration of partial ELUSLex rules**

```

<digit> -> [0..9][ 0 .. 9]; //define Arabic numerals
<integer> ::= {<digit>+}; // define Arabic Integer
<real> ::= <integer>(.|·|点)<integer>; // decimal fraction
<day> -> <integer>日; // define day
<month> -> <integer>月; // define month
<year> -> <digit><integer>年; // define year
<date> ::= <year><month><day>; // define date

```

In order to provide a kind of convenient and powerful description ability, some meta descriptions are assigned to the meta language.

- ✓ Permitted meta rules: <Non-terminator>, terminator, {Loop block}, {Loop block+}, {Loop block\*}, [Range block] (e.g. [a..z], ["a".."z"]), |, (Optional block), (Optional block +), (Optional block \*).
- ✓ Transferred meaning : if the token in the meta rule is the terminator, one needs to transfer its meaning, so one can use double quotation marks to bracket the terminator when it present ambiguity. *e.g.* “(“, “[“, “)“.

- ✓ Rule type: “->” is a temporary generative rule, and “:=” is a real generative rule or a detected rule. This method makes the rule easily written.

The authors built an FT rule compiler to convert all the FT generative rules into a DFA. Obviously, this method makes the system easy to be transferred into a different word segmentation definition, such as from PKU to MSRA. In fact, the authors have used it in SIGHAN 2005 and SIGHAN 2006. Correspondingly, the DFA is represented by the matrix [Jiang 2007], and a run API is provided to make this method easily used. FT detection is important in building the word lattice in word segmentation and also important in the POS tagging task.

The proposed system tries to deal with five main categories of morphologically derived words in real application, the same as Wu [2003] and Gao [2004]: 1) Affixation : 老师们 (teachers), 朋友们(friends); 2) Reduplication: 高高兴兴(happily); 3) Splitting:玩会球(play ball for a while) , 洗了澡(already wash), 吃了饭(already ate); 4) Merging: “进出口” comes from “进口”(importation) and “出口”(exportation); 5) Head Particle: “走出去”comes from “走”(walk) and “出去”(out).

The authors collate the possible MDW into a morphological dictionary from a large corpus, according to the morphological categories mentioned above. Then, some manual selections are needed to select fitting MDW words. As the segmentation specifications of all kinds of corpora are usually different, one needs to collect the corresponding MDW words.

### **2.3 The Data Structure of Lexicon**

The data structure of a lexicon affects the efficiency of word segmentation, as the word candidate in the word lattice is generated through searching the lexicon. When given a sentence string, the candidate comes from matching the substring (starting from the current Chinese character), and judging whether this substring exists in the lexicon. The authors represent lexicon words as a set of TRIEs, which is a tree-like structure. Words starting with the same character are represented as a TRIE, where the root represents the first Chinese character, and the children of the root represent the second characters, and so on, as shown in Figure 3.

The lexical word starts from the “Start state”, and ends in the “End state”. When matching the input sentence and generating the word candidate in the word lattice, each time “End State” is passed, a word candidate is formed and the properties of the current word represented in the “End State” are filled into word lattice.

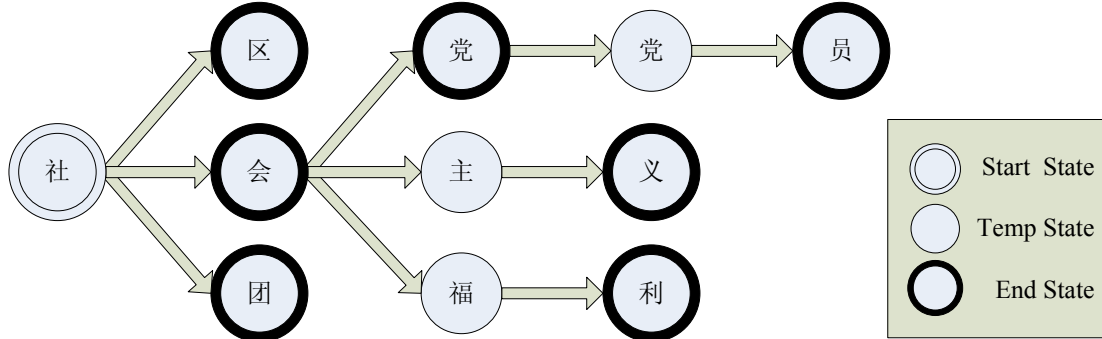


Figure 3. The example of data structure in the lexicon (TRIEs)

Since each Chinese character in the input sentence needs to match the word candidate, the authors build many TRIEs, as shown in Figure 3, to form a lexicon. The example in Figure 3, “社会主义” (socialism), is a word, and this tree is used to match the candidate from the start to the end in the sentence. If one constructs a word lattice in the opposite direction, the tree needs to be built correspondingly, e.g. “义主会社”. This data structure can improve speed in generating the word lattice.

## 2.4 The Disambiguation

It is necessary to effectively exploit the context in the disambiguation process. The authors have proposed using rough sets to extract complicated features and long distance features for disambiguation, which has been reported in previous work [Jiang 2006A]. In that paper, the authors proposed introducing a variable precision Rough Set in feature extraction, in order to acquire a balance of features in disambiguation processing, along with attempting to process complicated and consecutive ambiguity segmentation in the paper. In this paper, the ambiguity segmentations come from the error-total results after evaluating the system.

In Rough Set theory, knowledge is represented via relational tables. An Information System can be defined as follows:  $I = (U, A, V_a, f_a)_{a \in A}$ , where  $U$  is a non-empty set of objects;  $A$  is a non-empty set of attribute  $a$ 's; for each attribute  $a \in A$ , there is an attribute value  $V_a$  set and an information function  $f_a: U \rightarrow V_a$ . An equivalence  $\theta$  on set  $U$  is called an indiscernible relation, and lower approximation for an object set  $X \subseteq U$  is defined as  $\underline{X}\theta = \{\theta x: \theta x \subseteq X\}$ . However, this formula is too strict to fit the requirements of Natural Language Processing. For this reason, the concept of  $\alpha$ -approximation is provided:

$$\underline{X}\theta(\alpha) = \bigcup \left\{ \theta x: \frac{|\theta x \cap X|}{|\theta x|} \geq \alpha \right\}, \text{ where } \alpha \text{ is an external parameter [Jiang 2006A].}$$

When extracting features,  $\alpha$ -Approximation will probably cause unbalanced support, since each segmentation of the ambiguities possibly has disproportionate distribution. In order to let all the features that were added in provide more evidence in guiding toward the correct segmentation,  $\lambda$ -Approximation is introduced in this model. Let filter parameter  $\alpha_d \in [0, 1]$ ,

and the  $n$ -order rough rule set of keyword  $t$  be noted as  $R_t^n$ , then  $R_t^n \in G_{t,n}$ , and defined as:  $R_t^n = \{r \in G_{t,n} \mid r \in \underline{X}_{d,\theta}^{(i)}(\alpha_d)\}$ , where  $n = |A_f| - 1$ ,  $i \in [1, K]$  and  $G_{t,n}$  represents generalized LIT. In  $G_{t,n}$ , indiscernible objects are merged, the objects of each equivalence classes are counted and potential rule precision is calculated. If one lets each  $\alpha_d$  have the same value, namely, let  $\alpha_d = \alpha$  to the decision attribute  $d$ , then  $\lambda$ -Approximation will revert back to the conventional definition of  $\alpha$ -Approximation.

In order to make effective use of contextual knowledge, the authors adopt the Maximum Entropy model (ME), which is a conditional probabilistic model, and relax the feature independent assumption. Disambiguation is regarded as a classifying problem in ambiguous words by the Maximum Entropy model, which is defined over  $H \times T$  in segmentation disambiguation, where  $H$  is the set of possible contexts around the target word that will be tagged, and  $T$  is the set of allowable tags. Then, the model's conditional probability is defined as:

$$p(t|h) = \frac{p(h,t)}{\sum_{t' \in T} p(h,t')} \tag{6}$$

where,

$$p(h,t) = \pi\mu \prod_{j=1}^k \alpha_j^{f_j(h,t)} \tag{7}$$

It has been pointed out that two kinds of ambiguities were dealt with. One is the simple two categories problem, such as “从/小学”(from elementary school) and “从小/学”(study since youth), where the tags are 0 and 1; here 0 represents the first segmentation and 1 represents the second.  $H$  includes the near context and long distance context. The former is comprised of two words around the target word, and the latter features can be obtained by Average Mutual Information, Information Gain, etc.

In fact, a rough statistical result showed that the “one segmentation error” occupied more than 90% of all errors when not considering the errors caused by Named Entity Recognition. Here, “one segmentation error” means that the segmentations surrounding this segmentation error are correct. So, the authors focus on “one segmentation error”, which may be seen in two types of Chinese segmentation ambiguities: overlapping ambiguity and combining ambiguity.

Rough rule features are added in the ME model as a new kind of feature:

$$f_j(a,b) = \begin{cases} 1 & \text{if } ((w = \text{KeyWord}) \text{ and } (A_f(r) = b) \text{ and } (a = d)) \\ 0 & \text{others} \end{cases} \tag{8}$$

where the formula  $A_f(r) = b$  represents that the conditional attribute of  $r$  can be reconstructed in the current context, and  $a = d$  represents the decision attribute of  $d$  is equal

to the tag of ambiguous word. (More details were reported in the paper “[Jiang 2006A]”.)

## 2.5 The Suffix Based New Word Detection

New word (NW) in this system refers to the out-of-vocabulary word that isn't an FT word, MDW word or NE word. The authors do not try to detect all the NW words, since the precision is not satisfactory based on the existing methods in some applications.

On the other hand, in some applications, it is meaningful to recognize some special new words. For instance, ”景观+灯” (sightseeing light), “海风+牌” (Sea Breeze brand). Since some prefixes or some suffixes are paid attention to by this system, such as “现代+化”(modernization), “x+式”(x+way), “x+灯”(x+light), the authors propose to apply a variance algorithm to acquire the prefix or suffix candidate, leaving some minor manual selections possibly required. Hereafter, this paper takes the suffix as an instance, and collects the new words, e.g. “日光+灯” (sunlight), “霓虹+灯” (neon light), “景观+灯” (sightseeing light), etc. Table 4 illustrates the method.

**Table 4. The variance method to obtain the suffix**

	S <sub>1</sub>	S <sub>2</sub>	...	S <sub>m</sub>
W <sub>1</sub>	c <sub>11</sub>	c <sub>21</sub>	...	S <sub>m1</sub>
W <sub>2</sub>	c <sub>12</sub>	c <sub>22</sub>	...	S <sub>m2</sub>
....	...	...	...	...
W <sub>n</sub>	c <sub>1n</sub>	c <sub>2n</sub>	...	S <sub>mn</sub>

Use S<sub>1</sub>..S<sub>m</sub> to represent m candidate suffixes, W<sub>1</sub>..W<sub>n</sub> represent n remained word with the suffix being razed. e.g. S<sub>1</sub> is “灯” (light), then W<sub>1</sub> represents “景观” (sightseeing), W<sub>1</sub>S<sub>1</sub> is the W<sub>1</sub>+S<sub>1</sub>=”景观灯” (sightseeing light). Now, suppose C<sub>xy</sub>=Count(S<sub>x</sub>,W<sub>y</sub>)<sup>2</sup>, and N<sub>xy</sub> is the existence of a co-occurring pair (S<sub>x</sub>,W<sub>y</sub>)<sup>3</sup>, then, one gets the following formula:

$$CV(S_x) = \sum_{i=1}^m N_{xi}, \quad \text{Sum}(S_x) = \sum_{i=1}^m C_{xi}, \quad \text{avg}(S_x) = \text{Sum}(S_x)/CV(S_x),$$

$$p_{xi} = C_{xi}/\text{Sum}(S_x), \quad V_{xi} = p_{xi} * (C_{xi} - \text{avg}(S_x)) * (C_{xi} - \text{avg}(S_x))$$

So, the variance V (S<sub>x</sub>)=  $\sum_{i=1}^m V_{xi}$  .

Besides the variance, one also needs to consider two other factors: (1) the occurrence count in the corpus; (2) the type count that this suffix has constructed words in the lexicon. By considering the above two factors in Sighan2005 evaluation [Jiang 2005], the researchers selected 25 new word suffixes, e.g. 制 (method), 牌 (brand), 型 (type) 、式 (way). These

<sup>2</sup> Here, Count(x,y) represents taking count of the co-occurrence of pair (x,y).

<sup>3</sup> Namely, if C<sub>xy</sub>>0 then N<sub>xy</sub> is 1, else N<sub>xy</sub> is 0.



suffixes also seem to be useful in the Information Retrieval task.

The detection process adopts the Local Maximum Entropy Model, and this process is similar to the NER module [Jiang 2007].

### 3. Named Entity Recognition

Named Entity Recognition (NER) is one of the common message understanding tasks. The objective is to identify and categorize all members of certain categories of "proper names". In MUC-7, there are seven categories: person, organization, location, date, time, percentage, and monetary amount. Named Entities (NE) are broadly distributed in original texts from many domains. In this work, the authors only focus on those more difficult, yet commonly used categories: PER, LOC and ORG. Other NE, such as times and quantities can be recognized simply via Finite State Automata (Section 2.2), and do not need to be aided by a disambiguation algorithm (Section 2.4).

The extensive evaluation of NER systems in recent years (such as CoNLL-2002 and CoNLL-2003) indicates the best statistical systems are typically achieved by using a linear (or log-linear) classification algorithm, such as the Maximum Entropy model, together with a vast amount of carefully designed linguistic features. This still seems true at present in terms of statistics based methods.

In this section, the authors adopt the ME model, which is a linear (or log-linear) classification, to identify the Named Entities, and the focus will be on the utilization of the features [Jiang 2006C]. In addition, the authors propose to build double-layer fixing models to detect the Named Entities, which has also been reported in another paper [Jiang 2007].

The authors use  $w_i$  ( $i=0,1,\dots,n$ ) to denote the input sequence, then every token  $w_i$  should be assigned a tag  $t_i$ . B-I-O encoding, *e.g.*, B-CPN, I-CPN as the beginning of Chinese person's name and the continued part of person's name, respectively, is adopted. Furthermore, in order to improve the ability of describing the rich tagging knowledge, part of the role tags [Zhang 2003] is appended, including the Named Entity prefix, suffix and infix. For example:

✓ 我/O 荣幸/O 地/O 拜访/B-PER\_PREFIX 孙/B-PER 桂/I-PER 平/I-PER  
女士/B-PER\_SUFFIX (Note: It's my honor to visit Ms. Sun Gui-Ping.)

As there are distinct differences between a Chinese person's name and the translation of the person's name in terms of the person construction, the person name is divided into Chinese Person Name (CPN) and the Translation Person Name (FPN). In addition, the authors do not distinguish the type of infix, so the tag number for NER in this system is:  $4 * 4 + 1$  (O) + 1 (INFIX) = 18.

### 3.1 The Context Features

The ‘‘Ugly Duckling Theorem’’ has denoted that there is no generic feature extraction method suitable for all kinds of tasks. The basic feature template is shown in Table 5.

**Table 5. Feature templates for Named Entity Recognition**

Type	Feature Template
one order feature	$w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}$
two order feature	$w_{i-1:i}, w_{i:i+1}$
NER tag feature	$t_{i-1}$

In addition, in order to solve the unstable feature collection problem caused by having no delimiters to separate Chinese words, inspired by the term extraction in text classification, the authors construct a novel feature template of ‘‘word->tag’’ to extract the trigger features, which have a flexible distance between the two units [Jiang 2006C].

Mutual Information (MI) measures the interdependence between a trigger word and a NE type, being defined as:

$$MI(W, C) = \log \frac{P(W \wedge C)}{P(W) \times P(C)} \quad (9)$$

where  $P(W)$  represents the probability of the trigger word, and  $P(C)$  is the probability of the corresponding NE category. However, this method does not consider the influence of lacking one point. In contrast, average mutual information (AMI) is defined as:

$$\begin{aligned} AMI(W, C) = & P(W, C) \log \frac{P(C|W)}{P(C)} + P(W, \bar{C}) \log \frac{P(\bar{C}|W)}{P(\bar{C})} \\ & + P(\bar{W}, C) \log \frac{P(C|\bar{W})}{P(C)} + P(\bar{W}, \bar{C}) \log \frac{P(\bar{C}|\bar{W})}{P(\bar{C})} \end{aligned} \quad (10)$$

MI in fact is point wise information, while AMI can look like a Kullback-Leibler (KL) divergence:

$$AMI(X, Y) = D(P(X, Y) || P(X) \times P(Y)) \quad (11)$$

Equation 11 measures the two different probability distributions between  $P(X, Y)$  and  $P(X) \times P(Y)$ . However, MI is only a point in the whole set of distributions.

Let  $m$  be the number of the possible categories count, the average mutual information is

$$AMI_{avg}(W) = \sum_{i=1}^m P(C_i) \times AMI(W, C_i) \quad (12)$$

or another optional formula adopted in this paper:

$$AMI_{\max}(W) = \text{MAX}_{i=1}^m AMI(W, C_i) \tag{13}$$

The authors select the top triggers with higher AMI value, and acquire the trigger words.

### 3.2 The Entity Features

Besides context features, entity features are also very important in the NER task, such as the suffix of Location or Organization. The authors performed statistical analysis of foreign resources, including the corpora and the collected entity name on the Internet. The authors built 8 kinds of dictionaries:

**Table 6. The resource dictionary for the Named Entity Recognition**

List Type	Lexicon	Examples
Word list	Place lexicon	北京, 纽约, 马家沟
	Chinese surname	张, 王, 赵, 欧阳
	Prefix of PER	老, 阿, 小
String list	Suffix of PLA	山, 湖, 寺, 台, 海
	Suffix of ORG	会, 联盟, 组织, 局
	Character for CPER	军, 刚, 莲, 茵, 倩
Character list	Character for FPER	科, 曼, 斯, 娃, 贝
	Rare character	滢, 肫, 薜

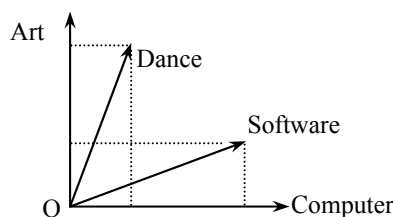
Table 6 gives several kinds of resource dictionaries used in this system. Take the “Suffix of ORG” as an example, the suffix “局”, “组织” is a good hint to detect the Organization Name, so the authors collected them into a “Suffix of ORG” dictionary. When used in the Maximum Entropy Model, this dictionary is used to judge the existing cases of the specified context feature.

### 3.3 The Feature Extension

Feature extension is used to overcome the sparse data problem and to increase robustness. In addition, semantic and pragmatic knowledge is useful in language processing, *e.g.*, if one knows “教授” (professor) is a good hint to label a person’s name, the similar words {老师 teacher), 助教(assistant), 讲师(lecturer)}, should have the same effect. So, one can build a semantic class by combining word clusters and using a thesaurus.

A vector for word *w* is derived from the close neighbors of *w* in the corpus. Close neighbors are all words that co-occur with *w* in a sentence or a larger context. The entry for word *v* in the vector for *w* records the number of times that word *v* occurs close to *w* in the corpus. The authors refer this vector space to as Word Space.

Figure 4 gives a schematic example of two words being represented in a two-dimensional space. This vector representation captures the typical topic or subject matter of a word. By looking at the amount of overlap between two vectors, one can roughly determine how closely they are related semantically. This is because related meanings are often expressed by similar sets of words. Semantically related words will, therefore, co-occur with similar neighbors and their vectors will have considerable overlap.



**Figure 4. A demonstration of word vectors**

The authors combine the basic semantic word in a thesaurus -- HOWNET2005 -- with the TF-IDF algorithm [Zhao 2005B], and use a frequency cutoff to select the 2000 words to serve as the dimensions of Word Space. Compared with the traditional TF-IDF method, this method increases the taxonomical information, so this method can give a better measure of the word similarity.

After constructing word vectors, the similarity can be measured by the cosine between two vectors. The cosine is equivalent to the normalized correlation coefficient:

$$\text{corr}(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^N v_i w_i}{\sqrt{\sum_{i=1}^N v_i^2 \sum_{i=1}^N w_i^2}} \quad (14)$$

The word cluster algorithm in the word vectors is used to measure the similarity by totaling the pragmatic knowledge from the corpora.

#### **4. Evaluation and Discussion**

The authors evaluated the system with two kinds of corpora: 1) The corpora in the International Chinese Word Segmentation Bakeoff; 2) The prior six-month corpora of Peoples' Daily (China) in 1998, which came from Peking University, and have been annotated with lexical tags, including word segmentation, POS tagging, and Named Entity Recognition tags.

#### 4.1 The International Chinese Word Segmentation Bakeoff

This system participated in the Second International Chinese Word Segmentation Bakeoff (SIGHAN-2005) held in 2005, and also participated in SIGHAN-2006.

The performance of ELUS in the SIGHAN-2005 bakeoff is presented in Table 7 and Table 8 respectively, in terms of Recall (R), Precision (P) and F score in percentages. The score software is standard and open by SIGHAN.

**Table 7. Closed test, in percentages (%)**

Closed	R	P	F	OOV	R <sub>ooov</sub>	R <sub>iv</sub>
PKU	95.4	92.7	94.1	5.8	51.8	98.1
MSR	97.3	94.5	95.9	2.6	32.3	99.1
CITYU	93.4	86.5	89.8	7.4	24.8	98.9
AS	94.3	89.5	91.8	4.3	13.7	97.9

This system has good performance in terms of F-measure in the simplified Chinese open test, including the PKU and MSR open tests. In addition, its In-vocabulary word (IV, namely, Lexical words) identification performance is remarkable, ranging from 97.7% to 99.1%, standing at the top or near the top in all the tests in which it has participated.

**Table 8. Open test, in percentages (%)**

Open	R	P	F	OOV	R <sub>ooov</sub>	R <sub>iv</sub>
PKU	96.8	96.6	96.7	5.8	82.6	97.7
MSR	98.0	96.5	97.2	2.6	59.0	99.0
CITYU	94.6	89.8	92.2	7.4	41.7	98.9
AS	95.2	92.0	93.6	4.3	35.4	97.9

This good performance in the R<sub>iv</sub> is due to the class-based Trigram, Absolute Discount Smoothing and Word Disambiguation module with the rough rule features. In this bakeoff, the Name Entity Recognition is a two layer mixing approach, which is reported in detail in a previous paper [Jiang 2007]. The Maximum Entropy Model in the mixing method is similar to that found in Section 3.

The performance of this system in the SIGHAN-2006 bakeoff is presented in Table 9.

**Table 9. MSRA test in SIGHAN2006 (%)**

MSRA	R	P	F	OOV	R <sub>ooov</sub>	R <sub>iv</sub>
Close	96.3	91.8	94.0	3.4	17.5	99.1
Open	97.7	96.0	96.8	3.4	62.4	98.9

The system has good performance in terms of R<sub>iv</sub> measure. The R<sub>iv</sub> measure in a closed test and in an open test was 99.1% and 98.9%, respectively. This good performance is due to a

class-based Trigram with the Absolute Smoothing and Word Disambiguation algorithm.

In this system, the following reasons illustrate why the open test had better performance than the closed test:

(1) Named Entity Recognition module is added into the open test system. And Named Entities, including PER, LOC, ORG, occupy the most of the out-of-vocabulary words.

(2) The system of closed test can only use the dictionary that is collected from the given training corpus, while the system of open test can use a better dictionary, which includes the words that exist in MSRA training corpus in SIGHAN-2005. As is known, the dictionary is one of the important factors that affects the performance, because the LW candidates in the word lattice are generated from the dictionary.

As for the dictionary, the authors compare the two collections in SIGHAN-2005 and SIGHAN2006, and in evaluating the SIGHAN-2005 MSRA closed test. There are less training sentences in SIGHAN-2006. As a result, there is at least a 1.2% performance decrease. So, this result indicates that the dictionary can have an important impact in a system.

## 4.2 The Detailed Evaluation of the System

In this section, some detailed evaluation results are presented. The authors mainly focus on two difficult sub-tasks in the word segmentation task, namely disambiguation and Named Entity Recognition. The measurements in the following experiments include: the precision  $P = \text{the right count} / \text{the model count}$ , the recall rate  $R = \text{the right count} / \text{the corpus count}$ , and  $F\text{-measure} = (2 * P * R) / (P + R)$ .

**Table 10. The comparison experiment for some ambiguities**

Ambiguity	Type	Train Count	Test Count	ME Precision	RS model Precision
才能	才能	704	190	90%	93%
	才/能	7612	300		
不要	不要	1421	150	91%	95%
	不/要	497	80		
从小学	从小学	170	40	88%	91%
	从/小学	260	70		
将来	将来	1200	200	92%	97%
	将/来	35	10		
个人	个人	1016	150	89%	94%
	个/人	819	120		

The authors firstly evaluate the disambiguation performance. Training was done with the preceding five month’s Corpus of the People’s Daily Newspaper, 1998, including 664,805 sentences, and the test corpus was the sixth month corpus, including 136,647 sentences. The authors applied the Rough Set (RS) theory to extract the rough rule features, and fused this theory into the Maximum Entropy Model. The basic feature templates are the  $w_{i-2}$ ,  $w_{i-1}$ ,  $w_i$ ,  $w_{i+1}$ ,  $w_{i+2}$ , furthermore, the rough rule features were fused into the ME disambiguation model [Jiang 2006A], the results are shown in Table 10.

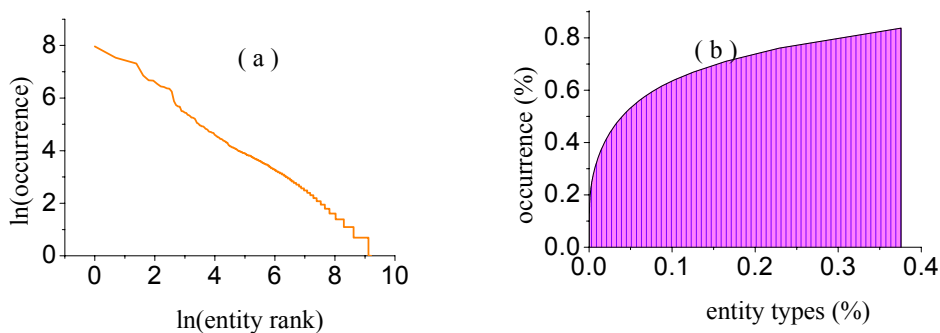
Table 10 demonstrates that RS model may achieve improvement over the baseline ME model. There are at least two main advantages in the proposed method: 1) As a conditional probabilistic model, ME can be fused to more effective features, which relaxes the features independent assumption that is suffered from by the N-Gram model; 2) The authors apply the rough set theory to extract complicated and long distance features. Due to how more effective features are utilized, the new method overcomes the sparse data problem to a certain extent.

Now, the authors evaluate the second group of difficult sub-tasks, namely, the NER module. The experimental corpora also came from the Chinese People’s Daily Newspaper in the first half-year of 1998. The overview of the entity distribution is shown in Table 11.

**Table 11. The entity distribution in People’s Daily**

Named Entity	CPN	FPN	LOC	ORG
By entities	27.54%	8.86%	41.53%	22.07%
By corpus	1.29%	0.41%	1.94%	1.03%
Occur Count	92941	29912	140162	74483

Figure 5 shows that the distribution of the entities complies with the Zipf’s law. As a result, the entities exhibit the sparse property; thereby bringing trouble to the model.



**Figure 5. The entities that exhibit Zipf’s law**

The authors compared several Named Entity Recognition Models, and Table 12 gives the evaluation result. The baseline result is obtained by selecting the NER tag that is most frequently associated with the current word. The authors add several tags in the tag set (Called adding “role”), including the entity prefix, infix and suffix. These tags are used to enhance the ability of the context repetition. In this experiment, HMM is one order model, and ME, CRF use the feature template:  $W_{-2}, W_{-1}, W_0, W_1, W_2, W_{-1:0}, W_{0:1}, T_{-1}$ .

Table 12 indicates that the ME + Role has achieved the best performance. Compared with Hidden Markov Model (HMM), ME can fuse more context features.

**Table 12. The comparison of several NER models**

Model	Precision	Recall	F-measure
BaseLine	68.99%	73.54%	71.19%
HMM	79.20%	79.96%	79.58%
ME	84.77%	83.23%	83.99%
HMM + Role	83.68%	85.20%	84.43%
ME + Role	87.95%	84.62%	86.25%

**Table 13. Trigger pairs draw from corpus**

Pair	AMI		MI	
	Value	Rank	Value	Rank
同志 CPN	3.9e-4	6	2.71	144
说 CPN	2.3e-4	11	1.85	885
主任 ORG	1.2e-4	23	2.63	181
会见 CPN	1.1e-4	27	2.43	269
举行 LOC	9.5e-5	34	1.61	1279
北部 LOC	3.9e-5	80	2.45	271
会议 ORG	3.8e-5	83	1.39	1650
教授 CPN	3.1e-5	96	2.21	463

In another experiment, the authors selected the pairs using two methods, one is to filter by the threshold, such as  $AMI > 0.001$ , the other method is to select the top pair after ranking the pair in descending order, e.g. selecting the top 500 pairs, having the maximum value. The partial pairs are shown in Table 13, including the MI, AMI value and their rank.

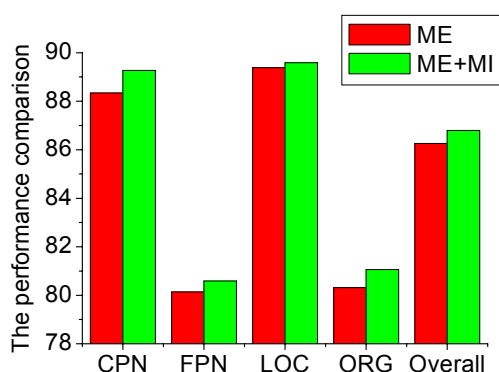
Then, the trigger features were collected, respectively, from above corpora. Taking AMI as an example, after being put in descending order, the top 500 features were selected. Table 14 shows the compared performance with trigger selected by AMI.



**Table 14. The performance with AMI trigger**

Entity type	ME (%)			ME+AMI(%)		
	P	R	F	P	R	F
CPN	84.54	77.71	80.98	86.36	82.41	84.34
FPN	73.27	53.21	61.65	78.50	56.90	65.97
LOC	86.95	76.53	81.41	87.57	77.62	82.30
ORG	74.87	55.29	63.61	74.08	60.95	66.88
Overall	82.81	69.74	75.71	83.60	72.97	77.92

Table 14 gives the detailed comparison between ME and ME with AMI trigger features. The overall improvement is 2.21% in terms of F-measure. Another experiment is done to compare ME with ME + MI model trained by five month corpora. The result is exhibited in Figure 6.

**Figure 6. The comparison about MI in F-measure**

The effectiveness of the proposed method has been confirmed. A similar result is also achieved for the IG approach. Experimental results show that the new method is more efficient.

In the last part of this section, the authors evaluate the word cluster performance. The word vectors method is performed in the large-scale corpora, in the 1998 and 2000 People's Daily Newspaper, the window of size  $k=8$  being used in this experiment.

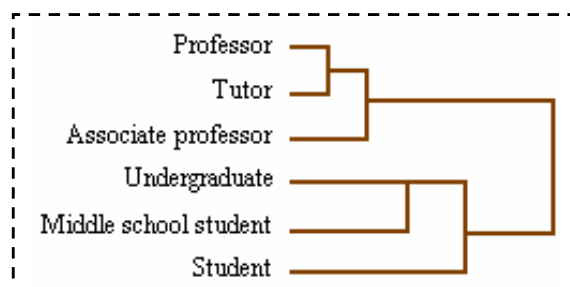
The hierarchical cluster analysis or other cluster analysis methods can be used to obtain the word cluster result. Table 15 demonstrates the proximity matrix, and Figure 7 gives its corresponding hierarchical cluster result. The authors used a synonym dictionary "Word Forest of The Synonym" to reduce the cluster space and increase prior knowledge. For instance, there are about 63 synonyms to the word "教授" (professor).

Though it is helpful to build the word classes for the NER task by combining the word cluster and the thesaurus, some manual correction is also needed, because the linguistic phenomenon is too complicated, therefore making it impossible to acquire all the perfect word

classes by only making statistical analysis of some corpora.

**Table 15. The proximity matrix<sup>4</sup>**

Case	Cosine of Vectors of Values					
	学生	教授	副教授	导师	大学生	中学生
学生	1.000	.352	.280	.288	.433	.331
教授	.352	1.000	.722	.815	.310	.174
副教授	.280	.722	1.000	.641	.216	.136
导师	.288	.815	.641	1.000	.226	.139
大学生	.433	.310	.216	.226	1.000	.674
中学生	.331	.174	.136	.139	.674	1.000



**Figure 7. The demonstration about hierarchical cluster**

Based on the analysis of the errors, one finds that the sparse data problem is the main problem [Jiang 2006A; Jiang 2007]. In this paper, the authors apply the Smoothing Algorithm, Word Cluster Method, etc. to overcome the sparse data problem.

## 5. Conclusion

A pragmatic Chinese word segmentation approach having balance between the precision, efficiency and model complication is described in this paper. The disambiguation and out-of-vocabulary detection are the two main difficulties found in the Word Segmentation task. Accordingly, a lot of work is done in order to improve the performance of the above two problems. The contributions of this research are:

1) Apply multiple models to build a word segmentation model, and a special sub-task can be effectively solved via an optimized language model.

2) The authors propose to apply Average Mutual Information, etc. to extract stable entity features, and also present a novel method to provide an auxiliary function in extending the

<sup>4</sup> 学生 student, 教授 professor, 副教授 associate professor, 导师 tutor, 大学生 undergraduate, 中学生 middle school student.

features by combining the word cluster and the thesaurus.

3) Rough Set theory is present to extract the complicated features and the long distance features for the segmentation disambiguation and for the Named Entity Recognition.

The work in the future will concentrate on two sides: improving the NER performance and exploring New Word Detection Algorithm.

### **Acknowledgements**

The authors thank Dr. Yan Zhao and Dr. Jian Zhao for their valuable suggestions in the proposed system. The authors also thank the members of the Natural Language Computing Group at School of Computer Science and Technology of the Harbin Institute of Technology. The authors especially thank the anonymous reviewers for their insightful comments and suggestions, based on which the paper has been improved.

### **References**

- Chen, S. F., and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, 13(4) 1999, pp. 359-394.
- Cheng, K.-S., G. Young, K.-F. Wong, "A study on word-based and integral-bit Chinese text compression algorithms," *Journal of the American Society for Information Science*, 50(3) 1999, pp. 218-228.
- Gao, J.-F., A.-D. Wu, M. Li, and C.-N. Huang, "Chinese word segmentation and named entity recognition: a pragmatic approach in Computational Linguistics," *Computational Linguistics*, 31(4) 2005, pp.531-574.
- Gao, J.-F., M. Li, A.-D. Wu, and C.-N. Huang, "Chinese Word Segmentation: A Pragmatic Approach," *Microsoft Research, Technical Report: MSR-TR-2004-123*, November 2004.
- Gao, J.-F., J. Goodman, M. Li, and K.-F. Lee, "Toward a unified approach to statistical language modeling for Chinese," *ACMTrans, Asian Language Information Process*, 1(1) 2002, pp. 3-33.
- Gao, J.-F., M. Li, and C.-N. Huang, "Improved source-channel model for Chinese word segmentation," *In the 41nd Annual Meeting of the Association for Computational Linguistics*, 2003, Sapporo, Japan, pp. 272-279.
- Hockenmaier, J., and C. Brew, "Error-driven Learning of Chinese word segmentation," *In the 12th Pacific Conference on Language and Information*, 1998, Singapore, pp. 218-229.
- Jiang, W., X.-L. Wang, Y. Guan, and J. Zhao, "Research on Chinese Lexical Analysis System by Fusing Multiple Knowledge Sources," *Chinese Journal of Computer*, January, 2007.
- Jiang, W., X.-L. Wang, Y. Guan, and G.-H. Liang, "Applying Rough Sets in Word Segmentation Disambiguation Based on Maximum Entropy Model," *Journal of Harbin Institute of Technology (New Series)*, 13(1) 2006A, pp. 94-98.

- Jiang, W., J. Zhao, Y. Guan, and Z.-M. Xu, "Chinese Word Segmentation based on Mixing Model," *In The 4th SIGHAN Workshop*, 2005, Jeju Island, Korea, pp. 180-182.
- Jiang, W., Y. Guan, and X.-L. Wang, "A Pragmatic Chinese Word Segmentation System," *In proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 2006B, Sydney, pp. 189–192.
- Jiang, W., Y. Guan, and X.-L. Wang, "Improving Feature extraction in Named Entity Recognition based on Maximum Entropy Model," *In the 2006 International Conference on Machine Learning and Cybernetics (ICMLC2006)*, 2006C, China, pp. 2630-2635.
- Jiang, W., Y. Guan, and X.-L. Wang, "An Improved Unknown Word Recognition Model based on Multi-Knowledge Source Method," *In the 6th International Conference on Intelligent Systems Design and Applications (ISDA'06)*, vol 2, 2006D, China, pp. 825-830
- Liang, N.-Y., "automatic word segmentation in written Chinese and an auto match word segmentation system-CDWS," (in Chinese) *Journal of Chinese information processing*, 1(2), 1987, pp. 44-52.
- Palmer, D., "A trainable rule-based algorithm to word segmentation," *In proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, 1997, Madrid, Spain, pp. 321-328.
- Peng, F.-C., F.-F Feng, and A. McCallum, "Chinese segmentation and new word detection using conditional random fields," *In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, 2004, Geneva, Switzerland, pp. 562-568.
- Peng, F.C., and D. Schuurmans, "A hierarchical EM approach to word segmentation," *In 6th Natural Language Processing Pacific Rim Symposium (NLPRS-2001)*, 2001, pp. 475-480.
- Schutze, H. "Automatic word sense discrimination," *Computational Linguistics*, 24(1) 1998, pp. 97-123.
- Sproat, R. C. Shih, G. William, and N. Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational linguistics*, 22(3) 1996, pp. 377-404.
- Wu, A.-D., and Z.-X. Jiang, "Word Segmentation in Sentence Analysis," *In 1998 International Conference on Chinese Information Processing*, 1998, Beijing, China, pp. 169-180.
- Xue, N.-W., and L.-B. Shen, "Chinese Word Segmentation as LMR Tagging," *In the Second SIGHAN Workshop on Chinese Language Processing*, 2003, Japan, pp. 176-179.
- Zhang, H.-P., Q. Liu, X.-Q. Cheng, H. Zhang, and H.-K. Yu, "Chinese Lexical Analysis Using Hierarchical Hidden Markov Model," *In the Second SIGHAN workshop affiliated with 4th ACL*, 2003, Sapporo Japan, pp. 63-70.
- Zhao, J., "Research on Conditional Probabilistic Model and Its Application in Chinese Named Entity Recognition," PhD thesis, *Harbin Institute of Technology, China*, 2006.
- Zhao, Y., "Research on Chinese Morpheme Analysis Based on Statistic Language Model," PhD thesis, *Harbin Institute of Technology, China*, 2005A.

Zhao, Y., X.-L. Wang, B.-Q. Liu, and Y. Guan, "Solution Strategies for Word Sense Problems Based On Vector Space Model and Maximum Entropy Model," (In Chinese), *Chinese High Technology Letters*, 15(1) 2005B, pp. 1-6.

