Towards a Principle-Driven Approach to Lexicography

Fusa Katada

School of Science and Engineering, Waseda University
3-4-1 Okubo Shinjuku-ku, Tokyo, Japan 169-8555
katada@waseda.jp

Abstract. This paper proposes that lexicography follow the received view in linguistics science that spoken language, as opposed to written language, offers fundamental characteristics of the faculty of human language. This proposal is in principle universally valid, and should be applicable to agglutinative languages in particular such as Japanese in which morpheme boundaries are not always visible in the written forms, hence when the written forms are adopted, losing one of the fundamentals of human language, computational effectiveness. As a grammatical model on which lexicography to be built, the paper proposes an orthography-free generative grammar which is morphology-based and principle-driven. The lexicography modeled on such a grammatical frame makes it possible to achieve computationally effective lexicon of the language. The script to be used should be phoneme-accommodating and the conversion to the orthography of a given language should be a post-lexical practice in this proposal.

1 Introduction

In any technological development related to many languages and cultures, it is important to formulate development strategies with firm theoretical linguistic foundations. In developing multilingual lexical databases, conceptual and formal foundations need to be first clarified. Among some of the factors to be discussed here are: the problem related to orthography and the problem related to a selection of a theoretical model. I propose that the script to be used should be phoneme-accommodating which allows the lexicon to be universally accessible. Also proposed is an orthography-free grammar, a generative grammar in specific, which realizes the lexicon to be morpheme-based and computationally effective. The idea of a primary lexicon in which feature-coding mechanism is adopted would allow the Papillon lexicon flexible enough to be modified for desired purposes of the users. The conversion of strings of lexical items to language-specific orthography should be done post-lexically.

Along the lines of this proposal, I look to the example of the verb paradigm in Japanese, and submit that due to its computational ineffectiveness, the predominantly influential, but orthography-based traditional grammar of Japanese must regress as a model on which its lexicography is built.

2 Conceptual and Formal Foundations

2.1 The Problem Related to Orthography

As we begin to address problems associated with developing the multilingual lexical databases for several languages, the problem of orthographical diversity comes to mind immediately. This is because the use of some form of the script is unavoidable in dictionary coding. The major newspapers of the world’s political regions show the extensive spread of once the script of Rome, but there are many others including Cyrillic, Arabic, Kanji, Bengali, Kana, Hangul, Thai, Lao [1], to name just a few. The lexicon compiled in the writing system of a given language may stand as authentic document by itself and in its own right; however, it may pose some problems: (i) communicating with other lexicons will be difficult if
they are compiled in varying orthographies, and (ii) computational effectiveness of the lexicon will be lost when morpheme boundaries are buried in the written forms of a language in question, of which Japanese is an example.

It is desirable to use common and phoneme-accommodating script in order to make the lexicons of many languages mutually intelligible and to make them morpheme-oriented so that the computational effectiveness is increased. Practically such script will be Roman alphabet, which has been employed widely to document pre-literate languages of the world.

2.2 The Problem Related to a Theoretical Model

The choice of phoneme-accommodating orthography accords with the notion of ‘knowledge of language’ posed by a generative grammar since Chomsky [2][3]. A generative grammar is a model or representation of grammatical knowledge mentally represented in one’s mind, which should be in the form accessible to a human infant’s pre-literate brain, and which allows one to understand spoken utterances and to verbalize his/her thoughts through speech. It is the knowledge of one’s spoken language that is universally more fundamental to a human mind, and the knowledge of written language is a subject of just another inquiry. It is worth noting in this regard that many languages and cultures of the world are intrinsically oral and non-literate, but not vice versa.

Under this view, the lexicon modeled on orthography-based grammar must regress, provided that a technical alternative based on spoken language is available. This submission is especially relevant in languages such as Japanese which pose discrepancies between spoken forms and written forms in significant ways for the purpose of achieving the effective lexicon. A generative grammar borrowed many of the concepts from mathematics; its aim is to formulate Universal Grammar (UG), a dynamic, computational theory which generates all and only possible sentences of a given language. One of the strengths of a generative grammar is to be able to explain the computationally creative aspect of language production, and thus well fits the nature of the Papillon project.

2.3 Feature Coding and the Idea of the Primary Lexicon

Flexibility. The conventional view of the lexicon has been that it is a list of words of a given language with the following information: phonetic and phonological information concerning pronunciation of the words and phrases, syntactic information concerning parts of speech, transitivity and the like, and semantic information concerning the meaning of the words. Though varying from culture to culture, it is estimated that the lexicon accommodates 45,000 to 60,000 words of a given language in the average person’s mind. If a lexical component is a simple listing of the words, its nature is static, no matter how big it is. The dynamic nature of the structured knowledge of language is left to other components of grammar.

However, the lexicon can be so structured that it can do much more work than a simple listing. How much workload we wish to impose on the lexicon will be decided according to an entire picture of what language model we are designing. Since the Papillon project is to offer lexical databases to be used for varying purposes depending on prospective users, the lexicon to be developed must be not only rich enough in both size and function but also flexible enough so that it can be modified, corrected, or altered from time to time by the users for their own purposes. The flexibility may be achieved by the feature-coding mechanism.

Feature Coding. The lexical items may be entered into the lexicon with their feature specifications. Besides the content meaning of the words that is static in nature, also suggested by generative grammar is to assign phonological, syntactic, and semantic features to lexical items. In this framework, adequate combinations of the word classes for well-formed sentences are detected by sets of syntactic features, called ‘subcategoryization’ frames, assigned to the words co-occurring in a sentence. Similarly, semantic coherence of a sentence is detected by sets of semantic features, called ‘selectional restriction’ frames, assigned to the words co-occurring in a sentence. The following examples in (1) illustrate the basic conception, in which [syntactic features / semantic features] (= [subcategoryization / selectional restriction] frame) are
tentatively assigned.

(1) mother [+noun / +animate]
    baby [+noun / +animate]
    idea [+noun / -animate, -colored]
    colorless [+adj / -color]
    green [+adj / +color]
    furiously [+adv / +emotion]
    sleep [+verb, -transitive / +animate subject, -emotion]

The lexicon will detect the sentence ‘the mother sleeps her baby’ as syntactically ill-formed since the subcategorization frame of ‘sleep’ is specified as [-transitive] and thus disallows the grammatical object to follow: i.e. ‘baby’ in this case which carries the feature [+noun]. Likewise, the lexicon will detect the famous sentence raised by generative grammar: ‘colorless green ideas sleep furiously’ as ill-formed semantically, though not syntactically. That is, the subcategorization frames of the words in adjacent positions do not show any conflicts; however, their selectional restriction frames do; that is, ‘sleep’ takes a [+animate] subject, which is a mismatch with ‘idea’ being [-animate]; the feature [-emotion] of ‘sleep’ is in conflict with its adverbial modifier ‘furiously’ that has [+emotion]; ‘colorless’ and ‘green’ are in conflict having the features [-color] and [+color] respectively.

Features identify natural classes of the lexical items. The more features each lexical item is assigned, the more restricted its distribution will be in relation to other items in a sentence. In other words, feature specifications will allow the lexicon to capture networking relationships of the lexical items. Features can be added or deleted as necessities arise for certain purposes so that distributional restrictions and/or relationships among lexical items may be altered. The lexicon sketched here is more dynamic rather than static; it is flexible enough to evolve from time to time not only in terms of its size but also its syntactic and semantic functions.

The Primary Lexicon. The basic lexicon to be developed which leaves room for its future evolution may be called the ‘primary lexicon’. I propose that the idea of the primary lexicon apply to any language in general and can be used as a general guideline for the Papillon lexicography. Any parameters specific to a certain language or to a certain purpose need to be worked out separately. Lexicographers’ task is then to define what features and how many features to assign to which lexical items of which language in the first place.

3 Morphology-Based Lexicography Made Possible: The Japanese Case

The conventional lexicon of Japanese is modeled on the traditional grammar of the language called Kokugo-gaku. Kokugo-gaku operates on the orthographical system called Kana. Focusing on the description of the verb paradigm for illustration I briefly demonstrate disadvantages of the orthography-oriented conventional lexicon, and propose a phoneme-accommodating alternative that would make the lexicon morphology-based and thus accord more with the computationally effective nature of knowledge of language discussed above.

3.1 The Kana Orthographical System

Japanese has a syllabary-based orthographical system called Gozyu-On-Kana ‘(lit.) 50-sound Kana characters’. The Kana characters in modern Japanese are arranged to form the basic matrix of [5 lines x 10 columns], as in (2). In each cell the Kana character is transliterated into the Roman alphabet called Roomaji, by which the Kana’s phonemic properties are represented. (Among the three slightly different invented forms of Roomaji, the Hepburn system is used here which is
most phoneme-accommodating. The syllabic nasal な[N] is left out of the matrix as irrelevant for the present purpose. For possible accounts for unfilled cells and other derived columns not presented here, see Katada [4].

(2) The Kana Matrix in Modern Japanese

|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|

Each Kana character expresses its corresponding syllabic unit called ‘mora’ whose basic structure is CV. Once acquired, ‘moras’ operate in grammar of Japanese as the psychologically real, smallest units; that is, C is inseparable from V. In other words the Kana matrix is blind to the internal structure of moras [5]. This is precisely what makes the Kana-based traditional grammar of Japanese, Kokugo-gaku, overlook some fundamental characteristics of the language. The verb paradigm of modern Japanese is one area which illustrates this point.

3.2 Descriptions of the Verb Paradigm under Kokugo-gaku

Kokugo-gaku poses six related forms of a verb, given in (3).

(3) Forms for conjugation

- mizen-kei: a form for negation
- renyou-kei: a form connecting to another declinable words
- shuushi-kei: an ending form
- renyou-kei: a form connecting to indeclinable words
- katei-kei: a form for definite conditions
- meirei-kei: a form for command

According to how the stem changes for these forms, three types of verbs given in (4) are recognized. This change is called ‘conjugation’ which operates on lines of the Kana matrix (1). The three types of verbs are decided depending on which column and which line(s) they conjugate. Examples are given in (5).

(4) Three types of verbs

- kami-ichi-dan: [x-column, upper one line]-conjugation verbs
- shimo-ichi-dan: [x-column, lower one line]-conjugation verbs
- yo-dan: [x-column, 4 lines]-conjugation verbs

(5) a. Examples of the verb conjugation paradigm for 6 forms in (3)

- ochiru ‘fall’ [ta-column, upper one (= 1)-line] (= chi), conjugating as:
  - ochi(nai) / ochi(masu) / ochiru / ochiru(koto) / ochiru(ba) / ochiro
wareru 'break (int.)' [ra-column, lower one (= e)-line] (= re), conjugating as:
wareru(nai) / wareru(masu) / wareru / wareru(koto) / warere(ba) / warero

taberu 'eat' [ba-column, lower one (= e)-line] (= be), conjugating as:
taberu(nai) / taberu(masu) / taberu / taberu(koto) / tabere(ba) / tabero

waru 'break (tr.)' [ra-column, 4 lines (= a, i, u, e)] (= ra, ri, ru, re), conjugating as:
waru(nai) / waru(masu) / waru / waru(koto) / ware(ba) / ware

yomu 'read' [ma-column, 4 lines (= a, i, u, e)] (= ma, mi, mu, me), conjugating as:
yomu(nai) / yomi(masu) / yomu / yomu(koto) / yome(ba) / yome

b. Examples of the auxiliary verb conjugation paradigm for 6 forms in (3)

rareru 'spontaneous, potential, honorific, passive' [ra-column, lower one (= e)-line] (= re),
conjugating as: rare(nai) / rare(masu) / rareru / rareru(koto) / rarere(ba) / rareero

saseru 'causative' [sa-column, lower one (= e)-line] (= se), conjugating as:
sase(nai) / sase(masu) / sasew / saseru(koto) / sasere(ba) / sasero

3.3 The Lexical Specifications under Kokugo-gaku

In the conventional lexicon of Japanese, each verbal item is entered in the form ‘shaushi-kei’ (the ending form), accompanied with the conjugation type specified. The dictionary forms are the ending forms entered in the Japanese orthography as in (6), or in the transliterated forms of Roomaji as in (7), which may be convenient for those who are unfamiliar with the Japanese orthography.

(6) 落ちる[た行上一段動詞]
割れる[ら行下一段動詞]
食べる[ば行下一段動詞]
割る[ら行四段動詞]
られる[ら行下一段助動詞]
させる[さ行下一段助動詞]

(7) ochirus: [ta-column, upper one-line] verb
warerus: [ra-column, lower one-line] verb
taberus: [ba-column, lower one-line] verb
warus: [ra-column, 4-line] verb
rarerus: [ra-column, lower one-line]aux. verb
saserus: [sa-column, lower one-line] aux. verb

In whichever form they appear, the fact remains that they are described as Kana-based, and we may never see the internal structure of moras since in this model, the moraic Kana units are the psychologically real smallest units.

3.4 Questions Raised

Kokugo-gaku operating on the Kana matrix constitutes authentic documentation by itself and in its own right for describing the Japanese language. However, questions still arise as to the notion of ‘knowledge of language’ described earlier, especially when the creative language product is wanted. First, lexical items are listed as discrete entries; how they are interconnected to form phrases and sentences is opaque. Second, a morpheme boundary discriminating the verb stem is unclear. Overall, it is hard to see in the form (3) – (7) the dynamic and computational nature of the knowledge. If the
productive nature of the language is not realized, such representations are most likely not in the form logically accessible to human infant’s pre-literate brain. The conclusion to be drawn is that the Kana-based lexicography modeled on Kokugo-gaku is likely to regress.

3.5. Proposing Morphology-Based Lexicography

Conjugation or Agglutination? Conjugational forms which Kokugo-gaku poses are not in the same nature as those in inflectional languages. In these languages such as English or Latin, the verbs alter their forms depending on mood, tense, voice, person, and number, via affixation or vowel alternation called ‘ablaut’. The English irregular verbs such as sing/sang/sung(song) present typical examples.

In this proper sense of conjugation, verbs in Japanese do not conjugate; they simply agglutinate in a certain order to derive surface outputs [7]. The operating notion here is a ‘morpheme’ (the smallest meaningful unit), which is perceptively accessible to a human infant’s pre-literate brain. This view introduces morphemic analyses to the Japanese verb paradigm, in which verbs are formed by several morphemes agglutinating in a certain order, as illustrated in (8). (The initial element in ( ) denotes a dropped sound.)

(8a) ta-be-ru / ta-be-rare-ru / ta-be-sase-ru / ta-be-reba / ta-be- (a)nai / ta-be (i)tagar (r)u
ta-te-ru / ta-te-rare-ru / ta-te-sase-ru / ta-te-reba / ta-te- (a)nai / ta-te (i)tagar (r)u
b. ka-k (r)u / ka-k (r)are-ru / ka-k (s)ase-ru / ka-k (r)eba / ka-k-anai / ka-k-itagar (r)u
war-(r)u / war-(r)are-ru / war-(s)ase-ru / war-(r)eba / war-anai / war-itagar (r)u

Agglutinating patterns are simple and straightforward:

(9a) verb stems are either:
   (i) vowel-ending as in (8a) ([tabe] ‘eat’, [tate] ‘build’), or
   (ii) consonant-ending as in (8b) ([kak] ‘write’, [war] ‘break’)

b. consonant-initial suffixes appear either:
   (i) in the full form when the preceding is V-ending ([ru], [rare], [sase], [reba]), or
   (ii) without the initial consonant when the preceding is C-ending ([a], [are], [ase], [eba])

c. vowel-initial suffixes appear either:
   (i) in the full form when the preceding is C-ending ([anai], [itagar]), or
   (ii) without the initial vowel when the preceding is V-ending ([nai], [tagar])

Note that this approach includes units that are not recognized by Kokugo-gaku; neither consonant-ending verb stems nor the drop of the initial consonant of the suffixes exists since C is inseparable from V in the Kana-based grammar. The patterns described in (9) do not reveal themselves in Kokugo-gaku.

Morphology-Driven Lexicography. The basic idea is the morphology-driven lexicography freed from being Kana-bound. The derived lexicon of Japanese will list verb stems, either V-ending or C-ending; it also lists other suffixal elements as separate entries in their full-forms. As we see in (10), the desired verbal outputs are obtained quite computationally, via agglutination mechanisms such as the deletion of the initial consonant of the suffixes (9b) or the initial vowel of the suffixes (9c).

(10a) mi-rare-ta > mirareta
    b. ka-kare-ru > ka-kare-ru > kakareru
c. *tabe-itagar-ru* > *tabe-(f)tagar-(r)* > *tabetagaru*

(d. *yom-sase-rare-itagar-ani* > *yom-(s)ase-rare-(f)tagar-ani* > *yomasaretagararanai*

Agglutinating patterns (9b/c) do not need to be stipulated; they derive from principles of the well-formedness conditions regulating the structure of moraic units of Japanese: that is, CV. The following are examples among principles and conditions posited by works in linguistics [4].

(11) Mora well-formedness conditions in Japanese

a. a head of a mora is right peripheral,

b. a head of a mora must be [+sonorant],

c. sonorancy within a mora must increase toward a head.

(11) is posited as only part of the lexical principles/conditions of Japanese for the purpose of illustration. Those principles and conditions regulating euphonic changes in the language, for example, should be additionally incorporated in the lexicon [4].

In short, the desired lexicon is morpheme-based; it is a storage of lexical items that are entered in the phonetic/phonological forms, by using phoneme-accommodating Roman alphabet. The lexicon is also equipped with operating principles and conditions that can be effectively drawn from works in linguistics [6][7][8]. The final outputs after agglutination has completed should then be converted to the Japanese orthography, not in the lexicon but at the post lexical level much closer to the surface.

4. Conclusion

This paper proposed the idea of the morphology-based principle-driven primary lexicon. Main ideas are: (i) it should be modeled on grammar of spoken language, (ii) feature coding should be adopted to achieve flexible, dynamic and evoking nature of the lexicon, and (iii) the lexicography utilizing the notion of morpheme accords more with the current linguistic inquires. The proposal is especially applicable to agglutinative languages such as Japanese in which the notion of morpheme is buried under the written forms of the language. The use of phoneme-accommodating script has been proposed. The conversion to language-specific orthography must be a post-lexical practice in this proposal.

References