Establishing a Word Boundary

Ichiro Yuhara
Tokyo Metropolitan University

1 Introduction

I would like to consider an analysis of the following simple transitive sentence 太郎が英語を勉強する ('Taro studies English'). It consists of six words (*taro ga eigo o benkyoo suru*), but for some linguists, especially for those trained in the framework of generative grammar, it seems to be a three-word sentence (*taro-ga eigo-o benkyoo-suru*). What makes this difference? With the multiply ambiguous word "word," I think there are four possible causes of drawing a differing word boundary; (i) Japanese *kana* syllabary, as well as Chinese characters, leaves no space between words (as opposed to the English orthography); (ii) Japanese traditional grammarians (*kokugogakusya*) write their works solely in Japanese, so that virtually no works of theirs are transliterated into English (no phonemic analysis); (iii) almost all linguists who write in English are from the English/American literature departments (where no opportunity of learning the structure of Japanese obtains); and finally (iv) many of them receive their scholastic training at major research institutions in North America, where the emphasis is placed on a mastery of intricate syntactic theories in search of the human language faculty. To put these four facts into other words, few linguists have received a substantive formal training in morphology, thereby mixing free forms (words) and bound forms (affixes), analogous to the English counterparts.[1]

The goal of this short paper is hence educational and remedial; it is to suggest the correct way to define a word boundary. After a brief note on orthographical, phonological, and semantic cues (Sections 2, 3, and 4), I will discuss formal criteria to isolate a word (Section 5) and then apply them to the transitive sentence above (Section 6). Concluding remarks follow (in Section 7).

2 Orthographical Cues

When asked how many words constitute the English expression *happy retirement!*, every one of us will answer "two words." If the logic behind this casual answer lies in the presence of a white space at each end of a word being equivalent of word boundary, that may not be an entirely correct answer. For one, a white space does not necessarily correspond to a slight pause in the airflow that our speech sounds make. No consistency in spelling compounds partially reflects this; *high school ~ high-school ~ highschool*. For another, two plosives, [p] and [t], make a complete obstruction of the breath stream, despite no white space in the middle of word. Then, *happy retirement* is supposed to be a four-word expression **ha ppy re tirement*.[2] Considering orthography is a product of cultural conventions, and our ability to hear boundaries between words are largely auditory hallucination, it is not entirely possible to base a word boundary on orthographical cues.

3 Phonological Cues

In many languages, there is an accentual system (a stress on the first, penultimate, or antepenultimate syllable) or a particular phonological device signaling a phonologically integrated word (ablaut, sequential voicing, sandhi form, vowel harmony); Furthermore, we would be certainly surprised if there is no overlap at all between a stoppage of our egressive airstream and a word boundary. To my knowledge, however, there is no language independent criterion to phonologically isolate a word.

* I would like to thank the editors Céleste Guillemot, Shinichiro Sano, and Seunghun J. Lee for inviting me to contribute to this Festschrift and their editorial assistance in getting this volume published.

[1] I must confess that I was clearly one member of this group until recently.

[2] For acceptability judgments, this paper uses a four-point scale with gradations that it refers to as "perfect (no mark)", "pretty good (?)", "pretty bad (??)", and "horrible (*)".

4 Semantic Cues

Given that we discriminate sounds no better than we label them, how did we come to have the notion "word"? One candidate is the semantic function of a word that caregivers use in our childhood. They use expressions such as *Daddy*, *gone*, and *byebye*, with a clear word boundary, and these words always carry a certain meaning and/or function within and across languages (e.g., referring to entities, describing actions and properties). Although the way we narrow down the range of possible meanings of such words depends much on our intentional state (thus we may later relearn another sense on the face of a discrepancy between our then lexical knowledge and things already evident in the here-and-now context), we first come to connect words with the world and then establish semantic relationships between words. Since we are born to be a part of social world, and language is first of all a system of expression, there is no room for contentless words and bound forms to make their way into our language development. We only hear and speak a semantically coherent whole (e.g., *doggies*, *walked*), which, undoubtedly, becomes the basis of our notion "word."

5 Formal Cues

As we grow and encounter many more expressions in language use, we reach a certain stage where we nurture a different kind of intuition on words. It concerns formal aspects of a word, or whether a sound sequence we recognize with phonological and semantic cues is part of word formation (morphology) or part of phrase formation (syntax). This knowledge on forms holds almost universally; it gives us a more reliable gauge to isolate a word without unhinging its empirical bedrock.

From this formal point of view, a word (but not an affix) is defined as a unit that is capable of making a phrase formation, or "syntactic atom" (Di Sciullo & Williams 1987). Then, a big difference in dividing a word (e.g., *students*) from a phrase (e.g., *the student*) lies in whether we can break a phonemic sequence into two parts and insert a free form, not a bound form, into the middle (I will return to free and bound forms below). If this is possible, as in *the good student*, the two separated phonemic arrays, *the* and *student* here, are respectively a word (i.e., **thestudent*). Likewise, if we can extract a series of phonemes and place it in a distinct position in an utterance (e.g., *I like students ~ students I like*), or if we can completely omit it (e.g., *Did you see two students there? Actually, I saw three ~~students~~*), the apparently moved and/or elided element can be a word. In fact, a phrase can also be displaced and deleted (that is, these formal manipulations are unable to isolate a word alone), but of importance is any bound form tightly concatenates itself with a stem, so that it could neither be detached nor displaced from a word, much less being omitted from a word -- in this respect, the British English words *abso-bloody-lutely* and *ex-fucking-pensive* are highly exceptional.[3] While operations such as Insertion, Movement, and Deletion act exclusively on phrase and sentence formations, they neither analyze nor apply to a unit smaller than a word. We call the inviolable domain from syntactic manipulations "morphologically integrated word," as well as the syntactic atomicity, and almost all words, including compounds and reduplicated forms, have this property. It is for this reason that we cannot say **student and a teacher-s* for the expression *students and a teacher*, by dividing *student-s* into *student* and *-s* and inserting *and a teacher* in between.

At this point, one caveat is in order on "free form" that I used in the above paragraph. Early in the 20th century, Leonard Bloomfield defines a word as "minimum free form" (Bloomfield 1933: 178), and it is considered one of the best characterization of words that is still available today. However, the expression "free" suggests "autonomous," which, in turn, implies being capable of standing on its own as an utterance. The term misleads us, as well as Bloomfield himself, to a hasty conclusion that a considerable number of words, including, but not limited to, articles (*the, a/an, ...*), the copula (*am, are, is, ...*), conjunctions (*and*, *if*, …), and prepositions (*with*, *on, ...*), are not words, for they cannot stand alone in utterances. In fact, they are all full-fledged words (i.e., free forms/syntactic atoms). It is just that these words require a complement to form a phrase. Crucially, this knowledge of ours involves phrase formation, or syntax. That knowledge should not be mixed with our knowledge of word formation, or a bound form being unable to stand on its own in an utterance.[4]

To recapitulate formal criteria for a word, one is whether a phonemic sequence in question can be used autonomously in syntax. This syntactic wordhood usually squares with orthographic, phonological, and semantic cues. Some complications arise with two kinds of elements that are not put into use independently in

[3] My colleague Robert Brock informed me that the two intensifiers that can split a multisyllabic adverb and adjective are *bloody and fucking*. They will even stack: *abso-bloody-fucking-lutely* but only in that order (as *fucking* is the stronger).

[4] In the late 20th century, Noam Chomsky discusses a possibility that we can treat syntax and (part of) morphology in the identical level of analysis.

language; one is bound forms, or affixes (**un-!* for *no*!), and the other is free forms that require a complement to occur in syntax. The former firmly attaches to a stem, forming a phonologically and morphologically unitary word (e.g., *untied*); it never gets separated from its host by a free form (**un almost tied*), and it is neither displaced nor elided from a word, either. On the other hand, the latter must precede or follow a complement, as in *at college* and *three years ago*; the preposition *at* and the postposition *ago* are both words here, despite our intuition saying otherwise. As a consequence, their complements are easily detached from them with another free form(s), as in *at community college* and *three years and a half ago*.

6 Wordhood in Japanese

With this much background, we are now in a position to analyze how many words we need to represent the proposition 太郎が英語を勉強する ('Taro studies English') in Japanese. Some linguists may say "three words" with the following hyphenation in gloss *taro-ga eigo-o benkyoo-suru*. In my view, this is a false analogy from English, or merely counting semantic words, but let us examine them and see whether their morphological integrity is maintained as they are alleged to be so.[5]

First, on two nominals 太郎(が) and 英語(を); if one uses a hyphen, the hyphenated expressions, *taro-ga* and *eigo-o*, are customarily single words. It then follows that が ('NOM') and を ('ACC') are assumed to be suffixes, having respectively attached to the preceding hosts, 太郎 and 英語, without a word boundary. What strikes us as odd is, it is not a problem to divide 太郎が and 英語を with free forms and say 太郎と花子が and 英語とラテン語を. This means that they have undergone a syntactic operation, forming a postpositional phrase. Furthermore, these case markers が and を often drop in response to *wh*-questions such as 誰が英語を専攻しているの？ ('who majors in English?') and 大学で何を勉強したの？('what did you study at college?'), as in 太郎(です) (but not *太郎がです) and 英語(です) (but not ??英語をです). This makes a stark contrast to corresponding answers in flectional languages like Latin, where a case ending is inseparable from a noun stem (e.g., *tarous* and *linguam anglicam*). That が and を are stacked with から (e.g., ここからが難しい), まで (e.g., ここまでを復習して), and か (e.g., 生きるか死ぬかが問題だ；どのように勉強したのかを尋ねてごらん) is another piece of evidence that their category is not so much bound forms as free forms. We may also add here the unacceptable conjunctive expressions *太郎がと花子が and *英語をとラテン語を in support of this conjecture. If case markers in Japanese are truly equivalent to affixes, a conjunctive expression is supposed to connect an inflected word form with an inflected word form (as in *they and I* as well as *them and me*), thereby resulting in an acceptable string, which is contrary to the fact.[6] Finally, a replacement possibility of が and を with an overlaying case such as は ('TOP'), さえ ('even') , and だけ ('only') (e.g., 太郎は英語だけ勉強する) strongly suggests so-called *zyosi* particles in traditional grammar are all postpositional words. It should be recalled that no affix, be it derivational or inflectional, supplants previously assigned bound forms.

Bringing these observations together, I have to conclude that the phonemic sequences 太郎が and 英語を are respectively a two-word expression as in *taro ga* and *eigo o,* as opposed to the widespread (inaccurate) practice of glossing *taro-ga* and *eigo-o*.

Second, on the predicative counterpart 勉強する ('studies'). Since Kuroda's (1965) and Kuno's (1973) foundational pieces in the generative studies of Japanese, many linguists have assumed that a verbal noun, *benkyoo* ('study'), and the verb, *suru* ('do'), are forming one word (i.e., *benkyoo-suru* or *benkyoosuru*). What *suru* follows is actually not limited to verbal nouns, but our concern here is whether 勉強する truly forms one morphological word.[7] (Not) surprisingly, all the tests we have applied to 太郎が and 英語を clearly show that 勉強する consists of two independent words, *benkyoo* and *suru*. It is easily possible to break down 勉強する into 勉強 and する by inserting a postpositional word such as *wa*, *sae*, and *dake* as in *benkyoo sae suru* ('even studies') and *benkyoo dake suru* ('only studies'). Moreover, the expression 勉強をする (now, to be spelled as *benkyoo o suru*) has been recognized very well in the literature as a semantic equivalence of 勉強する. Then, why has the one-word analysis more prevailed than a mere juxtaposition analysis of two words *benkyoo suru*? In my view, it has its source in two assumptions on the side of researchers that (i) *benkyoo o suru* and 勉強する should be accounted for in a derivational relation and (ii) *benkyoo* is a formally noun, which is later moved into the transitive verb *suru* in the similar manner of "noun incorporation" in polysynthetic languages like

[5] If the reader finds something useful in this section, the credit should all go to Ueno (2016).

[6] The unacceptable strings *太郎がと花子が and *英語をとラテン語を in the text show nothing more than が and を being not suffixes. The fact that other postpositions such as *kara* ('from') and *made* ('to') permit conjunctive expressions as in 国分寺からと武蔵境から and 国分寺までと武蔵境まで suggests a lexical category, here postposition, is far from a monolithic theoretical construct.

[7] Included to verbal nouns are V-*tari* (*tabetari*), A-*tari* (*uresikattari*), V-*i*+V-*i* (*yomikaki*), onomatopoeia (*puripuri*), loanwords from English (*oopun*), and clipped words (*kopipe*).

Greenlandic. The unacceptable expression *英語の勉強する is thus taken as circumstantial evidence of "syntactic word formation" (Sadock 1980) -- the complement *eigo no* is stranded after its head noun *benkyoo* being incorporated into the transitive verb *suru* at some point in derivation (i.e., **eigo no _ benkyoo-suru*). (Otherwise, this example may be taken as evidence of "no phrase constraint" on word formation; the noun phrase *eigo no benkyoo* conflicts with the empirical claim for syntax/phrase-free morphology.) For this kind of syntactic analysis, the lack of morphological integrity in *benkyoo sae suru* and *benkyoo dake suru* may be explained by invoking the light verb *suru* that only fulfills the sentence-final position as in *eigo o manabi sae suru* and *eigo o manabi dake suru*. For example, Kageyama (1993) adopts an approach along this line, and argues to the effect that unlike the transitive verb *suru*, the light verb *suru* (underlined above) is merely phonological and as such, unable to assign the case marker *o* to the preceding form, as in *(英語を)学びを(さえ)する and *(先生に)出会いを(さえ)する (*ibid*., 257).

Plausible as it may sound, there are at least two problems that cannot be worked around in this way. One problem is that such an analysis fails to capture the categorial generalization that the case marker *o* is a postpositional word, and *wa*, *sae*, and *dake* among others are formally same in this respect. It leaves unanswered why some postpositions allow incorporation while others do not. The other problem is that case marking, whether it takes a form of declension (*linguam anglicam*) or a phrase formation (*eigo o*), necessarily applies to nominal expressions. It has nothing to do with whether the (light) verb *suru* in question is capable or incapable of assigning a case marker to an adjacent verb or verb phrase (hence *[$_{VP}$先生に出会い]をする/*[$_{VP}$先生に出会う]をする).

As an alternative to the "noun incorporation" analysis, I would like to introduce and put forth the "non-conjugation verb" analysis, which Ueno (2016) recently details in the framework of Automodular Grammar (Sadock 2012).[8] Suppose that *benkyoo* (in *eigo o benkyoo suru*) is not a noun but a non-conjugation verb (NV), and the light verb *suru* only provides its verbal conjugation in taking an NV phrase as its complement (e.g., [$_{VP}$ [$_{NVP}$ *eigo o benkyoo*] *suru*]). Then, the oddity of **eigo no benkyoo suru* is straightforwardly explained in either (i) the light verb *suru* taking not an NV phrase but a noun phrase as its complement (*[$_{VP}$ [$_{NP}$ *eigo no benkyoo*] *suru*]) or else (ii) its complement *eigo* being marked with the postposition *no*, instead of *o*, much like the unacceptability of **eigo no manabu*. In this NV analysis, a double *o* sequence (e.g., **eigo o benkyoo o suru*) cannot possibly occur. This is simply because the light verb *suru* is unable to assign any case marker to nominals. It is clear that *suru* in *eigo no benkyoo o suru* is a transitive verb, taking the postpositional phrase *eigo no benkyoo o* as its complement. The expression *eigo o benkyoo sae suru* neither forms a complex predicate nor undergoes any argument transfer, as proposed in Grimshaw & Mester (1988). It suffices to stipulate that *wa*, *sae*, and *dake* among others take an NV phrase as its complement without passing up its categorial feature (i.e., [$_{VP}$ [$_{NVP}$ [$_{NVP}$ *eigo o benkyoo*] *sae*] *suru*]).

What accounts for the apparently unacceptable expressions such as **eigo o manabi suru* and **sensei ni deai suru*, then? Both *manabi* and *deai* are NVs here, for *eigo o manabi sae suru* and *sensei ni deai dake suru* are perfectly acceptable expressions with *manabi* and *deai* taking the direct objects *eigo* (*o*) and *sensei* (*ni*) respectively. Following many morphologists, I will claim that this is where "blocking effect" in Aronoff (1976) is at work. Our mental lexicon already includes *manab* and *deaw* as lexicalized words, which makes **manabi suru* and **deai suru* semantically superfluous, or possible expressions. Otherwise, the light verb *suru* is allowed to follow any sorts of NV and provide its verbal conjugation as in *tabetari suru*, *uresikattari suru*, *tabetari-nondari suru*, *mukamuka suru*, *oopun suru*, and *kopipe suru*.

Likewise, the cause of the unnatural expressions ??*eigo no manabi o suru* and ??*sensei to no deai o suru* should not be sought in some deep grammatical principle. Given the semantic property of the transitive verb *suru* (i.e., some activity that affects its direct object), it makes a perfect sense that the transitive *suru* is more selective about which nouns (including converted ones) to include in its complement. It seems to me that the odd-or-horrible sounding examples ??*tabetari o suru*, **uresikattari o suru*, ??*tabetari-nondari o suru*, **mukamuka o suru*, ??*oopun o suru*, and ?*kopipe o suru* demand a derivational relation between *benkyoo suru* and *benkyoo o suru* be reassessed on the basis of a declarative relation between *NV suru* and *N o suru*.[9]

7 Concluding Remarks

In this short paper, I have pointed out the widespread practice of analyzing 太郎が英語を勉強する as a

[8] Ueno (2016) gives the linguist Daizaburo Matsushita (1878-1935) credit for this non-conjugation verb analysis.

[9] I by no means intend to claim that all the puzzles of the *suru* constructions are solved here. For one, I have not mentioned an issue of where the goal phrase comes from in the example ?*Taro wa Tokyo e ryokoo o suru* ('Taro travels to Tokyo'). A more serious analysis needs to recognize another two types of *suru*; the pro-verb *suru* (e.g., 花子はよく勉強するが太郎はあまりしない), and the control verb *suru* (日比谷先生をお待ちする).

three-word sentence is incorrect. All the particles that I referred to in this paper (*ga, o*, *kara*, *made, ka*, *to*, *wa*, *sae*, *dake*, *no, ni*, and *e*) are postpositional words. (Traditional grammar correctly groups them as *zyosi* words, but it is not immune to problems, since it also misclassifies affixes such as *-te*, *-temo*, *-tari*, *-tutu*, and *-nagara* into the same category as non-conjugational dependent word class.) I have also discussed that as opposed to some eminent linguists claiming otherwise, 勉強する is a two-word expression, thus *benkyoo suru*, but not *benkyoo-suru*. It is suggested that the technical term "verbal noun" (reportedly, due to the linguist Sameul E. Martin) may be a misnomer, and that there is a promising research line that analyzes it as "non-conjugation verb" in the way that Ueno (2016) resuscitates from the works by Daizaburo Matsuhita. In summary, I conclude 太郎が英語を勉強する is a six-word sentence, nothing more and nothing less.

References

Aronoff, Mark. (1976) *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

Bloomfield, Leonard. (1933) *Language*. Chicago: University of Chicago Press.

Du Scilullo, Anna-Maria. & Edwin Williams. (1987) *On the Definition of Word*. Cambridge, MA: MIT Press.

Grimshaw, Jane. & Armin Mester. (1988) Light verbs and theta-marking. *Linguistic Inquiry* 19(2), 205-232.

Kageyama, Taro. (1993) *Bunpoo to Gokeisei* (文法と語形成). Tokyo: Hitsuzi Syoboo（ひつじ書房）.

Kuno, Susumu. (1973) *The Structure of the Japanese Language*. Cambridge, MA: MIT Press.

Kuroda, S.-Y. (1965) *Generative Grammatical Studies in the Japanese Language*. Ph.D. thesis, MIT.

Sadock, Jerrold M. (1980) Noun incorporation in Greenlandic: A case of syntactic word formation. *Language* 56(2), 300-319.

Sadock, Jerrold M. (2012). *The Modular Architecture of Grammar*. Cambridge: Cambridge University Press.

Ueno, Yoshio. (2016). *Gendainihongo no Bunpookoozoo: Keitaironhen* (現代日本語の文法構造：形態論編). Tokyo: Waseda University Press（早稲田大学出版）.