An Activation-based Sentence Processing Model of English*

Kei TAKAHASHI[†], Kiyoshi ISHIKAWA[‡], and Kei YOSHIMOTO[†]

[†]Tohoku University
{kei-ta,kei}@linguist.jp
[‡]Hosei University and the University of Edinburgh
kiyoshi@i.hosei.ac.jp

Abstract. In this paper, we argue that linear order plays crucial role in an adequate account of acceptability judgment and further that the linear order effects we observe should be seen in terms of real-time processing, rather than static syntax. To capture the linear-order effects, we formulate a memory-based model that predicts fine-grained degrees of acceptability.

1 Introduction

In the theoritical syntactic literature, hierarchical structures have played a central role in accounting for various phenomena. However, that does not mean that linear order has no role to play. Given that sentences are processed from left to right, it would be rather surprising if linear order played absolutely no role when native speakers make (un)acceptability judgments. In fact, some researchers (e.g. Hawkins 1994) have attempted to explain syntactic phenomena in terms of real-time processing. In this paper, we will deal with certain phenomena that could only be explained in terms of word order. Some of the linear order effects we deal with have already been pointed out in the literature, but to the best of our knowledge there exists no formalized account of the observed linear order effects. This paper reports our ongoing attempt to construct a formalized model which is based on the notion of working memory and activation value, and demonstrate how our model successfulluy accounts for the data in question. The structure of this paper is as follows: first, we summarize the kinds of linear order effects to be accounted for by our model, pointing out that they could not be accounted for in structural terms. Then we illustrate the observational generalizations in processing terms. In section 3, we propose a formalized sentence processing model and in section 4, we discuss a relation between our model and (un)acceptabilities. In section 5 we mention how our model parse sentence and in section 6, we demonstrate how our model accounts for the data in question uniformly. In section 6, we mention some remaining problems and concludes the paper.

2 Background

2.1 Problems for Grammar-based Accounts

Kaplan and Bresnan (1982) pointed out the contrast in (1a-b), which is a problem for a movement-based analysis of topicalization.[1] An obvious explanation for the unacceptability of (1b) would be that *about* cannot take a *that*-clause as its complement. In movement-based analyses, topicalization of the *that*-clause is not expected to alter the sentence's (un)acceptable status. However, this expectation is betrayed by (1b).[2]

* This study is supported in part from the Tohoku University 21st Century Center of Excllence Program In Humanities (Stragetic Research and Education Center for an Integrated Approach to Language, Brain and Cognition; The LBC Research Center; http://www.lbc21.jp) and Center for Interdisciplinary Research (CIR, Tohoku University).

[1] There is disagreement among native speakers about whether there is such a contrast. In this paper we follow Kaplan and Bresnan's judgments, on the assumption that a complete grammar of English should *be able to* account for the intuition of those native speakers for whom there is a contrast.

[2] (1) is cited from Bresnan (2000;17)

(1) a. [That he was sick], we talk about for days.
b.*We talk about [that he was sick] for days.

The solution proposed in the LFG literature (Kaplan and Zaenen 1989; Bresnan 2000; Falk 2001) is based on the LFG assumption that complement selection is stated in terms of grammatical function (GF), instead of part of speech (POS).

However, such LFG accounts fail in predicting the contrasts in the following coordination examples.

(2) a. John was thinking about [his girlfriend] and [that he was stupid].
b.*John was thinking about [that he was stupid] and [his girlfriend].

(3) a. Ken agreed with, but John denied, that Mike was wrong.
b.*John denied, but Ken agreed with that Mike was wrong.

(2a-b) differ only in the order of the conjuncts (which are bracketed).[3] Also, the order of conjuncts in (3) alters the (un)acceptability status. Generally, GF is not determined by the order of conjuncts, then we can say that linear order can be a trigger which affects the acceptabilities of sentences. This leads us to assume that it would be plausible to model a theory of acceptability based on linear order.

2.2 The Intuitive Generalization

Through these examples, we can observe one generalization, which we call "The Linear Order Effect"[4]:

(4) **The Linear Order Effect**:
The syntactic requirement the head imposes on an argument is effective only to the extent that the argument is "sufficiently close" to the head in linear order.

This generalization covers (1)-(3) uniformly. For example, the head in (1a) imposes its requirements on *that*-clause since the *that*-clause appears in the cannonical complement position of the prepositional head *about*. In contrast, the *that*-clause in (1b) is sufficiently far from the head so as to evade obeying the syntactic constraints imposed by the head.[5].

It might be possible to propose a non-linear-order-based account for the contrast in (1), as has been done by Kaplan and Bresnan (1982) and Kaplan and Zaenan (1989), but non-linear-order accounts for (2) and (3) are absolutely inconceivable. Moreover, accounts ignoring linear order are absolutely impossible for examples like (5):

(5) a.?Ken was thinking about, (pause) that he was stupid
b. Ken was thinking about, by the way, that he was stupid.

The observation is that the insertion of a pause improves the acceptability (5a), while the insertion of *by the way* makes the sentence fully acceptable (5b). On the standard assumption, a pause and *by the way* only affect timing, not syntax. However, the observation is, at least intuitively, exactly the same as the one we found for (1)-(2); the syntactic head *about* fails to exert its constraints on its complement when the complement is far enough from the head. Thus, syntactic accounts fail to capture our intuitive generalization.

Based on this observation, we propose a memory-based account [6].

[3] At first sight (2a) seems to allow an alternative parse, in which *about his girlfriend* and *that he was stipud* are coordinated. However, we have confirmed that the contrast in (2a-b) remains even when our inteded parses are forced.

[4] A similar generalization has already been pointed out by Moosally (1998) and Sadock (1988). However, they have not formulated a non-stipulatory account. Also note that their observations are limited to coordination examples and therefore do not cover the contrast in (1).

[5] Note that semantic constraints are fully imposed.

[6] This idea is based on the discussion in Takahashi and Ishikawa (2004) and Takahashi and Yoshimoto (2006)

(6) **Memory-based Sentence Processing Model** (General Idea):
The syntactic information is deactivated on the following conditions:
condition (i): when the parser assumes that the predicate-argument structure has been constructed by the parser
condition (ii): passage of processing time or overt phrase is inserted

Condition (i) states that the syntactic information is fully deactivated when the semantic content is assumed to have been obtained. Likewise, condition (ii) states that the syntactic information is gradually deactivated by the passage of time. Note that the degree of deactivation due to condition (i) is not the same degree of deactivation due to condition (ii). This difference is motivated by the assumption that the syntactic information is necessary only for constructing the predicate-argument structure. On the other hand, deactivation due to the passage of time is an unpreferable but unavoidable consequence of the limited capacity of working memory. Also, this model is based on the assumption that syntactic information is needed only to obtain the semantic content, and that the capacity of working memory is severely limited. These lead us naturally to expect that syntactic information is deactivated rapidly from working memory as soon as it has played the role of constructing the semantic content.

When the processor takes in a word, it constructs a corresponding node, and as the tree is successfully constructed, the processor predicts the forthcoming input on a look-ahead basis and expands the tree. If the input is a head, the processor immediately constructs its complement node on a look-ahead basis and imposes various syntactic/semantic constraints on the complement, which has not been actually encountered yet. Note that grammatical knowledge is crucially involved in this process. If a phrase with expected part-of-speech information appears and the predicate-argument content has been constructed, then the whole syntactic information of the PP is deactivated rapidly and only the semantic structure is left.

As stated, this model is not formalized yet. Our next task is to formalize it so that we would be able to implement it on a computer, if one wishes. It is this task that we turn to in the next section.

3 Formalization

Now, let us formalize our memory-based model.

The activation value is represented as a natural number and we define a counter av which calculates the activation value. av takes two arguments; If variables are typed, the type declaration of $av(X, \mathrm{CAT})$ is:

(7) If $\alpha = av(X, \mathrm{CAT})$, then $\alpha \in \mathbb{N}$

X is a node and CAT is a syntactic catogory label of X. In other words, X states the information of the node itself including linguistic expression which X dominates, and the category label of X is CAT. We illustrate a tree annotated by activation values in fig.1.

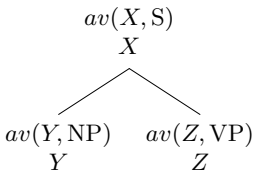

Fig. 1. A tree annotated with activation values for "S → NP VP"

Let n be the number of the possible parsings or ambiguities at a given point, ($av(X, \mathrm{CAT}_1)$ ... $av(X, \mathrm{CAT}_n)$), and T be the sum of the activation values of $av(X, \mathrm{CAT}_i)$.

(8) $$T = \sum_{i=1}^{n} av(X, \mathrm{CAT}_i)$$

Then T is a constant.

Also, we assume that the amount of activation values (T) at a given point is equal or greater than 0 and equal or less than 1.

(9) $$0 \leq T \leq 1$$

In our model if the input to the parser is head, its mother node and the complement node are constructed on a look-ahead basis. For example, given that the input is *think*, its complement node is constructed and two syntactic category labels are possible to be assigned – PP and CP [7]. Given that $av(Z,\mathrm{PP}) = 0.5$, then $av(Z,\mathrm{CP}) = 0.5$ by (8) and (9). If PP is actually encountered to the parser, because the category label is identical to the predicted one, then activation value of the other prediction ($av(Z,\mathrm{CP})$) is transfered to $av(Z,\mathrm{PP})$, resulting in $av(Z,\mathrm{PP}) = 1.0$. More generally, let the cannonical category label CAT_k, another possible label CAT_i, and the actually encountered category CAT_j, then the activation value of node Z is as follows: [8]

(10) $$av(Z, \mathrm{CAT}_j) = av(Z, \mathrm{CAT}_k) + \sum_{i=1}^{n} av(Z, \mathrm{CAT}_i) \quad (j = k, k \neq i)$$

On the other hand, if the input category such as VP is different from the predicted category, its activation value is 0 because the activation value is limited and fully assigned. More generally the activation value of the appeared category which is different from predicted one is:

(11) $$av(Z, \mathrm{CAT}_j) = 1 - \sum_{i=1}^{n} av(Z, \mathrm{CAT}_i) \quad (i \neq j)$$

Note that $av(Z, \mathrm{CAT}_i)$ in (11) represents all the predicted values by the parser.

We have proposed the deactivation of syntactic information which is regulated by two distinct conditions (i) and (ii) in (6). This means that the amount and manner of deactivation differs depending on whether it is caused by (i) or (ii). Furthermore, among those cases where (ii) causes deactivation, we also want to distinguish cases where there is only a silence and cases where there is an overt lingusitic expression.

Thus, we define three deactivation functions. The argument of each function is an activation value. These functions are applied while $av(X, \mathrm{CAT}) > 0$.

(12) a. f: applied when the predicate-argument structure is assumed to have been obtained
b. g: applied by the passage of time
c. h: applied when overt phrases are inserted

When the parser assumes that the semantic content of the relevant part of the sentence has been constructed, f is applied and deactivates the activation value and lowers to 0. The function g is applied whatever the parser is doing and decreases a from the activation value of syntactic information which are in working memory. The third function h is applied when the overt expression such as modifier is inserted and decreases b from the activation value of the selecting head. These functions are defined as follows:[9]

(13) a. $f =_{\mathrm{def.}} \lambda X.\mathtt{id}(X)$
$av(X, \mathrm{CAT}_i) = 0$
b. $g =_{\mathrm{def.}} \lambda X.\mathtt{id}(X)$
$av(X, \mathrm{CAT}_i) = av(X, \mathrm{CAT}_i) - a$
c. $h =_{\mathrm{def.}} \lambda X.\mathtt{id}(X)$
$av(X, \mathrm{CAT}_i) = av(X, \mathrm{CAT}_i) - b$

[7] We use CP for *that*-clause for convenience.
[8] Strictly saying, we use "=" here as substitution, not equality.
[9] The description in (13) is partially refered to `lisp` style. `id` means the identity function.

The function g is defined so that, with an auxiliary assumption, we would be able to model the degree of deactivation through the passage of time. The auxiliary assumption is that, whenever an overt linguistic input is encountered, g is applied *iteratively* to the syntactic category of the node the parser constructs based upon the input on a look-ahead basis. Let c be the time the application of g takes and t be the amount of time that has passed since the node in question has been constructed. Then the the amount of deactivation caused by the iterative application of g is:

(14) $$\text{cycle} =_{\text{def.}} \frac{t}{c}$$

The number of cycle is treated as the coefficient of a in function g. Let $av_0(X, \text{CAT})$ be the activation value of the category lable of a node when it has just been constructed. Then, when n cycles has passed since the construction, given that no other overt linguistic expression has been encountered, the activation value $av_n(X, \text{CAT})$ in question has now become:

(15) $$av_n(X, \text{CAT}_i) = av_0(X, \text{CAT}_i) - na$$

We determine the value of the constants in (13) through psycholinguistic experiments and linear regression analysis of the results.

If the each activation value of possible parsings predicted on a look-ahead basis is deactivated by the application of functions in (12), the activation values of all other categories are averaged and equally raised because the amount of values of predicted categories is decreased to less than 1, then:

(16) $$av(Z, \text{CAT}_m) = \frac{1 - \sum_{i=1}^{n} av(Z, \text{CAT}_i)}{m}$$
where m is the number of categories.

4 Relation between Activation Values and the (Un)acceptabilities

In this section we relate the notion of activation value to the (un)acceptable status of a sentence.

First, let the maximum activation value be 1 and minimum activation value be 0, namely:

(17) $$\forall X, Y[0 \leq av(X, Y) \leq 1]$$

Consider what happens when a *that*-clause is used as a complement of *about*. If the syntactic category of a *that*-clause is CP, and if Y is the node dominating it, and if the "NP" specification on the node is fully active, we have:

(18) a. $av(Y, \text{NP}) = 1$
b. $av(Y, \text{CP}) = 0$

(18a) arises as a result of the look-ahead construction of a node induced by *about*, while (18b) comes from the overtly expressed complement. Clearly, (18a-b) contradict each other, a contradiction that usually results in unacceptability. Here we define a function $\texttt{wf}_1$, which takes in a node and returns its well-formedness, and which obeys (19) (irrespective of the well-formedness of its daugheters; see the definition of $\texttt{wf}_2$ below); the well-formedness is a real number.

(19) $$0 \leq \texttt{wf}_1(Z) \leq 1$$

In the case of the node Y in (18), we assume:

$$\begin{aligned} \texttt{wf}_1(Y) &= av(Y, \text{CP}) \\ &= 1 - av(Y, \text{NP}) \\ &= 1 - 1 = 0 \end{aligned}$$

Note that this value is equal to the activation value of the other possibile parses. On the other hand, if that (18b) is $av(Y, \text{NP}) = 1$, in other words, if the predicted category and the encountered category is identical with each other, then:

$$\begin{aligned}\mathtt{wf}_1(\mathrm{Y}) &= av(Y, \mathrm{NP}) \\ &= 1\end{aligned}$$

More generally, we assume that all the category labels are linearly ordered (for technical convenience) and assigned an integer.

(20) $\mathtt{wf}_1(Z) =_{\mathrm{def.}} av(Z, \mathrm{CAT}_i)$

We distinguish well-formedness values from activation values since we assume that the former is necessary for gaining the whole well-formedness of phrases and this is not altered once the value is obtained, while the former is changed by the conditions in (6).

(20) gives the well-formedness of each node, putting aside the effects of the well-formedness of its daughters. Next we assume another function $\mathtt{wf}_2$, which takes in a node and returns its well-formedness as determined by the well-formedness of its daughters. This function is defined as:

(21) $\mathtt{wf}_2(Z) =_{\mathrm{def.}} \Pi\mathtt{wf}_1(Z_i)$
where Z_i refers to each daughter of Z

That is, the mother's well-formedness is the product of the well-formedness of its daughters. Finally we define yet another function $\mathtt{wf}$ (without a subscript), which takes in a node and returns its well-formedness as determined by *both* its category label specifications *and* the well-formedness of its daughters:

(22) $\mathtt{wf}(Z) =_{\mathrm{def.}} \mathtt{wf}_1(Z) \times \mathtt{wf}_2(Z)$

Thus if $av(Y, \mathrm{NP})$ is not fully active (say, equals 0.3) when the parser learns that the overt complement is a CP, we have:

(23) a. $av(Y, \mathrm{NP}) = 0.3$
b. $av(Y, \mathrm{CP}) = 1 - 0.3 = 0.7$

and hence

$$\mathtt{wf}_1(Y) = av(Y, \mathrm{CP}) = 0.7$$

If Z is the mother node dominating both *about* and the *that*-clause,

$$\begin{aligned}\mathtt{wf}_2(Z) &= \mathtt{wf}_1(Z_1) \cdot \mathtt{wf}_1(Z_2) \\ &= \mathtt{wf}_1(Z_1) \cdot \mathtt{wf}_1(Y) \\ &= 1 \cdot 0.7 \\ &= 0.7\end{aligned}$$

Here Z_1 is the node immediately dominating *about*.[10] If there is no other "label" clash in the sentence, we will also have:

$$\mathtt{wf}(Z) = 0.7$$

which will be "carried over" to the topmost S node, which we equate with the degree of acceptability of the whole sentence.

5 Parsing Process

In this section we roughly mention how our model parse sentence.[11]

In the cannonical sentence, as a non-ambiguous head is read in the parser, only one parse is possible and fully activated. Then as the predicted input is actually appeared, the activation value is 1 without problem. On the other hand, if the unpredicted category is inputted in the parser,

[10] We assume that the $\mathtt{wf}_1$ value of a lexical node projected from a lexical entry is always 1.

[11] In the psycolonguistics and computational linguistics there are other parsing strategies. In this paper, we don't argue them in detail because of the limitation of pages.

the activation value of the input phrase or word is 0 by (11). However, if the processing time has passed without occupying the constructed node, the predicted node is deactivated gradually, then the activation values of unpredicted categories are raised as formalized in (16).

If an ambiguous head is read in the parser, n parses are possible and activation values are assigned to each parse. Note that the amount of each activation value is 1. Then each parse constructs the forthcoming node on a look-ahead basis. As the overt phrase is actually appeared, the activation value of the parse which predicted the same category as the input category is fully activated by (10) and the activation values of unchosen parse are automatically decreased by the constraint (9).

We consider the difference of "highest weighted" parse and the other "lower weighted" parses represents the degree of garden-path effect. In other words, the larger the values of chosen parse and unchosen parses differs, the more the time complexity by disambiguation costs. We consider this corresponds to the delay of reading time on psychological experiments.

6 Demonstrations

6.1 Topicalization

In our memory-based model, (1b) is predicted to be unacceptable for the reason outlined in the end of section 4. On the other hand, (1a) is predicted to be acceptable since the "CP" label of the topicalized phrase has been so deactivated, ($av(X, \mathrm{CP})$ has become 0, if X is the node dominating the topicalized phrase) when the parser encounters *about*, which requires its label to be "NP" ($av(X, \mathrm{NP} = 1)$). Then:

$$\begin{aligned}\mathtt{wf}_1(X) &= av(X, \mathrm{CP}) \\ &= 1 - 0 \\ &= 1\end{aligned}$$

$$\begin{aligned}\mathtt{wf}_2(Y) &= \mathtt{wf}_1(Y_1) \times \mathtt{wf}_1(Y_2) \\ &= \mathtt{wf}_1(Y_1) \times \mathtt{wf}_1(X) \\ &= 1 \cdot 1 \\ &= 1\end{aligned}$$

$$\mathtt{wf}(Y) = 1$$

This value is carried over to S. The degree of acceptability is so high in our model that (5a) is predicted to be acceptable.

Note that the semantic constraints are fully imposed on the topicalized phrase because the semantic content is not deactivated from working memory.

6.2 Complement Coordination

We assume that, when dealing with a constituent coordination structure with two conjuncts, the human parser initially constructs a structure containing only the first conjunct before reading the conjunction (*and*) and after that, it combines with the second conjunct; the structure is subsequently modified into a coordinate structure when the conjunction and the second conjunct are encountered. This assumption is experimentally supported by Sturt and Lombardo (2005).

With this assumption in hand, the pattern in (2) can be explained as follows.

Upon encountering *about*, the parser constructs its complement X with the label "NP" ($av(X, \mathrm{NP})$ =1). Since $av(X, \mathrm{NP}) = 0$ at this point, subsequent coordination of the NP conjunct with the forthcoming CP does not cause a label clash. Thus the sentence is predicted to be acceptable.

6.3 Insertion

In our model, the partially or completely acceptable status of (5a-b) are due to the presence of a time interval between the head and the complement (the pause in (5a)) or the inserted adjunct in (5b). The difference between (5a) and (5b) is the difference between the deactivation values; in (5a), the deactivation value is "$\frac{t}{c}a$" while the degree of deactivation value of the "NP" label in (5b) is "$\frac{t}{c}a + b$". Thus, the acceptability status of (5b) is better than (5a).

6.4 RNR

In our model, the Right Node Raising examples (3a-b) are accounted for as follows. In (3a), *with* and the "raised" phrase *that Mike was wrong* are adjacent, while in (3b) they are not. In our model, the constraints imposed by the head *with* are loosened or deactivated by the application of the function g while processing the second conjunct. Moreover, since the second conjunct is an overt linguistic expression, h is also applied. Hence we have:

$$av_i(X, \mathrm{NP}) = av_0(X, \mathrm{NP}) - \frac{t}{c} \cdot (a + b)$$

where $av_i(X, \mathrm{NP})$ is the av value when the parser encounters the *that*-clause and $av_0(X, \mathrm{NP})$ is the av value immediately after *with* is encountered. This is not a high enough value to cause a label clash with a non-NP label of the "raised" phrase, and hence (3a) is correctly predicted to be acceptable, in contrast to (3b), where the "NP" specification due to the lexical requirement of *with* is active enough when the "CP" specification of the "raised" phrase, resulting in label clash.[12]

7 Conclusion

We have observed certain phenomena that could not be accounted for in structural terms and argued for an observational generalization in terms of linear-order. We formulated the generalization as a memory-based model by defining the activation value, deactivation function, and the relation between the acceptability degrees and the activation values. Also, we reconsidered and succeeded in accounting for the data which are difficult for grammar-based accounts in terms of this formalized model. We think this model is intuitively natural for human sentence processing.

We leave it to our future work (i) to extend the coverage of the present model, and (ii) to extend the model so that it also accounts for various degrees of garden path effects exhibited by various sentences (in various contexts).

References

Bresnan, Joan. 2000. *Lexical-Functional Syntax*. Oxford: Blackwell.

Falk, Yehuda N. 2001. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Stanford: CSLI Publications.

Hawkins, John. 1994. *Performance Theory of Order and Constotuency.* Cambridge: Cambridge University Press.

Kaplan, Ronald M., and Joan Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. *The Mental Representation of Grammatical Relations*, ed. Joan Bresnan, 173–281. Cambridge, MA: MIT Press.

Kaplan, Ronald M., and Annie Zaenan 1989. Long-Distance Dependencies, Constituent Structure, and Functional Uncertainty. *Alternative Conception of Phrase Structure,* eds. Mark R. Bartin and Anthony S. Kroch, 17–42, Chicago: University of Chicago Press.

Moosally, Michelle J. 1999. Subject and Object Coordination in Ndebele: An HPSG Analysis. *WCCFL 18*, eds. S.Bird, A Camnie, J. Haugen, P. Norquest, 379–392. MA: Cascadilla Press.

Sadock, Jerrold M. 1998. Grammatical Tension. *CLS 34: The Panels*, 179–198.

Sturt, Patrick, and Vincenzo Lombardo. 2005. Processing Coordinated Structures: Incrementality and Connectedness. *Cognitive Science* **29**, 291–305.

Takahashi, Kei, and Kiyoshi Ishikawa. 2004. An Adjacecy Constraint on Argument Selection. *Language, Information and Computation: Proceedings of the 18th Pacific Asia Conference*. eds. Hiroshi Masuichi, Tomoko Ohkuma, Kiyoshi Ishikawa, Yasunari Harada, and Kei Yoshimoto. 23–34. Tokyo: The Logico-Linguistic Society of Japan.

Takahashi, Kei, and Kei Yoshimoto. 2006. A Memory-based Sentence Processing Model. *The fifth International Workshop on Evolutionary Cognitive Sciences Human Sentence Processing and Production* (Cosponsored by The Technical Group of Thought and Language (The Institute of Electronics, Information and Communication Engineers)). IEICE Technical Report. TL-2006-11, 25–31.

[12] In the present model, the number of the intervening overt expressions has no effect; all that counts is whether some overt lingusitic expression or other intervenes. This is against our intuition, and we intend to remedy this defect in the future development of our model.