How meaning and predictability affect the duration of Japanese homophonous words

MOTOKI SAITO

Eberhard-Karls-Universität Tübingen
Seminar für Sprachwissenschaft
motoki.saito@uni-tuebingen.de

RUBEN VAN DE VIJVER

Heinrich-Heine-Universität
Institut für Linguistik
ruben.vijver@hhu.de

Although a number of studies have challenged the traditional assumption that there are truly homophonous words, they are exclusively based on English, a language in which duration is not phonemic. This study extends research on the phonetic properties of homophonous words to Japanese, a language in which duration is phonemic. Based on a corpus of spontaneous Japanese speech, we found that 1) homophonous words differed in duration among the members of the homophonous group, 2) durations of homophonous words were positively correlated with degrees of certainty about the predicted word forms predicted from the word meanings, and 3) mora duration was sensitive to the contextual predictability of the mora within the word, while word duration was not sensitive to its contextual predictability.

Keywords: Word duration, meaning, homophones, Japanese, speech production, computational modeling, corpus

1. Introduction

Sounds are used to express meanings, and meanings are differentiated by sounds. In most psycholinguistic models, these two parts of language are separated by at least one other component such as syntax (e.g., Dell 1986, Levelt & Wheeldon 1994, Dell et al. 1997, Levelt et al. 1999, Dell et al. 2007). As a consequence of this extra component, meanings of words cannot be used to directly predict their sounds. If meanings would indeed not predict phonetic properties of a word, it would mean that homophones are produced the same.

Although homophones are not the rarest phenomenon in languages, it remains to be investigated how they are disambiguated. Obviously, homophones are disambiguated by context to some extent. If context is the only cue to disambiguate homophones, they do not have to differ in their phonetic realizations (e.g., Levelt et al. 1999).

However, this assumption has been questioned with evidence from systematic relations between phonetic realizations of words and their part-of-speech properties (Lohmann 2018a), their morphological status (Walsh & Parker 1983, Sproat & Fujimura 1993, Hay 2007, Sugahara & Turk 2009, Smith et al. 2012, Song et al. 2013, Strycharczuk & Scobbie 2016, Zimmermann 2016, Ben Hedia & Plag 2017, Plag et al. 2017, Plag & Ben Hedia 2018, Seyfarth et al. 2018, Li et al. 2020, Schmitz et al. 2021a, Zuraw et al. 2021), frequency of occurrence (Gahl 2008, Lohmann 2018b), and also their meanings (Baayen et al. 2019, Chuang et al. 2021, Saito et al. 2021, Schmitz et al. 2021b, Saito et al. 2023, Gahl & Baayen in press, Saito et al. under revision). For example, homophones have been found to systematically differ from each other, depending on their frequencies (Gahl 2008, Lohmann 2018b), their morphological make-up (Hay 2007, Plag et al. 2017, Schmitz et al. 2021a), and semantic distances between homophonous pairs (Gahl & Baayen in press).

Especially relevant for the topic of the current study, semantics has been found to influence phonetic realizations in several ways. Baayen et al. (2019) found that uncertainty in a word form given the word's meaning led to longer duration of the word. Gahl & Baayen (in press) reported longer duration of homophonous words when the speaker is more certain about the form discriminated by its meaning. Saito et al. (under revision) extended these findings and found that these effects of semantics can also influence coarticulatory tongue movements. In addition, Chuang et al. (2021) showed that semantic effects can also be observed for pseudowords, due to overlaps between pseudowords and existing words in terms of phonology. These observations suggest that phonetic realizations of homophonous words are more differentiated than assumed by traditional models and can contribute to disambiguating homophonous words.

The studies discussed so far have been based exclusively on English. In English, the durational differences between vowels are not phonemic and correlated with vowel quality. As a consequence, vowel length itself is not contrastive. In addition, English is said to be a so-called stress-timed language, in which the intervals between stressed syllables stay roughly stable. In other words, duration can easily be adjusted to accommodate phonological environments in English. It would not be a serious problem to distinguish homophones, because different vowel qualities are enough to discriminate homophonous words from each other.

However, in Japanese, duration is contrastive: 席 [seꜜki] 'seat' vs. 世紀 [seꜜːki] 'century'[1]. In addition, Japanese is a so-called mora-timed language, where each mora occupies roughly the same duration. For example, いい [iꜜː] 'good' is about twice as long as 胃 [i] 'stomach'. As a consequence, duration (especially of each mora) cannot be lengthened or shortened so easily in Japanese as in English. With such phonological constraints, it is still an open question whether, and if so, how Japanese homophonous words are disambiguated. One possibility is that Japanese homophones are disambiguated solely by context

[1] Japanese is a pitch-accent language, and different lexical items can be distinguished by where the pitch drops. The downstep symbol (i.e., ꜜ) indicates the pitch drop. Not all Japanese words have a pitch accent. For those words, the downstep symbol ꜜ is not indicated.

due to tight phonological constraints on vowel length. It is also possible that Japanese homophonous words show systematic durational differences in their phonetic realizations even with the phonological constraints on vowel length. In order to generalize across languages the observations that homophonous words can be phonetically realized with systematically different durations, an investigation in the Japanese language is important. The first aim of the current study is, therefore:

Aim 1: To investigate whether Japanese, a language with stricter phonological constraints on durational differences, would show systematic differences in duration for homophones.

The literature also indicates that homophones are distinguished not only by their (word) durations but also by durations of their constituents, such as morphemes (Sproat & Fujimura 1993, Hay 2007, Sugahara & Turk 2009, Plag et al. 2017, Plag & Ben Hedia 2018, Seyfarth et al. 2018, Schmitz et al. 2021a) and segments (Smith et al. 2012, Ben Hedia & Plag 2017). Systematic differences in duration in units smaller than words are not expected from the traditional perspectives of speech production (e.g., Dell et al. 1997, Levelt et al. 1999). According to traditional models of speech production, durations should not differ systematically according to systematic differences in higher levels such as morphological makeup, as long as the same segmental makeup is shared (as is the case for homophones), whether the duration of interest is the duration of a word or that of its smaller units.

Although such a conceptualization of the speech production process is dominant, some alternative viewpoints have been proposed. The Discriminative Lexicon Model (DLM) (Baayen et al. 2019) is one such model. DLM predicts durations of sublexical units directly from the meaning of words. It, therefore, straightforwardly predicts that homophones can be different in their word- and constituent-durations, as long as they have different meanings. The second aim of the current study is therefore to investigate whether systematic differences in phonetic realizations among homophones are tied to lexical items and how they are stored. If systematic influences from upper levels, such as semantics, affect units smaller than words, as predicted by DLM, homophones should also differ in the durations of their sublexical forms. See Section 2 for more details about DLM.

Aim 2: To investigate if effects of upper-level information such as semantics are tied to lexicality of words.

For these two aims, the current study investigated Japanese homophones with respect to duration of words and moras. The choice of moras was motivated by the fact that Japanese is based on moras, rather than syllables, as is reflected in the Japanese writing system, which uses hiragana and katakana and the On and Kun readings of kanji, all of which are mora-based. Each mora takes up roughly the same duration. Each of word- and mora-duration could significantly differ in durations among homophones. Therefore, four possible outcomes could be possible: 1) word duration and mora duration both significantly differ, 2) only word duration significantly differs, 3) only mora duration significantly differs, and 4) Neither of them shows significant differences in durations. The first possibility is predicted by DLM, and the last possibility is predicted by traditional speech production models. The second and third possible outcomes are not predicted by either of the frameworks.

Table 1. The possible combinations of observations and the predictions by the models.

Outcome	WordDur	MoraDur	PredictedBy
H1	✓	✓	DLM
H2	✓		Neither
H3		✓	Neither
H4			Traditional

2. Discriminative Lexicon Model

2.1. *Word production in DLM*

The Discriminative Lexicon Model (DLM) is a mathematical/computational model of speech comprehension and speech production, based on discriminative learning (Baayen et al. 2011, 2019). In most applications, the model has only three components, which are word forms, word meanings, and associations between them[2].

A form of a word is represented as a vector of numbers in DLM. For example the form vectors of *cat* [kæt], *rat* [ɹæt], and *hat* [hæt] can be defined as follows, where # indicates a word boundary[3]:

[2] More than three components can be conceptualized. For example, visual and auditory inputs could theoretically have their own components. Furthermore, words do not have to be the basic unit on which the model operates. Any other size of linguistic units can be used. It does not have to be based on linguistic units. For example, acoustics can be adopted directly as word forms (Shafaei-Bajestan et al. 2021). It is an open and empirical question what components should be set up in the model of speech comprehension and production and what units should be used in what way

[3] The current study adopted triphone-based representations of word forms, although it is not a requirement of DLM. For more details, see Appendix B.

$$\mathbf{c}_{\text{cat}} = \begin{array}{cccccccc} & \text{\#kæ} & \text{\#ræ} & \text{\#hæ} & \text{kæt} & \text{ɹæt} & \text{hæt} & \text{æt\#} \\ \text{cat} & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{array} \tag{1}$$

$$\mathbf{c}_{\text{rat}} = \begin{array}{cccccccc} & \text{\#kæ} & \text{\#ræ} & \text{\#hæ} & \text{kæt} & \text{ɹæt} & \text{hæt} & \text{æt\#} \\ \text{rat} & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{array} \tag{2}$$

$$\mathbf{c}_{\text{hat}} = \begin{array}{cccccccc} & \text{\#kæ} & \text{\#ræ} & \text{\#hæ} & \text{kæt} & \text{ɹæt} & \text{hæt} & \text{æt\#} \\ \text{hat} & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{array} \tag{3}$$

Similarily, word-meanings are also represented in terms of vectors in DLM. The meanings of *cat*, *rat*, and *hat* may be defined as follows[4]:

$$\mathbf{s}_{\text{cat}} = \begin{array}{ccccc} & \text{<ANIMATE>} & \text{<OBJECT>} & \text{<PREDATOR>} & \text{<PREY>} \\ \text{cat} & 1 & 0 & 1 & 0 \end{array} \tag{4}$$

$$\mathbf{s}_{\text{rat}} = \begin{array}{ccccc} & \text{<ANIMATE>} & \text{<OBJECT>} & \text{<PREDATOR>} & \text{<PREY>} \\ \text{rat} & 1 & 0 & 0 & 1 \end{array} \tag{5}$$

$$\mathbf{s}_{\text{hat}} = \begin{array}{ccccc} & \text{<ANIMATE>} & \text{<OBJECT>} & \text{<PREDATOR>} & \text{<PREY>} \\ \text{hat} & 0 & 1 & 0 & 0 \end{array} \tag{6}$$

Associations between word-meanings and word forms, namely which aspects of meanings are related to which triphones to what extent, can be expressed by a matrix. For the current example with *cat*, *rat*, and *hat*, the association matrix **G** would be as follows:

$$\mathbf{G}_0 = \begin{array}{r|ccccccc|} & \text{\#kæ} & \text{\#ræ} & \text{\#hæ} & \text{kæt} & \text{ɹæt} & \text{hæt} & \text{æt\#} \\ \text{<ANIMATE>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<OBJECT>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREDATOR>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREY>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{7}$$

Each cell in the association matrix represents an association weight from a certain aspect in semantics to a certain triphone. All the cell values in the matrix are

[4] For explanatory reasons we use one-hot encoded semantic vectors, but in our models we will use real valued semantic vectors. The interpretation of real-valued semantic vectors is explained on page 12.

initialized to zero for now, which might be conceptualized as the hypothetical state of linguistic knowledge before learning has been started. To make clear that all the associations are initialized to be zero in the current state (in Equation 7), the suffix '$_0$' is added to '**G**'.

In DLM, word production is conceptualized as predicting word forms based on word-meanings as well as the associations between meanings and forms. This process can be expressed mathematically as follows:

$$\mathbf{s}\mathbf{G}_0 = \hat{\mathbf{c}} \tag{8}$$

The hat (i.e., " ˆ ") indicates that the form vector (i.e., $\hat{\mathbf{c}}$) is a prediction, not a correct word form. In fact, with such a blank weight matrix as laid out in Equation (7), a prediction of a word form will be far away from the correct vector. For example, with the blank matrix, the word form of *hat* will be predicted as follows:

$$\mathbf{s}_{\text{hat}}\,\mathbf{G}_0 = \hat{\mathbf{c}}_{\text{hat}} \tag{9}$$

$$= \begin{array}{c} \\ \text{hat} \end{array} \left[\begin{array}{ccccccc} \text{\#kæ} & \text{\#ræ} & \text{\#hæ} & \text{kæt} & \text{ɹæt} & \text{hæt} & \text{æt\#} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right] \tag{10}$$

The system needs to learn that, in this small example world with only three words, the meaning of <OBJECT> is associated with the triphones #hæ, hæt, and æt#. This can be achieved by considering what errors the system made in predicting the form vector of *hat* in the current example, namely:

$$\mathbf{c}_{\text{hat}} - \hat{\mathbf{c}}_{\text{hat}} = \begin{array}{c} \\ \text{hat} \end{array} \left[\begin{array}{ccccccc} \text{\#kæ} & \text{\#ræ} & \text{\#hæ} & \text{kæt} & \text{ɹæt} & \text{hæt} & \text{æt\#} \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{array} \right] \tag{11}$$

The error vector indicates which sublexical forms should have been predicted how. In Equation (11), it is indicated that #hæ, hæt, and æt# should have been predicted with greater positive values, namely 1. The error vector (Equation 11), however, does not tell us which associations exactly should be fixed. Since this error occurred when the input semantic vector was $\mathbf{s}_{\text{hat}}$, which contained 1 only in the <OBJECT> dimension (see Equation 6), the associations from <OBJECT> to #hæ, hæt, and æt# should be updated. This reasoning can be achieved mathematically by multiplying the error vector with (the transpose of) the input semantic vector $\mathbf{s}_{\text{hat}}$ as below in Equations (12–15).

$$\mathbf{s}_{\text{hat}}^{\top}(\mathbf{c}_{\text{hat}} - \hat{\mathbf{c}}_{\text{hat}}) \tag{12}$$

$$= \begin{array}{r c} & \text{hat} \\ \texttt{<ANIMATE>} & \\ \texttt{<OBJECT>} & \\ \texttt{<PREDATOR>} & \\ \texttt{<PREY>} & \end{array} \!\!\! \left[\begin{array}{c} 0 \\ 1 \\ 0 \\ 0 \end{array}\right] \begin{array}{c c c c c c c} \texttt{\#kæ} & \texttt{\#ræ} & \texttt{\#hæ} & \texttt{kæt} & \texttt{ɹæt} & \texttt{hæt} & \texttt{æt\#} \\ \left[\,0\right. & 0 & 1 & 0 & 0 & 1 & \left.1\,\right] \end{array} \tag{13}$$

$$= \begin{array}{r|c c c c c c c|} & \texttt{\#kæ} & \texttt{\#ræ} & \texttt{\#hæ} & \texttt{kæt} & \texttt{ɹæt} & \texttt{hæt} & \texttt{æt\#} \\ \text{<ANIMATE>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<OBJECT>} & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ \text{<PREDATOR>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREY>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{14}$$

$$= \Delta\mathbf{G}_1 \tag{15}$$

The operation effectively ‘expands’ the row of the error vector from just a single word (i.e., *hat*) to the semantic dimensions. The end result (i.e., $\Delta\mathbf{G}$) has the same shape and dimensions as the weight matrix $\mathbf{G}_0$ (Equation 7) and represents which associations should be fixed how much in which directions. $\Delta\mathbf{G}$, therefore, tells us that the associations from <OBJECT> to #hæ, hæt, and æt# should be fixed in the positive direction by 1. The ‘fix’ of the $\mathbf{G}_0$ matrix is achieved by adding this error matrix (i.e., $\Delta\mathbf{G}$) to the previous state of the $\mathbf{G}$ matrix (i.e., $\mathbf{G}_0$ in Equation 7), namely:

$$\mathbf{G}_0 + \Delta\mathbf{G}_1 \tag{16}$$

$$= \begin{array}{r|ccccccc|} & \texttt{\#kæ} & \texttt{\#ræ} & \texttt{\#hæ} & \texttt{kæt} & \texttt{ɹæt} & \texttt{hæt} & \texttt{æt\#} \\ \text{<ANIMATE>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<OBJECT>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREDATOR>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREY>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{17}$$

$$+ \begin{array}{r|ccccccc|} & \texttt{\#kæ} & \texttt{\#ræ} & \texttt{\#hæ} & \texttt{kæt} & \texttt{ɹæt} & \texttt{hæt} & \texttt{æt\#} \\ \text{<ANIMATE>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<OBJECT>} & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ \text{<PREDATOR>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREY>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{18}$$

$$= \begin{array}{r|ccccccc|} & \texttt{\#kæ} & \texttt{\#ræ} & \texttt{\#hæ} & \texttt{kæt} & \texttt{ɹæt} & \texttt{hæt} & \texttt{æt\#} \\ \text{<ANIMATE>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<OBJECT>} & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ \text{<PREDATOR>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREY>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{19}$$

$$= \mathbf{G}_1 \tag{20}$$

The suffix '$_1$' indicates that it is the updated state from $\mathbf{G}_0$ by one step by adding $\Delta\mathbf{G}$. In this simple example, the updated association matrix $\mathbf{G}_1$ would produce the perfect prediction for the word form of *hat*. However, in a real-world application with much more words, smaller updates in association weights at a time would produce a better performance. The size of updates in association weights can be tuned by adding a sufficiently small number to the error matrix such as $\Delta\mathbf{G}_1$ (Equation 15), e.g. 0.1, as below:

$$\mathbf{G}_0 + \Delta\mathbf{G}_1 \cdot 0.1 \tag{21}$$

$$= \begin{array}{r|ccccccc|} & \text{\#kæ} & \text{\#ræ} & \text{\#hæ} & \text{kæt} & \text{ɹæt} & \text{hæt} & \text{æt\#} \\ \text{<ANIMATE>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<OBJECT>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREDATOR>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREY>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{22}$$

$$+ \begin{array}{r|ccccccc|} & \text{\#kæ} & \text{\#ræ} & \text{\#hæ} & \text{kæt} & \text{ɹæt} & \text{hæt} & \text{æt\#} \\ \text{<ANIMATE>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<OBJECT>} & 0 & 0 & 0.1 & 0 & 0 & 0.1 & 0.1 \\ \text{<PREDATOR>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREY>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{23}$$

$$= \begin{array}{r|ccccccc|} & \text{\#kæ} & \text{\#ræ} & \text{\#hæ} & \text{kæt} & \text{ɹæt} & \text{hæt} & \text{æt\#} \\ \text{<ANIMATE>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<OBJECT>} & 0 & 0 & 0.1 & 0 & 0 & 0.1 & 0.1 \\ \text{<PREDATOR>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{<PREY>} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{24}$$

$$= \mathbf{G}_1 \tag{25}$$

The updating rule described above (Equations 7–25) is summarized below:

$$\mathbf{G}_{t+1} = \mathbf{G}_t + \mathbf{s}_i^\top (\mathbf{c}_i - \hat{\mathbf{c}}_i) \cdot \eta \tag{26}$$

where *i* represents the index of a certain word and η represents a learning rate.

After a sufficient amount of learning events, associations can come to an equilibrium state, where associations do not get updated much any more. The equilibrium state can efficiently estimated by stacking up word form vectors and word-meaning vectors into matrices. The matrix of stacked-up form vectors is conventionally expressed as **C** and that of stacked-up semantic vectors is expressed as **S**.

$$\mathbf{C} = \begin{array}{c} \\ \text{cat} \\ \text{rat} \\ \text{hat} \end{array} \begin{array}{ccccccc} \texttt{\#kæ} & \texttt{\#ræ} & \texttt{\#hæ} & \texttt{kæt} & \texttt{ɹæt} & \texttt{hæt} & \texttt{æt\#} \\ \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array}\right. & \begin{array}{c} 0 \\ 1 \\ 0 \end{array} & \begin{array}{c} 0 \\ 0 \\ 1 \end{array} & \begin{array}{c} 1 \\ 0 \\ 0 \end{array} & \begin{array}{c} 0 \\ 1 \\ 0 \end{array} & \begin{array}{c} 0 \\ 0 \\ 1 \end{array} & \left.\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right] \end{array} \tag{27}$$

$$\mathbf{S} = \begin{array}{c|cccc} & \texttt{<ANIMATE>} & \texttt{<OBJECT>} & \texttt{<PREDATOR>} & \texttt{<PREY>} \\ \hline \text{cat} & 1 & 0 & 1 & 0 \\ \text{rat} & 1 & 0 & 0 & 1 \\ \text{hat} & 0 & 1 & 0 & 0 \end{array} \tag{28}$$

The weight matrix **G** is unknown. The cell values of **G** can be estimated as follows[5]:

$$\mathbf{SG} = \mathbf{C} \tag{29}$$

$$\mathbf{S}^{\top}\mathbf{SG} = \mathbf{S}^{\top}\mathbf{C} \tag{30}$$

$$(\mathbf{S}^{\top}\mathbf{S})^{-1}(\mathbf{S}^{\top}\mathbf{S})\mathbf{G} = (\mathbf{S}^{\top}\mathbf{S})^{-1}\mathbf{S}^{\top}\mathbf{C} \tag{31}$$

$$\mathbf{G} = (\mathbf{S}^{\top}\mathbf{S})^{-1}\mathbf{S}^{\top}\mathbf{C} \tag{32}$$

Though these equations might look daunting to those not versed in matrix algebra, the logic is straightforward. Since we know **S** and **C**, all we need to do is to move them to one side of the equal sign and move **G** to the other side. Just like solving the equation: $4x = 8$. A weight matrix **G** estimated this way (i.e., Equation 32) converges to the equilibrium state which would be reached eventually by learning association weights through a series of events (i.e., Equation 26). This way of estimating association weights (i.e., Equation 32) is also be called the ENDSTATE-LEARNING, focusing on the fact that the endstate (i.e., equilibrium state) of learning is estimated directly.

The **G** matrix for the simple example lexicon above would be as follows:

[5] The weight matrix **G** can also be estimated, using the Moore-Penrose pseudo-inverse, as follows: $\mathbf{G} = \mathbf{S}^{-1}\mathbf{C}$. The estimation by the Moore-Penrose pseudo-inverse can differ from that by Equation (32) only when **S** has more columns than rows, namely when the semantic space is defined by a larger number of semantic dimensions than the number of words.

$$\mathbf{G} = \begin{array}{r|rrrrrrr|} & \texttt{\#kæ} & \texttt{\#ræ} & \texttt{\#hæ} & \texttt{kæt} & \texttt{ɹæt} & \texttt{hæt} & \texttt{æt\#} \\ \textsc{<animate>} & 0.33 & 0.33 & 0.00 & 0.33 & 0.33 & 0.00 & 0.67 \\ \textsc{<object>} & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 1.00 \\ \textsc{<predator>} & 0.67 & -0.33 & 0.00 & 0.67 & -0.33 & 0.00 & 0.33 \\ \textsc{<prey>} & -0.33 & 0.67 & 0.00 & -0.33 & 0.67 & 0.00 & 0.33 \end{array} \tag{33}$$

Conceptually, the weight matrix **G** can be interpreted as a mature linguistic knowledge. For example, in the example **G** matrix above, the meaning of <ANIMATE> is learned to be associated with #kæ and #ræ, namely the word-initial *ca-* and *ra-*. The meaning <ANIMATE> alone cannot determine between *cat* and *rat*. Therefore, the association weights are split between them (i.e., 0.33). In contrast, the meaning of <OBJECT> co-occurs unambiguously with #hæ, namely the word-initial *ha-*. Consequently, the association weight from <OBJECT> to #hæ is 1.00.

With 'mature' linguistic knowledge about associations between word forms and word-meanings, the speaker can produce a word form based on a word-meaning. The production of a word form is expressed in DLM as the multiplication of a certain semantic vector $\mathbf{s}_i$ and the association matrix **G** as below:

$$\mathbf{s}_i\mathbf{G} = \hat{\mathbf{c}}_i \tag{34}$$

Although it is not likely that one speaker has to produce all the words in the language, it would be convenient to have a trained model to produce all the words to see the model's accuracy of predictions. For this purpose, it is possible to give a model semantic vectors one by one, having it produce a single form vector for each semantic vector. However, such a lengthy and redundant process can be simplified by performing the multiplication in Equation (34) with the semantic matrix **S**, instead of semantic vectors, as follows:

$$\mathbf{S}\mathbf{G} = \hat{\mathbf{C}} \tag{35}$$

Note that, in (35), each row of **S** is taken out and multiplied with **G** to produce a form vector in the same row of $\hat{\mathbf{C}}$. Each row in $\hat{\mathbf{C}}$ therefore represents a predicted form vector for a certain word (for the row), based on the word's meaning and the learned associations between word forms and word-meanings. The $\hat{\mathbf{C}}$ matrix for the example lexicon above would be as follows:

$$\hat{\mathbf{C}} = \begin{array}{c} \\ \text{cat} \\ \text{rat} \\ \text{hat} \end{array} \begin{array}{c} \begin{array}{ccccccc} \texttt{\#kæ} & \texttt{\#ræ} & \texttt{\#hæ} & \texttt{kæt} & \texttt{ɹæt} & \texttt{hæt} & \texttt{æt\#} \end{array} \\ \left[\begin{array}{ccccccc} 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{array}\right] \end{array} \tag{36}$$

For this small example lexicon, the predicted form vectors in $\hat{\mathbf{C}}$ are all exactly the same as the correct (gold-standard) form vectors defined in **C**, namely 100% of prediction accuracy.

In the example above, word-meanings were defined only with 1 and 0, based on hand-made semantic dimensions. However, it is not realistic and highly likely to be impracticable to define enough semantic dimensions in order to distinguish all the words in a language. Since the current study is mainly concerned with speech production, rather than defining a complete set of semantic features, it is enough and better from the perspective of replicability to define semantics with an algorithm that has been known to be able to approximate enough what we call 'meanings', namely a word-embedding model (Landauer & Dumais 1997) such as fastText (Bojanowski et al. 2017). Consequently, semantic vectors and a semantic matrix made of those semantic vectors are all real-valued. The interpretation of each number is the strength of the association of a word with a semantic dimension. As a word embedding customarily has a vector of length 300, each of which represents a semantic dimension, it is no longer possible to interpret each semantic dimension. However, the whole process and interpretations explained above, regarding the production process in DLM, are all identical.

2.2. *Word comprehension in DLM*

The previous section was concerned with the word-production process in DLM, where word forms were produced based on word-meanings. The comprehension process in DLM operates in the opposite direction, namely to produce a word-meaning given a word form. The comprehension process has its own association weight matrix, which is conventionally called **F**. The weight matrix **F** has sublexical forms as rows and semantic dimensions as columns, and it can conceptually be understood as a learned linguistic knowledge about what word forms mean. The weight matrix **F** can be estimated in the same way explained in the previous section for the production weight matrix **G**, namely either incrementally, in which associations are learned word by word, or analytically, in which an equilibrium state of **F** is estimated at once. The latter way of estimating F is illustrated below.

$$\mathbf{CF} = \mathbf{S} \tag{37}$$

$$\mathbf{C}^{\top}\mathbf{CF} = \mathbf{C}^{\top}\mathbf{S} \tag{38}$$

$$(\mathbf{C}^{\top}\mathbf{C})^{-1}(\mathbf{C}^{\top}\mathbf{C})\mathbf{F} = (\mathbf{C}^{\top}\mathbf{C})^{-1}\mathbf{C}^{\top}\mathbf{S} \tag{39}$$

$$\mathbf{F} = (\mathbf{C}^{\top}\mathbf{C})^{-1}\mathbf{C}^{\top}\mathbf{S} \tag{40}$$

The **F** matrix for the example above with *cat*, *rat*, and *hat* would be as follows. Note that the associations in **F** and **G** are not identical.

$$\mathbf{F} = \begin{array}{l c c c c} & \text{<ANIMATE>} & \text{<OBJECT>} & \text{<PREDATOR>} & \text{<PREY>} \\ \texttt{\#kæ} & 0.30 & -0.10 & 0.40 & -0.10 \\ \texttt{\#ræ} & 0.30 & -0.10 & -0.10 & 0.40 \\ \texttt{\#hæ} & -0.20 & 0.40 & -0.10 & -0.10 \\ \texttt{kæt} & 0.30 & -0.10 & 0.40 & -0.10 \\ \texttt{ræt} & 0.30 & -0.10 & -0.10 & 0.40 \\ \texttt{hæt} & -0.20 & 0.40 & -0.10 & -0.10 \\ \texttt{æt\#} & 0.40 & 0.20 & 0.20 & 0.20 \end{array} \tag{41}$$

With an estimated **F**, the listener can ‘understand’ a word-meaning from the sublexical forms of the word. Using the matrix notation, predicted meanings from word forms by the model can be expressed as follows. The $\hat{\mathbf{S}}$ matrix for the example lexicon above will be identical to **S**, namely the perfect accuracy, as was the case for the estimation of $\hat{\mathbf{C}}$.

$$\mathbf{CF} = \hat{\mathbf{S}} \tag{42}$$

2.3. *Production of homophones in DLM*

DLM predicts that homophones are easier to produce while difficult for comprehension. To demonstrate, suppose another tiny toy lexicon, which contains only two Japanese words. They are 書く [kakɯ] ‘write’ and 角 [kakɯ] ‘a certain type of a piece in the Japanese chess’. They are homophonous and made up of two moras, as evident from the same words written in hiragana (another writing system in Japanese): かく ‘write’ and かく ‘a piece in the Japanese chess’. Each hiragana character represents one mora of sound. For this toy lexicon, the **C** matrix can be set up as below, using trimoras as the basic sublexical unit. Note that both of the words are assumed to share the exactly same phonetic realizations.

$$\mathbf{C} = \begin{array}{c} \\ \text{書く} \\ \text{角} \end{array} \begin{array}{c} \begin{array}{cc} \text{\#かく} & \text{かく\#} \end{array} \\ \left[\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array}\right] \end{array} \tag{43}$$

For simplicity, suppose that each of the two words has its own symbolic meaning with the meanings of 書く ‘write’ and 角 ‘a piece in the Japanese chess’ being <WRITE> and <CHESS> respectively. Then, the **S** matrix can be set up as below.

$$\mathbf{S} = \begin{array}{c} \\ \text{書く} \\ \text{角} \end{array} \begin{array}{c} \begin{array}{cc} \texttt{<WRITE>} & \texttt{<CHESS>} \end{array} \\ \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right] \end{array} \tag{44}$$

Based on the word forms and meanings defined this way, the association matrix in the comprehension side will be as follows. Note that this estimation is carried out in the endstate-learning method (see Section 2.1 for more detail).

$$\mathbf{F} = \begin{array}{c} \\ \text{\#かく} \\ \text{かく\#} \end{array} \begin{array}{c} \begin{array}{cc} \text{<WRITE>} & \text{<CHESS>} \end{array} \\ \left[\begin{array}{cc} 0.25 & 0.25 \\ 0.25 & 0.25 \end{array}\right] \end{array} \tag{45}$$

Since the two words in the toy lexicon are homophonous, namely made up of the same set of trimoras, none of the two trimoras can be associated with either of the two meanings. This complete ambiguity leads to the complete ambiguity in predicting word-meanings from the word forms, as seen in the $\hat{\mathbf{S}}$ matrix below.

$$\hat{\mathbf{S}} = \begin{array}{c} \\ \text{書く} \\ \text{角} \end{array} \begin{array}{c} \begin{array}{cc} \texttt{<WRITE>} & \texttt{<CHESS>} \end{array} \\ \left[\begin{array}{cc} 0.5 & 0.5 \\ 0.5 & 0.5 \end{array}\right] \end{array} \tag{46}$$

where both of the meanings are supported equally, based on the word form かく [kakɯ].

In contrast, homophony does not deteriorate the production process. Based on the same **C** and **S** matrices, the association weights on the production side (i.e., **G**) are estimated as follows.

$$\mathbf{G} = \begin{array}{r} \\ \text{<WRITE>} \\ \text{<CHESS>} \end{array} \begin{array}{c} \begin{array}{cc} \text{\#かく} & \text{かく\#} \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{array} \tag{47}$$

Both of the meanings are associated maximally to both of the trimoras. This is because the same form is to be predicted always whether the input meaning is <WRITE> or <CHESS>. Because of the complete associations between the meanings and the trimoras, the predictions of the model about the word forms are also perfect with no sign of deterioration due to homophony (Equation 48).

$$\hat{\mathbf{C}} = \begin{array}{r} \\ \text{書く} \\ \text{角} \end{array} \begin{array}{c} \begin{array}{cc} \text{\#かく} & \text{かく\#} \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{array} \tag{48}$$

The resistance of the production process against homophony is again due to the fact that the model only needs to produce the same form, regardless of the input meaning.

The example above involves one artifact because of simplification. The meanings of 書く 'write' and 角 'a piece in the Japanese chess' were defined with their own symbolic meanings, which were orthogonal (i.e., completely unrelated) to each other[6]. However, the assumption of orthogonality among word meanings is unrealistic. Some words can be similar to each other than others. For example, another Japanese word 記す [ʃiɾɯsɯ] 'write' shares the meaning of <WRITE> with 書く [kakɯ] 'write'. With this word added to the toy lexicon above, the **C** and **S** matrices will be updated as follows.

[6] Orthogonality between the two meanings in the example can be shown by calculating cosine similarity between the two meanings, for example, namely the row vectors in the **S** matrix (Equation 44). It will be

$$\mathrm{CosSim}(\mathbf{s}_{書く}, \mathbf{s}_{角}) = \frac{\mathbf{s}_{書く} \cdot \mathbf{s}_{角}}{\left\|\mathbf{s}_{書く}\right\| \left\|\mathbf{s}_{角}\right\|} = \frac{\begin{bmatrix}1 & 0\end{bmatrix} \cdot \begin{bmatrix}0 & 1\end{bmatrix}}{\left\|\begin{bmatrix}1 & 0\end{bmatrix}\right\| \left\|\begin{bmatrix}0 & 1\end{bmatrix}\right\|} = \frac{1 \cdot 0 + 0 \cdot 1}{\sqrt{1^2 + 0^2}\sqrt{0^2 + 1^2}} = \frac{0}{1} = 0$$

The cosine similarity of 0 represents no association between the two vectors.

$$\mathbf{C} = \begin{array}{c|ccccc|} & \text{\#かく} & \text{かく\#} & \text{\#しる} & \text{しるす} & \text{るす\#} \\ \text{書く} & 1 & 1 & 0 & 0 & 0 \\ \text{角} & 1 & 1 & 0 & 0 & 0 \\ \text{記す} & 0 & 0 & 1 & 1 & 1 \end{array} \tag{49}$$

$$\mathbf{S} = \begin{array}{c|cc|} & \text{<WRITE>} & \text{<CHESS>} \\ \text{書く} & 1 & 0 \\ \text{角} & 0 & 1 \\ \text{記す} & 1 & 0 \end{array} \tag{50}$$

Shared meanings such as <WRITE> in the example make production more difficult, while it does not deteriorate comprehension. It is the opposite case to the problem homophones cause for the comprehension process. In comprehension, the model needs to predict the meaning of <WRITE>, whether the input form is 書く or 記す, thus a simpler task for comprehension. Below are the **F** and $\hat{\mathbf{S}}$ matrices, based on the **C** and **S** matrices above.

$$\mathbf{F} = \begin{array}{c|cc|} & \text{<WRITE>} & \text{<CHESS>} \\ \text{\#かく} & 0.25 & 0.25 \\ \text{かく\#} & 0.25 & 0.25 \\ \text{\#しる} & 0.33 & 0.00 \\ \text{しるす} & 0.33 & 0.00 \\ \text{るす\#} & 0.33 & 0.00 \end{array} \tag{51}$$

$$\hat{\mathbf{S}} = \begin{array}{c|cc|} & \text{<WRITE>} & \text{<CHESS>} \\ \text{書く} & 0.5 & 0.5 \\ \text{角} & 0.5 & 0.5 \\ \text{記す} & 1.0 & 0.0 \end{array} \tag{52}$$

In the **F** matrix, the new trimoras are associated with the meaning of <WRITE> without ambiguity, leading to the perfect comprehension, as displayed in the third row of the $\hat{\mathbf{S}}$ matrix. The other cell values in the **F** and $\hat{\mathbf{S}}$ matrices are the same as those in the same matrices without the additional word 記す (i.e., Equations 45 and 46), indicating the addition of the new word 記す does not deteriorate the comprehension process.

For the production process, however, shared meanings pose a problem. Since the meaning of <WRITE> is shared by different word forms, the model cannot determine the correct word form only by the meaning of <WRITE>. Below are the **G** and $\hat{\mathbf{C}}$ matrices, based on the **C** and **S** matrices of the toy lexicon with 書く, 角, and 記す.

$$\mathbf{G} = \begin{array}{c} \\ \text{<WRITE>} \\ \text{<CHESS>} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{#かく} & \text{かく#} & \text{#しる} & \text{しるす} & \text{るす#} \end{array} \\ \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 1.0 & 1.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \end{array} \tag{53}$$

$$\hat{\mathbf{C}} = \begin{array}{c} \\ \text{書く} \\ \text{角} \\ \text{記す} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{#かく} & \text{かく#} & \text{#しる} & \text{しるす} & \text{るす#} \end{array} \\ \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 1.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix} \end{array} \tag{54}$$

As illustrated in the **G** matrix above, the association weights from the meaning of <WRITE> are split to the trimoras of かく (which is the pronunciation of 書く and 角 both) and those of しるす (i.e., 記す). This indeterminacy in the **G** matrix is reflected in indeterminacy in the $\hat{\mathbf{C}}$ matrix, where the word forms of 書く and 記す are not fully supported by their meanings due to ambiguity created originally by the meaning of <WRITE>.

In summary, in the framework of DLM, different meanings of homophones can lead to different ways of their realizations, being interacted with other words with similar meanings. Homophones would be predicted to be identical only when they share the exactly same meaning, which is not likely in reality. In fact, recent studies have reported word-specific phonetic realizations (Chuang et al. published online 11 May 2024, Lu et al. published online 28 August 2024). With this in mind, we can now turn to an explanation of semantic support. More specifically, different meanings of homophones should be reflected in different values in predicted form vectors (i.e., $\hat{\mathbf{C}}$).

2.4. Semantic support

2.4.1. Unconditional semantic support

Each value in $\hat{\mathbf{C}}$ can conceptually be understood as a measure of how much a certain sublexical form (e.g., a triphone) is supported by the meaning of a particular word. For example, in (36), the model is quite certain that `#kæ`, `kæt`, and `æt#` should be the components of the word form, given the meaning of *cat*.

Semantic support values can be considered for each sublexical form or for the entire word. For example, in the example above, the triphone `#kæ` has a semantic support value of 1.00 from the meaning of *cat*, as can be seen in Equation 36. Since *cat* is defined to be made of `#kæ`, `kæt`, and `æt#` and they all receive a semantic value of 1.00, the word *cat* receives a total semantic support value of 3.00. Formally, semantic support for a sublexical form and for the entire word can be defined below as Equation (55) and (56) respectively:

$$\text{SemSup}_{i,j} = \hat{\mathbf{C}}_{i,j} \tag{55}$$

$$\text{SemSup}_{i} = \sum_{k \in C} \text{SemSup}_{i,k} \tag{56}$$

where i is the index of the word, j is the index of the sublexical form, and C represents a set of cues (e.g., triphones) that constitutes the word i.

Semantic support, regardless whether it is for the entire word or for a particular sublexical form, has been found to be positively correlated with duration (Saito et al. 2023, Gahl & Baayen in press, Saito et al. under revision).

2.4.2. Conditional semantic support

Semantic support values are cell values of the predicted form matrix $\hat{\mathbf{C}}$. In the original and most basic setup of DLM, a word form vector is produced at once. In other words, all constituent sublexical forms are produced altogether, regardless of their positions in the word. Sublexical forms can be defined, so that the word-initial or word-final properties can be visible to the analyist, for example, by using # (e.g., `æt#`). These are simply labels for rows and columns of the matrices involved and invisible to DLM. The order of sublexical forms is only implicitly inferred from a set of sublexical forms at hand. For example, only one order is possible from the set of `#kæ`, `kæt`, and `æt#`.

This issue of ordering sublexical forms has been addressed so far by means of graph theory and the technique called positional learning (Baayen et al. 2018, Heitmeier et al. 2024). According to this solution, additional weight matrices are estimated for each position separately (e.g., a matrix for the word-initial position, another matrix for the second-in-word position, and so on). There need to be as many matrices as the number of ngrams in the longest word in the lexicon. If the longest word in the lexicon has 10 ngrams, there need to be 10 matrices. Each of these matrices represents how likely it is for each sublexical form to occur in each intra-word position. Based on these matrices, possible strings of sublexical forms are taken into account, which must pass a certain threshold. The threshold essentially helps to reduce the number of candidates. After constructing all the possible concatenations of sublexical forms, the most optimal path is determined with help of the graph theory. Listing up of all the possible candidates and selection of the most likely candidate with the graph theory need to be executed for production of each word.

Although this solution is very effective (Heitmeier et al. 2024), it is not completely free from shortcomings. First of all, it may not be very realistic from the cognitive perspective that the length of the longest word in the lexicon needs to be known prior to learning the positional matrices. Secondly, it may be too restrictive to assume that all candidate forms must be listed first before one of them is eventually selected. Such a solution would lead to better accuracy and might therefore be preferred from an engineering perspective. However, it may not be very realistic to assume that the listener lists up in their head all the words such as *ant*, *anterior*, *antenna*, *anarchy*, and so on, upon hearing the word-initial [æn-]. These candidates can be reduced by setting a hyper-parameter of a threshold. However, it is not clear how cognitively realistic to assume such a hyper-parameter, although use of hyper-parameters would not be a problem from the engineering perspective.

The issue of ordering sublexical forms can be solved without resort to graph theory or positional learning. The current study proposes one way of solving this issue by considering the interaction between the speaker and the listener. The speaker does not produce word forms for themselves. They usually speak to someone else. The speaker speaks louder in a noisy environment to make their speech more audible to the listener. The speaker may repeat what they have said again if they think the listener could not hear what they had said. These facts suggest that the speaker modifies their speech in accordance with the listener's understanding, or more precisely, in accordance with what the speaker thinks the listener has understood so far. This process of self-monitoring by the speaker can be integrated in the process of producing a word form in DLM in the following manner:

$$\hat{\mathbf{c}}_{i,t} = \mathbf{s}_t \cdot \mathbf{G}_t \tag{57}$$

Similarly to the original production process of DLM, which was shown in Equation (34), the modified production process (57) predicts a word form based on a mapping between a semantic vector and the association matrix **G**. The only modification in (57) is that each term has its own state at time t. $\hat{\mathbf{c}}_{i,t}$ represents which sublexical forms are activated at time t.

$\mathbf{s}_t$ represents a semantic vector of the target word with the speaker's assumption about the listener's understanding taken into account. It is defined as a difference vector between the target semantics $\mathbf{s}_i$ and the speaker's assumption about the listener's understanding so far, namely $\hat{\mathbf{s}}_{i,t-1}$, as in Equation (58). $\mathbf{s}_t$ conceptually represents what the speaker thinks they should say in order to make the listener understand what they want them to understand.

$$\mathbf{s}_t = (\mathbf{s}_i - \hat{\mathbf{s}}_{i,t-1}) \tag{58}$$

$\hat{\mathbf{s}}_{i,t-1}$ represents the speaker's model of the listener's understanding, based on what the speaker has said. It is only a model of the speaker about the listener's understanding, because the speaker cannot know exactly what the listener really heard and understood so far. For ease of exposition, we assume that the speaker assumes that the listener has perceived all the sublexical forms the speaker has provided so far correctly. Mathematically, this can be expressed as a form vector that contains all the sublexical forms that the speaker has produced so far, with 1 in the cell values of these sublexical forms and 0 for the other sublexical forms, namely $\mathbf{c}_{i,t-1}$, multiplied by the comprehension weight matrix $\mathbf{F}$, as below:

$$\hat{\mathbf{s}}_{i,t-1} = \mathbf{c}_{i,t-1}\mathbf{F} \tag{59}$$

Equation (57) also contains the production weight matrix $\mathbf{G}$ at time t. $\mathbf{G}_t$ is essentially the same as the usual association matrix $\mathbf{G}$, but it is modified, so that irrelevant sublexical forms are not produced in conjunction with the previous sublexical form. This modification reflects physical and physiological restrictions. Regardless of the size of sublexical forms, each one is expected to have certain co-articulatory characteristics, which serve to concatenate the sublexical forms implicitly (see Appendix B.1 for more details). For example, if the current sublexical form is #æn, then the next sublexical form has to have æn in the beginning (e.g., ænt). It is because #æn already has some coarticulatory characteristics of the upcoming [n] (e.g., the tongue tip rising toward the offset of [æ]). Choosing any other triphone that does not contain æn in the beginning would represent a physiologically impossible situation where articulatory characteristics executed and prepared for certain segments have vanished suddenly. In other words, the speaker only needs to think about the next possible tongue positions, which are physiologically possible. In order to reflect this reasoning, irrelevant cell values are turned off to be 0 in the association matrix $\mathbf{G}$, which is expressed in Equation (60), where $\mathrm{Diag}(\mathbf{v}_{t-1})$ represents this process of 'turning-off' unnecessary cell values.

$$\mathbf{G}_t = \mathbf{G} \cdot \mathrm{Diag}(\mathbf{v}_{t-1}) \tag{60}$$

Diag is intended to be an operator that converts a vector into a diagonal matrix, whose diagonal values correspond to the input vector. This operation by Diag can also expressed as in Equation (61). In Equation (61), $\mathbf{1}$ is all-one vector of the same size as $\mathbf{v}_{t-1}$, $\odot$ represents element-wise multiplication, and $\mathbf{I}$ is an identity matrix of the same shape as $(\mathbf{v}_{t-1}^{\top} \cdot \mathbf{1})$.

$$\mathrm{Diag}(\mathbf{v}_{t-1}) = (\mathbf{v}_{t-1}^{\top} \cdot \mathbf{1}) \odot \mathbf{I} \tag{61}$$

$\mathbf{v}_{t-1}$ is a vector that has 1 only for the sublexical forms that are physiologically possible given the last sublexical form. It is a row vector of an additional matrix $\mathbf{V}$, which lists all possible sequences of sublexical forms (Equation (62)).

$$\mathbf{V} = \begin{array}{c} \\ \texttt{\#kæ} \\ \texttt{\#ræ} \\ \texttt{\#hæ} \\ \texttt{kæt} \\ \texttt{ɹæt} \\ \texttt{hæt} \\ \texttt{æt\#} \\ \phi \end{array} \begin{array}{c} \begin{array}{ccccccc} \texttt{\#kæ} & \texttt{\#ræ} & \texttt{\#hæ} & \texttt{kæt} & \texttt{ɹæt} & \texttt{hæt} & \texttt{æt\#} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{62}$$

The rows of **V** are the current sublexical forms, and the columns represents the possible next sublexical forms. So, for example, the first entry in the first column of V, #kæ, can only be continued by kæt, which is why only kæt has a 1 in the row of #kæ, and all other values in that row are at 0. The last row of **V**, i.e. ϕ, represents the word-initial position.

Based on this algorithm of form-production (i.e., Equation (57)), a predicted form vector can be produced at each time step. Each predicted form vector will represent which sublexical forms are supported to what degree by the meaning of the target word and by what has been produced by the speaker so far. We call these values in predicted form vectors produced with this algorithm of the incremental production CONDITIONAL SEMANTIC SUPPORT, because they are semantic support values in the sense that they are values in predicted form vectors, but they are conditional on context (i.e., what has been produced so far).

As can be understood from Equations (57) and (58), conditional semantic support values are determined by the certainty of sublexical forms to represent the meaning of the target word and by syntagmatic predictability. The former is a major property in DLM (in the production side). When certain meanings occur unambiguously with certain cues (i.e. sublexical forms), the certainty of these cues for expressing the meanings increases. Certainty of sublexical forms is reflected in the associations in the **G** matrix. Higher certainty among sublexical forms will appear as higher values in the **G** matrix, and those high values in the **G** will be mapped onto high values in predicted form vectors. Since the **G** matrix is part of the definition of conditional semantic support, certainty of cues is integrated and captured in conditional semantic support. This certainty is also captured and represented also by UNCONDITIONAL SEMANTIC SUPPORT, which, however, does not reflect the conditional certainty that is captured by the conditional semantic support.

On the other hand, syntagmatic predictability is only captured by conditional semantic support and not by unconditional semantic support. When the sublexical forms early in the word already convey most of the meaning of the entire word, the rest of the sublexical forms later in the word do not have to be pronounced clearly. This can be the case, especially for a long word such as *encyclopedia*, for example.

This kind of syntagmatic predictability effect occurs for conditional semantic support because the predicted form vector at a certain time is produced based on DIFFERENCES between the target semantic vector (i.e. $\mathbf{s}_i$) and the semantic vector of the listener's understanding (i.e., $\hat{\mathbf{s}}_{i,t-1}$) (see Equation 58). When sublexical forms early in the word already convey most of the meaning of the target word, the listener's understanding, namely $\hat{\mathbf{s}}_{i,t-1}$, should already be close enough to the target meaning (i.e., $\mathbf{s}_i$). In other words, the differences between $\hat{\mathbf{s}}_{i,t-1}$ and $\mathbf{s}_i$, namely $\mathbf{s}_t$, should already be small. $\mathbf{s}_t$ with smaller values will, in turn, produce smaller values in $\hat{\mathbf{c}}_{i,t}$ through its mapping with $\mathbf{G}_t$ (i.e., Equation 57).

In summary, conditional semantic support values are determined between the two opposing forces. They are higher, when predicted sublexical forms are well supported by semantics. They become smaller, when certain sublexical forms (later in the word) are not so important any more, in the sense that the meanings they can convey are already conveyed other sublexical forms earlier in the word. In other words, conditional semantic support values for certain sublexical forms get smaller when those sublexical forms are predictable from other sublexical forms earlier in the word.

This algorithm of the incremental production is explained in more details with a concrete example in Appendix B.4.

3. METHODS

3.1. *Data*

The current study investigates phonetic realizations of Japanese homophones at the mora level as well as the word level. For this purpose, durations of homophonous words were collected from the core section of Corpus of Spontaneous Japanese (CSJ). CSJ contains a total of about 661.6 hours of spontaneous and read-aloud speech in Japanese, including a portion of the entire dataset that amounts to about 44 hours, for which annotations for a series of different linguistic units were corrected by hand. In CSJ, annotations were generated first with forced-alignment and subsequently checked manually by two phoneticians to ensure validity of the annotations (The National Institute for Japanese Language 2006). For its reliability of annotations, this section of CSJ was adopted for our investigation. The dataset contained about 500,000 words from approximately 44 hours of recording of speech in Japanese, which consisted of formal monologues of spontaneous speech by 177 speakers, formal dialogues of spontaneous speech by 18 speakers, and read-aloud speech of books by 6 speakers.

In Japanese, there are many homophones, compared to English or other languages. It is not uncommon that a single pronunciation is shared by more than two homophonous words. For example, the pronunciation こうしょう [koːʃoː] is shared by at least 54 words[7]. This is partially due to the relatively small inventory of

[7] They are 交床, 交渉, 交睫, 交鈔, 厚相, 厚賞, 公傷, 公娼, 公相, 公称, 公証, 咬傷, 口承, 口

phonemes and its simple phonotactics. These homophonous words have different meanings, such as 書く [kakɯ] 'write' and 掻く [kakɯ] 'scratch'.

In the current dataset, 310,574 word tokens were involved in one of the homophonous word types. In type counts, 20,971 word types were identified to be involved in homophonous word types. Homophonous words were identified on the basis of their phonetic transcriptions. Consequently, different tokens from the same word type with different pitch-drop positions were not counted as homophones, because they are not exactly homophones. For example, for the word 昨日 'yesterday', two different pitch contours were observed, which were [kino$^{\downarrow}$ː] and [kinoː], where '$^{\downarrow}$' indicates a drop in pitch.

However, these numbers could be inflated by function words that were made of only one mora. For example, the word に [ni] is extremely flexible. It can be a locative particle similar in meanings to *to*, a word for *two*, another word for *alike*, and many more. Those counts might also be inflated by the fact that pronunciations are represented phonetically, which would separate word tokens with phonetically different realizations. By limiting to nouns and phonological representations, where the above-mentioned issues are expected to be minimized, the current dataset has 108,076 noun tokens and 8,205 noun types. Of these, approximately 8% of the noun types (653 noun types) were involved in at least one other homophonous noun, and about 35% of the noun tokens (38,075 noun tokens) were found to be involved in at least one other homophonous noun.

3.2. Estimation of DLM matrices

DLM requires a matrix for word forms and another matrix for word-meanings. They are conventionally called **C** and **S**. For **C**, we encoded word forms in terms of sequence of three moras. We used sequences of three moras instead of triphones, because moras are the minimal phonological unit in Japanese that matches the intuition of the native speakers (Port et al. 1987, Cutler & Otake 1994, Han 1994, Kubozono 2017).

Japanese makes use of three writing systems. One of them is Chinese characters and called Kanji. Kanji is an ideographic system, in which each character represents a meaning, rather than sounds. The other two are phonographic systems, in which each character represents the equivalent of one or two IPA sounds. They are called Hiragana and Katakana, and each character in Hiragana and Katakana represents a mora in Japanese. Most content words can be written in any of, or in mixture of, the three writing systems[8]. Function words are usually written in Hiragana, and

誦, 哄笑, 好尚, 幸勝, 公勝, 工匠, 工商, 工廠, 巧匠, 巧笑, 康正, 康尚, 後章, 後証, 校章, 洪鐘, 甲匠, 紅晶, 綱掌, 翺翔, 考証, 行省, 行粧, 行障, 行賞, 鉱床, 講頌, 講誦, 降将, 高小, 高升, 高声, 高姓, 高尚, 高昇, 高承, 高昌, 高商, 高唱, 高蹤, and 黄鐘.

[8] Most content words are written in Kanji, loanwords and scientific names are written in katakana, and grammatical functions are written in hiragana. But authors have a great deal of freedom to use whatever writing system they want, most likely depending on a nuance in the meaning they want to convey. As far as we know there has been no research on this interesting topic.

they cannot be written in Kanji.

The CSJ corpus provides transcriptions in the usual transcription with three writing systems mixed and also one only in Katakata. The latter is provided in order to transcribe pronunciations in a clearer manner. Since each character in Katakana represents a mora, sequences of three moras of each word were encoded using the transcription in Katakana. For example, the Japanese word for ‘language’ is written as 言語 ([geŋgo]) with Chinese characters. Its pronunciation is ゲンゴ [ge ŋ go], where each character represents one mora (in the IPA the sounds of each katakana is separated by a space). Using the transcription in Katakana in the CSJ corpus, the word was defined with the following trimoras: #ゲン, ゲンゴ, and ンゴ#, where # indicates the word boundary.

For the semantic matrix **S**, a pre-trained model of fastText (Bojanowski et al. 2017) for Japanese was adopted (Grave et al. 2018)[9], which was trained on Japanese Wikipedia pages. Each semantic vector had 300 semantic dimensions.

These semantic vectors were based on orthographic representations. In Japanese, the same word written in Chinese characters (i.e., Kanji) can have different pronunciations. For example, the Japanese word for ‘Japan’ can be written as 日本 ([nihoɴ]) with Chinese characters. There are two readings. It can be either にほん [nihoɴ] or にっぽん [nip˺poɴ]. Different pronunciations were distinguished in the CSJ corpus. Consequently, different pronunciations sharing the same orthography were assigned with the same semantic vector from the pre-trained fastText model (Grave et al. 2018).

Based on these form and semantic matrices, the weight matrices **F** and **G** were estimated, using end-state of learning, which analytically estimates an equilibrium state of the association weights.

For the setup of these matrices, all words with a frequency greater than 1 in the CSJ dataset were included. Some words found in the CSJ dataset were not a part of the words the pre-trained fastText model was trained on. We excluded these. In addition, all the words with only one mora were excluded. It was because most of the one-mora words were function words such as case particles (e.g., に [ni] ‘to’) and also because one-mora words had to be defined in terms of only one sequence of three moras, which made it difficult to distinguish the mora-level phenomena from the word-level (i.e., lexical) phenomena (e.g., に [ni] ‘to’ would be defined only by #ニ#, in which ニ is the katakana of the hiragana に, both pronounced as [ni].

After these data exclusion procedures, 99,776 word tokens (data points) remained available for the word-level analysis, consisting of 1,586 word types in orthography and 1,200 word types in phonetic transcriptions. For the mora-level analysis, 213,399 data points (i.e., moras) of 1,586 word types in orthography and 1,200 phonetic word types in phonetic transcriptions were available.

[9] The pre-trained model was downloaded from the following website: https://fasttext.cc/docs/en/crawl-vectors.html

3.3. Analysis

To test the non-linear effects of semantic support measures with other covariates and factors, we used Generative Additive Mixed-effects Models (GAMMs) (Wood 2017). They were fitted with word duration (i.e., `WordDur`) and mora duration (i.e., `MoraDur`) as the dependent variables for word-level analysis and for mora-level analysis respectively. `WordDur` and `MoraDur` were distributed skewed and therefore logarithmically transformed prior to analysis in order to approximate the normal distribution for each of the two dependent variables.

Semantic support measures for each mora and its sum for each word are of the most interest in the current study. Semantic support for each sublexical forms (i.e., bimoras) can be calculated with and without preceding contexts, namely preceding sublexical forms. We call semantic support with context taken into account conditional semantic support, and we call semantic support without context taken into account unconditional semantic support. Unconditional semantic support is a cell value in a predicted form matrix.

Conceptually, semantic support represents how well a certain sublexical form is unambiguously supported from the meaning of the word that contains the certain sublexical form, whether it is unconditional or conditional semantic support. Unconditional semantic support does not take into account what sublexical forms precede the sublexical form of interest. Conditional semantic support does take into account how predictable the sublexical form of interest is from the other sublexical forms preceding it. See Section 2.4.2 for more details. All the four semantic support measures, namely uSemSup, cSemSup, uSemSupWord, and cSemSupWord, showed a skewed distribution and were therefore log-transformed in prior to analysis.

Speech rate affects duration. Faster speech rates are often correlated with shorter durations (Kuehn & Moll 1976, Kelso et al. 1985, Gahl et al. 2012, Cohen Priva 2015, Malisz et al. 2018). In the current study, speech rate was included in the analysis (i.e., `SpRate`), as defined as the number of moras in an utterance divided by the duration of the utterance. Regarding the definition of utterances in the current study, we made use of the INTER-PAUSAL UNIT (IPU) available in the CSJ corpus. Inter-pausal units are a continuous stretch of speech bound by at least 0.2 seconds. Based on this definition of utterances, the words at the utterance-initial and utterance-final positions were marked (i.e., `UttInitial` and `UttFinal`).

In addition, word frequency was calculated, based on the phonetic transcriptions available in the CSJ corpus, to control baseline (prior) probabilities of words (i.e., `Freq`). `Freq` was log-transformed prior to analysis. The current study aimed to investigate the mora durations as well as word durations. In order to take into account baseline (prior) probabilities of sublexical forms, bimora probabilities were also calculated, and the sum of the bimora probabilities for each word was included in the analysis (i.e., `BimoraFreq`). `BimoraFreq` was correlated with lengths of words, because `BimoraFreq` was the sum of the probabilities of the bimoras that make up the word. To normalize word length, `BimoraFreq` was divided by the word length, before being log-transformed and included in the analysis.

Different syntactic classes have been found to show systematically different durations (Lohmann 2018a). In order to account for these systematic differences among syntactic classes, parts-of-speech were included as an additional factor variable. The reference level was adjective.

To also take into account speaker differences, gender was included as a factor variable (i.e., `Gender`) . In addition, speakers were included as a random effect, as well (i.e., `Speaker`). The reference level of `Gender` is female. Birth place and year were also considered. However, their effects were not significant and were therefore excluded from the analysis.

With the set of variables introduced above, four GAM models were constructed for each combination of the two dependent variables (i.e., either `WordDur` or `MoraDur`) and the two types of semantic support measures. These four models had the same model structure, except for their dependent variables and semantic support measures. The structures of these four models are illustrated below, following the syntax adopted by the `mgcv` package (Wood 2017) in `R` (R Core Team 2022):

Model 1: WordDur ~ s(uSemSupWord, k=3) + Covariates
Model 2: WordDur ~ s(cSemSupWord, k=3) + Covariates
Model 3: MoraDur ~ s(uSemSupWord, k=3) + Covariates
Model 4: MoraDur ~ s(cSemSupWord, k=3) + Covariates
Covariates: s(SpRate, k=3) + s(Freq, k=3) + s(BimoraFreq, k=3)
+ UttBgn + UttEnd + PoS + Gender + s(Speaker, bs='re')

4. Results

4.1. *The word-level analysis*

Word durations were significantly different among the members of each homophonous pair ($V = 2241903, p < 0.001$). The median difference was 0.039 seconds. In addition, unconditional and conditional semantic support values were also found to be significantly different among the members of homophonous pairs ($V = 2241903, p < 0.001$).

The GAM model with unconditional semantic support for word duration (i.e. Model 1) outperformed the model with conditional semantic support (i.e., Model 2) in AIC (ΔAIC = 1079.090). Therefore, the results based on Model 1 will be reported for word duration below.

Unconditional semantic support was estimated to have a positive relationship with word duration (Figure 1). The relationship was estimated to be almost linear. The estimates of the model are summarized in Table 2.

Faster speech rate was associated with shorter duration (Figure 2). Word frequency was correlated with shorter duration, but bimora frequency was correlated with longer duration. However, word frequency and bimora frequency were correlated to each other ($r = 0.695, p < 0.001$), and therefore, their estimated effects will not be interpreted or discussed.

Table 2. The summary of the model with unconditional semantic support (Model 1) for the word duration data.

(A. Parametric)	β	SE	t	p
(Intercept)	-1.366	0.011	-125.593	<0.001
UttBgn=TRUE	-0.045	0.004	-10.345	<0.001
UttEnd=TRUE	0.410	0.004	105.348	<0.001
POS=adnominal	-0.272	0.009	-28.750	<0.001
POS=adverb	0.044	0.009	4.930	<0.001
POS=auxverb	0.041	0.008	5.103	<0.001
POS=conjunction	0.088	0.016	5.643	<0.001
POS=determiner	-0.155	0.041	-3.758	<0.001
POS=english	-0.209	0.040	-5.249	<0.001
POS=interjection	-0.376	0.028	-13.632	<0.001
POS=noun	0.083	0.008	10.865	<0.001
POS=particle	0.141	0.009	15.643	<0.001
POS=prefix	-0.112	0.013	-8.837	<0.001
POS=pronoun	0.025	0.010	2.606	0.009
POS=suffix	-0.141	0.008	-16.703	<0.001
POS=verb	-0.226	0.008	-29.980	<0.001
Gender=male	-0.041	0.010	-4.209	<0.001
(B. Smooth)	edf	Ref.df	F	p
s(SpRate)	1.814	1.965	5888.294	<0.001
s(Freq)	1.906	1.991	2067.443	<0.001
s(BimoraFreq)	1.995	2.000	261.796	<0.001
s(Speaker)	125.726	135.000	33.584	<0.001
s(uSemSupWord)	1.001	1.001	1342.198	<0.001

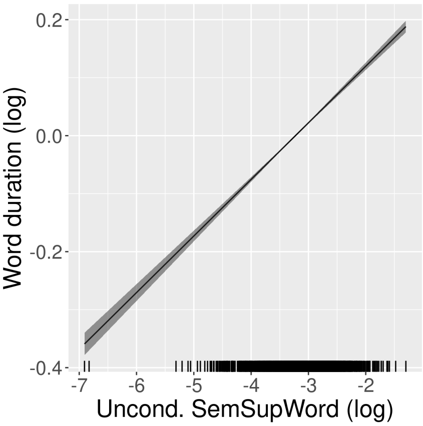


Figure 1. Estimated partial effects of unconditional semantic support for word on word duration.

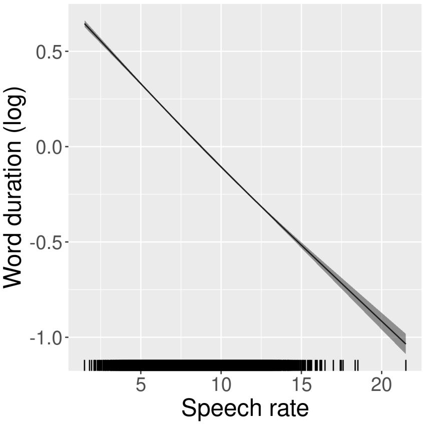


Figure 2. Estimated partial effects of speech rate on word duration.

4.2. The mora-level analysis

Mora durations were also found to be significantly different among the members of each group of homophonous words with the median difference 0.018 seconds ($V = 2241903, p < 0.001$). These differences were echoed by unconditional and conditional semantic support as well with the median difference being 0.017 ($V = 2239786, p < 0.001$).

The mora duration model with conditional semantic support (i.e., Model 4) outperformed the model with unconditional semantic support (i.e., Model 3) in

AIC (ΔAIC = 203.691). Therefore, the model with conditional semantic support will be reported below for mora duration. Estimates for the model with conditional semantic support are summarized in Table 3

Table 3. The summary of the model with conditional semantic support (Model 4) for the mora duration data.

(A. Parametric)	*β*	SE	*t*	*p*
(Intercept)	-2.285	0.010	-220.894	<0.001
UttBgn=TRUE	-0.056	0.004	-13.621	<0.001
UttEnd=TRUE	0.409	0.004	105.422	<0.001
POS=adnominal	-0.189	0.009	-20.573	<0.001
POS=adverb	0.058	0.009	6.793	<0.001
POS=auxverb	0.172	0.008	22.255	<0.001
POS=conjunction	0.248	0.016	15.032	<0.001
POS=determiner	-0.087	0.041	-2.133	0.033
POS=english	-0.063	0.040	-1.566	0.117
POS=interjection	-0.130	0.028	-4.685	<0.001
POS=noun	0.103	0.007	14.202	<0.001
POS=particle	0.221	0.009	24.947	<0.001
POS=prefix	0.108	0.013	8.608	<0.001
POS=pronoun	0.022	0.009	2.387	0.017
POS=suffix	0.044	0.008	5.446	<0.001
POS=verb	-0.192	0.007	-26.464	<0.001
Gender=male	-0.036	0.009	-4.064	<0.001
(B. Smooth)	**edf**	**Ref.df**	***F***	***p***
s(SpRate)	1.899	1.989	6835.057	<0.001
s(Freq)	1.999	2.000	398.516	<0.001
s(BimoraFreq)	1.999	2.000	2180.071	<0.001
s(Speaker)	123.952	135.000	142.704	<0.001
s(cSemSupMora)	1.987	2.000	660.770	<0.001

Effects of conditional semantic support on mora duration were estimated to be non-linear with a clear positive relationship with mora duration for where most data points are concentrated. The distribution of the data points is illustrated in Figure 3 as small vertical black lines at the bottom of the figure. The non-linearity of the estimated effects was mainly due to the sparseness of the data for smaller values of conditional semantic support for mora, for which there are fewer data points. The data sparsity made the estimation unreliable, indicated by wider confidence intervals.

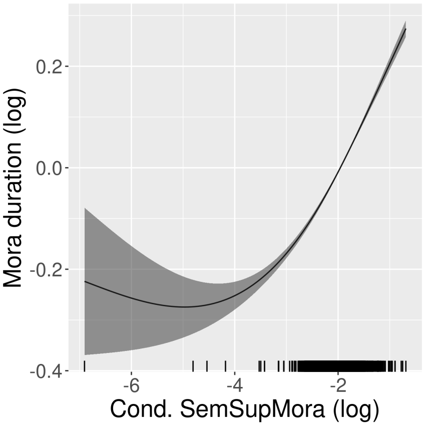


Figure 3. Estimated partial effects of conditional semantic support for mora on mora duration.

Speech rate had a negative correlation with mora duration. Faster speech rates were associated with shorter mora duration (Figure 4).

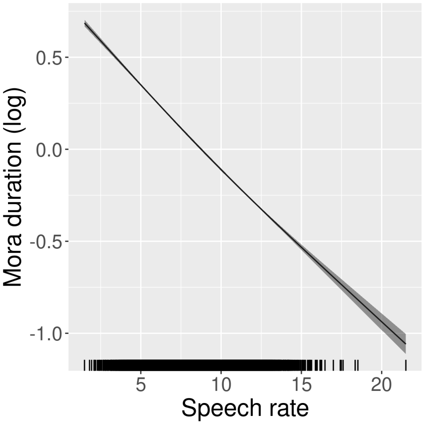


Figure 4. Estimated partial effects of speech rate on mora duration.

For low to mid word frequency values mora duration decreased. From mid frequency to high frequency mora duration increased. The opposite patterns were observed for bimora frequency. Mora duration was longest for the mid bimora frequency and the shortest for high bimora frequency. However, these two variables were correlated with each other ($r = 0.648, p < 0.001$), and therefore they will not be interpreted any further.

5. Discussion

We investigated the effects of semantics on the duration of homophones in Japanese, a language where duration is phonemic (Aim 1). In addition, we investigated mora duration as well as word duration in order to investigate whether semantic effects were tied to the lexicality of words (Aim 2). Systematic differences in word duration and mora duration as an effect of semantics are predicted by DLM (Baayen et al. 2019) while no such effects are predicted by traditional modular-based feed-forward speech production models (e.g., Levelt et al. 1999), as summarized in Table 1.

Based on a spontaneous speech corpus of Japanese (CSJ), homophone durations were found to be systematically longer when they were supported better by the meaning of the word. This observation indicates a clear relationship between sounds and meanings. Homophones systematically differ in duration as a function of semantics even in a language in which vowel duration is phonemic (see Aim 1 in Section 1). Furthermore, durational differences due to semantics are likely to occur at the sublexical level, indicated by a significant relationship found between mora duration and conditional semantic support in the current study (see Aim 2 in Section 1). With respect to the hypotheses (Table 1), our findings–duration effects at the word and mora level–are explained by DLM (Baayen et al. 2019).

In the current study, word duration was better predicted by the measure of semantic support without contextual information (i.e., unconditional semantic support), while mora duration was better predicted by the measure of semantic support with contextual information (i.e., conditional semantic support). These distinct effects of unconditional and conditional semantic support suggest that these two measures capture different aspects of durational realizations in Japanese. Differences between conditional and unconditional semantic support usually appear more clearly toward the end of a word. This is because more context has become available toward the end of the word. With contextual information being accumulated toward the end of a word, the sublexical forms toward the end of the word become more predictable, thus smaller conditional semantic support values. In contrast, unconditional semantic support values do not necessarily decrease from the beginning to the end of a word. Better performance of conditional semantic support for predicting mora duration suggests that this decreasing trend in duration from the beginning to the end of each word was captured well by conditional semantic support. If word duration is only the sum of mora duration, conditional semantic support should win over unconditional semantic support to predict word duration as well. The better performance of unconditional semantic support for word duration observed in the current study, therefore, suggests that each word has its own target duration. Word-specific durational targets have, in fact, been suggested by the literature regarding the mora-timing system in Japanese (Port et al. 1987, Han 1994). According to the literature, segment duration can be stretched or compressed within and across moras to achieve certain word-durational targets in Japanese. In the current study, a post-hoc simulation confirmed better performance of unconditional and conditional semantic support for word and mora duration respectively when each word has its own durational target in addition to decreasing mora durations (Appendix C).

Regardless of the choice of word- and mora-durations, the relationship between duration and semantic support was consistently positive. Semantic support reflects (un)certainty among forms and meanings (see section 2.4 for more detail) with greater semantic support associated with higher certainty in forms based on semantics. Accordingly, the positive relationship found in the current study between semantic support and duration suggests a positive relationship between certainty and duration. When the speaker is more certain about the pronunciation, it results in a more careful and precise pronunciation (Kuperman et al. 2007, Cohen 2014, Tomaschek et al. 2019, Tucker et al. published online 20 March 2019, Tomaschek et al. 2021).

In addition, the current observations are in line with, and therefore add to, a growing number of recent studies that have documented systematic relations between sounds and meanings (Baayen et al. 2019, Chuang et al. 2020, Gahl & Baayen in press, Saito et al. under revision). The direct relationship between sounds and meanings also dovetails well with the literature on sound symbolism and iconicity (Dingemanse et al. 2016, Dingemanse & Thompson 2020). Non-arbitrary relationships between sounds and meanings were reported as early as in the late 1920s both in linguistics and psychology (Fischer 1922), and supporting evidence was repeatedly observed since then across languages (Ćwiek et al. 2022). Nevertheless, phenomena such as sound symbolism have only recently started being integrated into linguistic theory, due to the deeply entrenched assumption about the arbitrariness of the relationship between sound and meaning (de Saussure 1916).

In DLM, certain sublexical forms can be more strongly supported by semantics, when certain sublexical forms occur more unambiguously with certain meanings. For example, the word-initial *prz* (e.g., *Przwalskihorse*) does not occur in so many words in English, and all of the English words with the word-initial *prz* are associated with the meaning of <PRZWALSKIHORSE>. Consequently, DLM predicts, the constant co-occurrence of *prz* and <PRZWALSKIHORSE> makes the sublexical word form *prz* strongly supported by the meaning of <PRZWALSKIHORSE>. The degree of semantic support is continuous, as it is real-valued, not binary. It implies that a certain sublexical form can be associated with a certain meaning only partially. DLM allows for an intermediate association between sounds and meanings, which might be called phonaesthemes. The word-initial *gl-* often appears in the words with the meanings related to light (e.g., *glow*, *gleam*, *glisten*), but not necessarily (e.g., *glitch*). DLM predicts a stronger relation between *gl-* and the meaning of light if they occur together a substantial amount of times, allowing for the possibility that the same sublexical form is used as a part of the words that do not mean anything about light. Contribution of each sublexical form to a certain meaning is always partial.

In summary, the current study indicates that phonetic realizations can be determined by semantics at least partially (Baayen et al. 2019, Chuang et al. 2021, published online 11 May 2024, Lu et al. published online 28 August 2024), challenging traditional views of speech production that limit such effects of semantics on phonetic realizations (e.g., Levelt et al. 1999). Especially, degrees of certainty in forms based on meanings, which we quantified as semantic support, were found in the current study to lead to clearer and more careful speech (Kuperman et al. 2007, Cohen 2014, Tomaschek et al. 2019, Tucker et al. published online 20 March 2019, Tomaschek et al. 2021). By providing evidence for the continuous nature of form-semantic relations, the current study opens up quite a few possibilities to investigate relationships between sounds and forms such as sound symbolism, iconicity, and phonaesthemes (Dingemanse et al. 2016, Dingemanse & Thompson 2020). Once accepting partial contributions of meanings, intermediate components between sounds and meanings may not be absolutely necessary.

Competing interests

The authors declare none.

Supplementary materials

The data and scripts of the present study can be found at https://osf.io/mcr9p/.

References

Baayen, R. H., Yu-Ying Chuang & James P. Blevins. 2018. Inflectional morphology with linear mappings. *The Mental Lexicon* 13(2). 230–268.

Baayen, R. H., P. Milin, D. Filipović Durdević, P. Hendrix & M. Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3). 438–481.

Baayen, R. H., Richard Piepenbrock & Leon Gulikers. 1996. Celex2. *Linguistic Data Consortium, Philadelphia* .

Baayen, R Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity* 2019.

Ben Hedia, Sonia & Ingo Plag. 2017. Gemination and degemination in english prefixation: Phonetic evidence for morphological organization. *Journal of Phonetics* 62. 34–49.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5. 135–146.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan & H. Lin (eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, vol. 33, 1877–1901. Red Hook, NY, USA: Curran Associates, Inc.

Cheng, Hei Yan, Bruce E. Murdoch, Justin V. Goozée & Dion Scott. 2007. Electropalatographic assessment of tongue-to-palate contact patterns and variability in children, adolescents, and adults. *Journal of Speech, Language, and Hearing Research* 50(2). 375–392.

Chuang, Yu-Ying, Melanie J. Bell, Yu-Hsiang Tseng & R. Harald Baayen. published online 11 May 2024. Word-specific tonal realizations in Mandarin. Published online on *arXiv*, 11 May 2024. https://arxiv.org/abs/2405.07006.

Chuang, Yu-Ying, Marie Lenka Vollmer, Elnaz Shafaei-Bajestan, Susanne Gahl, Peter Hendrix & R. Harald Baayen. 2020. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods* 1–32.

Chuang, Yu-Ying, Marie Lenka Vollmer, Elnaz Shafaei-Bajestan, Susanne Gahl, Peter Hendrix & R. Harald Baayen. 2021. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods* 53. 945–976. doi:10.3758/s13428-020-01356-w.

Cohen, Clara. 2014. Probabilistic reduction and probabilistic enhancement. *Morphology* 24(4). 291–323.

Cohen Priva, Uriel. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6(2). 243–278. doi:10.1515/lp-2015-0008.

Cutler, Anne & Takashi Otake. 1994. Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language* 33(6). 824–844. doi:https://doi.org/10.1006/jmla.1994.1039. https://www.sciencedirect.com/science/article/pii/S0749596X84710394.

Ćwiek, Aleksandra, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus et al. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B* 377(1841). 20200390.

Davidson, Lisa. 2005. Addressing phonological questions with ultrasound. *Clinical Linguistics and Phonetics* 19(6-7). 619–633.

Dell, Gary S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93(3). 283–321. doi:10.1037//0033-295x.93.3.283.

Dell, Gary S., Nadine Martin & Myrna F. Schwartz. 2007. A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. *Journal of Memory and Language* 56(4). 490–520. doi:10.1016/j.jml.2006.05.007.

Dell, Gary S., Myrna F. Schwartz, Nadine Martin, Eleanor M. Saffran & Deborah A. Gagnon. 1997. Lexical access in aphasic and nonaphasic speakers. *Psychological Review* 104(4). 801–838.

Dingemanse, Mark, Will Schuerman, Eva Reinisch, Sylvia Tufvesson & Holger Mitterer. 2016. What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language* 92(2). e117–e133. doi:10.1353/lan.2016.0034.

Dingemanse, Mark & Bill Thompson. 2020. Playful iconicity: structural markedness underlies the relation between funniness and iconicity. *Language and Cognition* 12(1). 203–224. doi:10.1017/langcog.2019.49.

Fischer, Siegfried. 1922. Über das Entstehen und Verstehen von Namen. *Archiv für deutsche Gestalt Psychologie* 42. 335–368.

Gahl, Susanne. 2008. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84(3). 474–496.

Gahl, Susanne & R. Harald Baayen. in press. Time and thyme again: Connecting spoken word duration to models of the mental lexicon. *Language*. Published online on the website of the institution of the second author, 2 July 2024. https://quantling.org/~hbaayen/publications/GahlBaayen2024.pdf.

Gahl, Susanne, Yao Yao & Keith Johnson. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66(4). 789–806. doi:10.1016/j.jml.2011.11.006. http://dx.doi.org/10.1016/j.jml.2011.11.006.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin & Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, .

Han, Mieko S. 1994. Acoustic manifestations of mora timing in Japanese. *The Journal of the Acoustical Society of America* 96(1). 73–82. doi:10.1121/1.410376. https://doi.org/10.1121/1.410376.

Hay, Jennifer. 2007. The phonetics of 'un'. In Judith Munat (ed.), *Lexical creativity, texts and contexts*, 39–57. Amsterdam: John Benjamins.

Heitmeier, Maria, Yu-Ying Chuang & R. Harald Baayen. 2024. *The Discriminative Lexicon: Theory and implementation in the Julia package JudiLing*. to appear with Cambridge University Press.

Howson, Phil J. & Melissa A. Redford. 2019. Liquid coarticulation in child and adult speech. *Proceedings of the 19th International Congress of Phonetic Sciences* .

Kelso, J. A. S., Eric Vatikiotis-Bateson, Elliot L. Saltzman & Bruce Kay. 1985. A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America* 77(1). 266–280. doi:10.1121/1.392268.

Kubozono, Haruo. 2017. Mora and syllable. In Natsuko Tsujimura (ed.), *The handbook of Japanese linguistics*, chap. 2, 31–61. John Wiley & Sons, Ltd. doi:https://doi.org/10.1002/9781405166225.ch2. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405166225.ch2.

Kuehn, David P. & Kenneth L. Moll. 1976. A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics* 4(4). 303–320. doi:10.1016/s0095-4470(19)31257-4.

Kuperman, V., M. Pluymaekers, M. Ernestus & R. H. Baayen. 2007. Morphological predictability and acoustic duration of interfixes in Dutch compounds. *Journal of the Acoustical Society of America* 121(4). 2261–2271.

Landauer, Thomas K & Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2). 211.

Levelt, Willem J. M., Ardi Roelofs & Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22(1). 1–38.

Levelt, Willem J. M. & Linda Wheeldon. 1994. Do speakers have access to a mental syllabary? *Cognition* 50. 239–269.

Li, Vivian G., Sejin Oh, Garima Chopra, Joshua Celli & Jason A. Shaw. 2020. Articulatory correlates of morpheme boundaries: Preliminary evidence from intra- and inter-gestural timing in the articulation of the English past tense. *Proceedings of ISSP 2020 - 12th International Seminar on Speech Production* .

Lohmann, Arne. 2018a. *Cut* (N) and *cut* (V) are not homophones: Lemma frequency affects the duration of noun–verb conversion pairs. *Journal of Linguistics* 54(4). 753–777. doi:10.1017/s0022226717000378.

Lohmann, Arne. 2018b. Time and thyme are not homophones: A closer look at Gahl's work on the lemma-frequency effect, including a reanalysis. *Language* 94(2). e180–e190. doi:10.1353/lan.2018.0032.

Lu, Yuxin, Yu-Ying Chuang & R. Harald Baayen. published online 28 August 2024. Form and meaning co-determine the realization of tone in Taiwan Mandarin spontaneous speech: The case of Tone 3 sandhi. Published online on *arXiv*, 28 August 2024. https://arxiv.org/abs/2408.15747.

Malisz, Zofia, Erika Brandt, Bernd Möbius, Yoon Mi Oh & Bistra Andreeva. 2018. Dimensions of segmental variability: Interaction of prosody and surprisal in six languages. *Frontiers in Communication* 3(25). doi:10.3389/fcomm.2018.00025.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (eds.), *Advances in neural information processing systems*, vol. 26, 3111–3119. Curran Associates, Inc.

Nittrouer, Susan, Michael Studdert-Kennedy & Richard S. McGowan. 1989. The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research* 32(1). 120–132.

Noiray, Aude, Martijn Wieling, Dzhuma Abakarova, Elina Rubertus & Mark Tiede. 2019. Back from the future: Non-linear anticipation in adults and children's speech. *Journal of Speech, Language and Hearing Research* 62(8S). 3033–3054.

Öhman, S. E. G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America* 39. 151–168. doi:10.1121/1.1909864.

Plag, Ingo & Sonia Ben Hedia. 2018. The phonetics of newly derived words: Testing the effect of morphological segmentability on affix duration. In Sabine Arndt-Lappe, Angelika Braun, Claudine Moulin & Esme Winter-Froemel (eds.), *Expanding the lexicon*, 93–116. Berlin & Boston: De Gruyter Mouton.

Plag, Ingo, Julia Homann & Gero Kunter. 2017. Homophony and morphology: The acoustics of word-final s in english. *Journal of Linguistics* 53(1). 181–216.

Port, Robert F., Jonathan Dalby & Michael O'Dell. 1987. Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America* 81(5). 1574–1585. doi:10.1121/1.394510. https://doi.org/10.1121/1.394510.

R Core Team. 2022. R: A language and environment for statistical computing. https://www.r-project.org/.

Repp, Bruno H. & Virginia A. Mann. 1982. Fricative–stop coarticulation: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America* 71(6). 1562–1567.

Saito, Motoki, Fabian Tomaschek & R. Harald Baayen. 2021. Relative functional load determines co-articulatory movements of the tongue tip. *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)* 210–213.

Saito, Motoki, Fabian Tomaschek & R. Harald Baayen. 2023. Articulatory effects of frequency modulated by semantics. In Marcel Schlechtweg (ed.), *Phonology and phonetics*, De Gruyter.

Saito, Motoki, Fabian Tomaschek & R. Harald Baayen. under revision. Interaction of frequency and inflectional status: An approach from discriminative learning. *Language and Speech* .

Saito, Motoki, Fabian Tomaschek, Ching-Chu Sun & R Harald Baayen. 2024. Articulatory effects of frequency modulated by semantics. In Marcel Schlechtweg (ed.), *Interfaces of phonetics*, vol. 38, 125. Walter de Gruyter GmbH & Co KG.

de Saussure, Ferdinand. 1916. *Course in general linguistics.* McGraw-Hill.

Schmitz, Dominic, Dinah Baer-Henney & Ingo Plag. 2021a. The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords. *Phonetica* 78(5-6). 571–616. doi:10.1515/phon-2021-2013.

Schmitz, Dominic, Ingo Plag, Dinah Baer-Henney & Simon David Stein. 2021b. Durational Differences of Word-Final /s/ Emerge From the Lexicon: Modelling Morpho-Phonetic Effects in Pseudowords With Linear Discriminative Learning. *Frontiers in Psychology* 12(680889). 1–20. doi:10.3389/fpsyg.2021.680889.

Seyfarth, Scott, Marc Garellek, Gwendolyn Gillingham, Farrell Ackerman & Robert Malouf. 2018. Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience* 33(1). 32–49.

Shafaei-Bajestan, Elnaz, Masoumeh Moradipour-Tari, Peter Uhrig & R. H. Baayen. 2021. Ldl-auris: A computational model, grounded in error-driven learning, for the comprehension of single spoken words. *Language, Cognition and Neuroscience* 1–28.

Smith, Rachel, Rachel Baker & Sarah Hawkins. 2012. Phonetic detail that distinguishes prefixed from pseudo-prefixed words. *Journal of Phonetics* 40(5). 689–705.

Song, Jae Yung, Katherine Demuth, Stefanie Shattuck-Hufnagel & Lucie Ménard. 2013. The effects of coarticulation and morphological complexity on the production of English coda clusters: Acoustic and articulatory evidence from 2-year-olds and adults using ultrasound. *Journal of Phonetics* 41(3-4). 281–295.

Sproat, Richard & Osamu Fujimura. 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics* 21(3). 291–311. doi:10.1016/s0095-4470(19)31340-3.

Strycharczuk, Patrycja & J. M. Scobbie. 2016. Gradual or abrupt? the phonetic path to morphologisation. *Journal of Phonetics* 59. 76–91.

Sugahara, Mariko & Alice Turk. 2009. Durational correlates of English sublexical constituent structure. *Phonology* 26(3). 477–524.

The National Institute for Japanese Language. 2006. *Construction of the corpus of spontaneous japanese*, vol. 124 The National Language Research Institute Research Report. Tokyo, Japan: The National Institute for Japanese Language.

Tomaschek, Fabian, Ingo Plag, Mirjam Ernestus & R. Harald Baayen. 2019. Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics* 1–39. doi:10.1017/S0022226719000203.

Tomaschek, Fabian, Benjamin V. Tucker, Michael Ramscar & R. H. Baayen. 2021. Paradigmatic enhancement of stem vowels in regular english inflected verb forms. *Morphology* 31(2). 171–199.

Tucker, Benjamin V., Michelle Sims & R. Harald Baayen. published online 20 March 2019. Opposing forces on acoustic duration. doi:10.31234/osf.io/jc97w. Published online on *PsyArXiv*, 20 March 2019.

Walsh, Thomas & Frank Parker. 1983. The duration of morphemic and non-morphemic /s/ in English. *Journal of Phonetics* 11(2). 201–206.

Wood, Simon N. 2017. *Generalized additive models: An introduction with R*. Boca Raton, Florida, USA: CRC Press 2nd edn.

Zharkova, Natalia, Nigel Hewlett & William J. Hardcastle. 2012. An ultrasound study of lingual coarticulation in /sV/ syllables produced by adults and typically developing children. *Journal of the International Phonetic Association* 42(2). 193–208.

Zimmermann, Julia. 2016. Morphological status and acoustic realization: Findings from New Zealand English. In *Proceedings of the sixteenth australasian international conference on speech science and technology (sst-2016)* December, 201–204. Canberra: Australasian Speech Science and Technology Association (ASSTA).

Zuraw, Kie, Isabelle Lin, Meng Yang & Sharon Peperkamp. 2021. Competition between whole-word and decomposed representations of English prefixed words. *Morphology* 31(2). 201–237. doi:10.1007/s11525-020-09354-6.

A. Japanese phonology

Japanese language is based on moras, rather than syllables. For example, 王 /ou/ ([oː]) 'king' is about twice as long as 尾 [o] 'tail', because the former consists of one mora while the latter consists of two moras. For another example, 勝った /katta/ ([kat˺ta]/[katːa]) 'won' is about 1.5 times as long as 肩 /kata/ ([kata]) 'shoulder' with one more mora.

In addition, in Japanese, words are also distinguished by the position of a pitch accent. Each word has its own pitch accent position, where the pitch drops suddenly. For example, 箸 /ha↓si/ ([ha↓ɕi]) and 橋 /hasi↓/ ([haɕi↓]) are distinguished by the different pitches of the second mora. In the examples, the position of the pitch drop is marked by the symbol of the downstep '↓'. The pitch drop after the final mora of the word (e.g., 橋 /hasi↓/) represents a lower pitch of the next mora after the word. For example, when the word 橋 /hasi↓/ 'bridge' is marked by the focus particle that can mark the syntactic subject, the particle will receive a lower pitch, namely 橋が /hasi↓ga/ 'bridge (nom)'. Some words do not have a pitch drop. For those words, the next mora after the word is realized with the same pitch as the last mora of the word. For example, 端 /hasi/ ([haɕi]) 'edge' does not have a pitch drop, and consequently it is said to have the same pitch contour as 橋 /hasi↓/ 'bridge' by itself. However, the focus particle が /ga/ of 端が /hasiga/ will be higher in pitch than the same particle が /ga/ of 橋が /hasi↓ga/, due to the absence of a pitch drop for 端 /hasi/.

Japanese has approximately 14 consonants (depending on different ways of counts) and 5 vowels from the phonological perspective. Phonotactics in Japanese is relatively simple. No consonant cluster is allowed, at least phonemically, and no coda consonant is allowed, either, except for nasals. These constraints result in very simple phonotactics, which allows mostly open syllables with a simple onset, and codas that can only contain nasals or the first part of a geminate.

B. DLM

The current study focused on triphone-based representations of word forms. These choices are, however, not requirements or any inherent limitation of DLM. In DLM, form vectors can be defined in many ways, as long as they are represented in the form of vectors (i.e., a sequence of numbers).

B.1. Triphones

In DLM, form vectors can be defined by any size of n-grams, not only by trigrams/triphones. The current study adopted triphones rather than other sizes of n-grams for better interpretability. A triphone can be understood as a contextual representation (informally, an allophone) of a certain SINGLE segment, rather than a string of three segments. For example, [bæb] and [dæd] differ in phonetic realizations of tongue tip movements. For [bæb], the tongue can stay low for [æ], because [b] does not require any tongue tip movement. In contrast, [d] requires that the tongue tip is raised for its articulation. When [æ] is sandwiched between [d], tongue tip positions are expected to be higher at the onset and the offset of [æ], compared to the onset and the offset of [æ] but sandwiched between [b]. These different realizations are due to carryover in anticipatory coarticulations (Öhman 1966, Repp & Mann 1982, Nittrouer et al. 1989, Davidson 2005, Cheng et al. 2007, Zharkova et al. 2012, Song et al. 2013, Howson & Redford 2019, Noiray et al. 2019).

B.2. Phonetic representations

DLM does not require to use any particular type of linguistic representations. Word forms can be represented in terms of orthographic letters as well as phonetic segments. Lower levels of representations can also be utilized such as acoustic properties (e.g., spectrogram) (Shafaei-Bajestan et al. 2021) or tongue positions (Saito et al. 2024).

The current study made use of phonetic segments as the basic descriptive unit for word forms, over orthographic representations. This was because the current study involves phonetic realizations. Acoustic representations were not adopted to enable simpler word-type-based modeling. Acoustic realizations are always different from one token to another. While it is absolutely the fact of everyday-use of language, the token-based modeling would require an incremental way of estimating association weights, which would be computationally very heavy. In addition, a sufficient quantity and quality of acoustic data would be required to reliably estimate association weights. Due to consideration of these issues, the use of simple triphones was adopted in the current study. Compared to acoustics, phonetic segments already abstract away non-linguistic variability such as pitch differences between male and female speakers, and therefore the same word can be informed properly as the same word to DLM.

B.3. Word as the basic unit

DLM requires that word forms and word-meanings are defined as vectors. Depending on theoretical interests and objectives, any size of linguistic units can be used, which includes syllables, morphemes, words, phrases, and even sentences, for example (Baayen et al. 2019). The requirement also suggests that the basic unit does not have to be based on any linguistic unit. It can be a fixed time window such as 100 ms, or certain frequent tokens as utilized in recent GPT algorithms (Brown et al. 2020), for example.

The current study focused on word-level units, nevertheless. This was mainly due to make the study feasible. In order to estimate associations between forms and meanings, they need to be defined in terms of vectors. While defining forms for different sizes of linguistic units than words would be a manageable task, defining semantics for different sizes of linguistic units would not be straightforward and it would be an empirical question by itself how they should be defined. In contrast, there are quite a few algorithms available to estimate meanings of words such as word2vec (Mikolov et al. 2013) and fastText (Bojanowski et al. 2017). Therefore, it was preferred for the current study to derive word-meanings from a well-known algorithm reliably.

B.4. Example of incremental production

Suppose there are only two words in the example toy lexicon, namely *pays* [peɪ̯z] and *paid* [peɪ̯d]. Using triphones, the form matrix **C** for this toy lexicon can be set-up as in Equation (63), where `1` represents the diphthong [eɪ̯] (Baayen et al. 1996).

$$\mathbf{C} = \begin{array}{r} \\ \textit{pays} \\ \textit{paid} \end{array} \begin{array}{c} \begin{array}{ccccc} \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \end{array} \\ \left[\begin{array}{ccccc} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \end{array}\right] \end{array} \tag{63}$$

In addition, suppose that the meanings of the two words are defined in terms of the meanings of their stems and suffixes, as in Equation 64.

$$\mathbf{S} = \begin{array}{c|ccc} & \textsc{<pay>} & \textsc{<-s>} & \textsc{<-ed>} \\ \hline \textit{pays} & 1 & 1 & 0 \\ \textit{paid} & 1 & 0 & 1 \end{array} \tag{64}$$

From these two matrices, the comprehension and production weight matrices **F** and **G** will be estimated as in Equations (65) and (66).

$$\mathbf{F} = \begin{array}{c|ccc|} & \langle\text{PAY}\rangle & \langle\text{-S}\rangle & \langle\text{-ED}\rangle \\ \texttt{\#pl} & 0.500 & 0.250 & 0.250 \\ \texttt{plz} & 0.250 & 0.375 & -0.125 \\ \texttt{lz\#} & 0.250 & 0.375 & -0.125 \\ \texttt{pld} & 0.250 & -0.125 & 0.375 \\ \texttt{ld\#} & 0.250 & -0.125 & 0.375 \end{array} \tag{65}$$

$$\mathbf{G} = \begin{array}{c|ccccc|} & \texttt{\#pl} & \texttt{plz} & \texttt{lz\#} & \texttt{pld} & \texttt{ld\#} \\ \langle\text{PAY}\rangle & 0.667 & 0.333 & 0.333 & 0.333 & 0.333 \\ \langle\text{-S}\rangle & 0.333 & 0.667 & 0.667 & -0.333 & -0.333 \\ \langle\text{-ED}\rangle & 0.333 & -0.333 & -0.333 & 0.667 & 0.667 \end{array} \tag{66}$$

Suppose the speaker intends to produce *paid*. In other words, the speaker intends to create the meaning (i.e., semantic vector) of *paid* in the listener's head. The 'goal' semantic vector is the second row of Equation (64), which is repeated below for clarity as a single vector (Equation 67).

$$\mathbf{s}_{\text{paid}} = \begin{array}{c|ccc|} & \langle\text{PAY}\rangle & \langle\text{-S}\rangle & \langle\text{-ED}\rangle \\ \textit{paid} & 1 & 0 & 1 \end{array} \tag{67}$$

The speaker knows that they have not produced anything yet, and therefore the listener has not received any cues (i.e., sublexical forms). The set of sublexical forms the speaker has provided so far is therefore a zero vector as in Equation (68).

$$\mathbf{c}_{\text{paid},t=0} = \begin{array}{|ccccc|} \texttt{\#pl} & \texttt{plz} & \texttt{lz\#} & \texttt{pld} & \texttt{ld\#} \\ 0 & 0 & 0 & 0 & 0 \end{array} \tag{68}$$

Given the set of sublexical forms the speaker has produced (i.e., Equation (68), the speaker infers what the listener has understood so far. The listener should not have understood anything yet, because the speaker has said nothing yet (Equation (69)).

$$\hat{\mathbf{s}}_{\text{paid},t=0} = \mathbf{c}_{\text{paid},t=0} \cdot \mathbf{F} = \begin{array}{|ccc|} \langle\text{PAY}\rangle & \langle\text{-S}\rangle & \langle\text{-ED}\rangle \\ 0 & 0 & 0 \end{array} \tag{69}$$

Note that the comprehension weight matrix **F** is the comprehension side of the linguistic knowledge possessed by the speaker, not by the listener. This is because all these processes happen in the speaker's head.

Subsequently, the speaker decides what to say, based on what they think the listener has understood so far, namely Equation (69). More specifically, the speaker considers what meanings should be emphasized or clearly expressed to the listener at the moment in order to create the goal semantic vector (i.e., the meaning of *paid*) in the listener's head. This reasoning can be expressed as the difference vector between the goal semantic vector (i.e., $\mathbf{s}_{\text{paid}}$) and the meanings/semantic vector assumed to be achieved in the listener's head (i.e., $\hat{\mathbf{s}}_{\text{paid},t=0}$), as expressed in Equation (70).

$$\begin{aligned}
\mathbf{s}_{t=1} &= (\mathbf{s}_{\text{paid}} - \hat{\mathbf{s}}_{\text{paid},t=0}) = \\
&= \begin{array}{c} \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \end{array} \\ \left[\begin{array}{ccc} 1 & 0 & 1 \end{array}\right] \end{array} - \begin{array}{c} \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \end{array} \\ \left[\begin{array}{ccc} 0 & 0 & 0 \end{array}\right] \end{array} \\
&= \begin{array}{c} \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \end{array} \\ \left[\begin{array}{ccc} 1 & 0 & 1 \end{array}\right] \end{array}
\end{aligned} \tag{70}$$

The semantic vector at the moment (i.e., $\mathbf{s}_{t=1}$) drives the speaker to decide on what to say next. This makes a contrast to the original production process of DLM, in which the whole meaning of the target word drives the speaker to produce the entire form vector. The semantic vector at the moment is then mapped onto a form vector through the weight matrix at the moment $\mathbf{G}_{t=1}$. The association matrix $\mathbf{G}$ also has its own temporary state at each time step to ensure physiological validity. Only the sublexical forms that have the coarticulatory characteristics of the word-initial position can be produced at the word-initial position physiologically (see Section 2.4.2 for more details). The temporary state of the $\mathbf{G}$ matrix, namely $\mathbf{G}_{t=1}$ is defined with help of the $\mathbf{V}$ matrix. For the current toy lexicon, the $\mathbf{V}$ matrix is set up as in Equation 71.

$$\mathbf{V} = \begin{array}{c} \\ \texttt{\#pl} \\ \texttt{plz} \\ \texttt{lz\#} \\ \texttt{pld} \\ \texttt{ld\#} \\ \phi \end{array} \begin{array}{c} \begin{array}{ccccc} \texttt{\#pl} & \texttt{plz} & \texttt{lz\#} & \texttt{pld} & \texttt{ld\#} \end{array} \\ \left[\begin{array}{ccccc} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{array}\right] \end{array} \tag{71}$$

At the word-initial position, no sublexical form has not been produced yet. In other words, nothing (i.e., ϕ) is the current sublexical form, which determines next possible sublexical forms. Therefore, the row of ϕ in the $\mathbf{V}$ matrix is taken out and converted into a diagonal matrix, as in Equations (72) and (73).

$$\mathbf{v}_{t=0} = \mathbf{v}_{\phi} = \begin{array}{c c} & \begin{array}{ccccc} \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \end{array} \\ \phi & \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \end{array} \right] \end{array} \tag{72}$$

$$\mathrm{Diag}(\mathbf{v}_{t=0}) = \begin{array}{c c} & \begin{array}{ccccc} \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \end{array} \\ \begin{array}{c} \texttt{\#p1} \\ \texttt{p1z} \\ \texttt{1z\#} \\ \texttt{p1d} \\ \texttt{1d\#} \end{array} & \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] \end{array} \tag{73}$$

The production weight matrix $\mathbf{G}$ is filtered for the current time step to be $\mathbf{G}_{t=1}$ by multiplying $\mathbf{G}$ by $\mathrm{Diag}(\mathbf{v}_{t=0})$ as in Equation 74.

$$\begin{aligned} \mathbf{G}_{t=1} &= \mathbf{G} \cdot \mathrm{Diag}(\mathbf{v}_{t=0}) \\ &= \begin{array}{c c} & \begin{array}{ccccc} \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \end{array} \\ \begin{array}{c} \text{<PAY>} \\ \text{<-S>} \\ \text{<-ED>} \end{array} & \left[\begin{array}{ccccc} 0.667 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.333 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.333 & 0.000 & 0.000 & 0.000 & 0.000 \end{array} \right] \end{array} \end{aligned} \tag{74}$$

With $\mathbf{s}_{t=1}$ and $\mathbf{G}_{t=1}$, a form vector is produced for the moment $t = 1$ by mapping the temporary semantic vector $\mathbf{s}_{t=1}$ onto $\hat{\mathbf{c}}_{\text{paid},t=1}$ through $\mathbf{G}_{t=1}$, as in Equation (75).

$$\hat{\mathbf{c}}_{\text{paid},t=1} = \mathbf{s}_{t=1} \cdot \mathbf{G}_{t=1} = \begin{array}{c} \begin{array}{ccccc} \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \end{array} \right] \end{array} \tag{75}$$

Equation (75) indicates that the segment [p] is selected for this time step. This segment [p] has carryover-coarticulatory characteristics at the word-onset position and anticipatory-coarticulatory characteristics with the following diphthong [eɪ̯], namely `#p1`.

At the next time step $t = 2$, the speaker updates his assumption about the listener's understanding. The speaker has produced `#p1` (i.e., $\hat{\mathbf{c}}_{\text{paid},t=1}$), and they assume that the listener has received the cue correctly (i.e., $\mathbf{c}_{\text{paid},t=1}$).

$$\hat{\mathbf{c}}_{\text{paid},t=1} = \mathbf{c}_{\text{paid},t=1} = \begin{array}{c} \begin{array}{ccccc} \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \end{array} \right] \end{array} \tag{76}$$

Given this set of cues, the listener should have understood the following meaning:

$$\hat{\mathbf{s}}_{\text{paid},t=1} = \mathbf{c}_{\text{paid},t=1} \cdot \mathbf{F} = \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \\ \left[\; 0.50 \right. & 0.25 & \left. 0.25 \;\right] \end{array} \tag{77}$$

Note that the meanings of <-s> and <-ED> are activated to the same degree. This is because the sublexical form that has been provided so far (i.e., `#p1`) does not provide any evidence about the upcoming suffix. In other words, the meaning of <PAY> has been delivered relatively well. Given this understanding by the listener (in the speaker's assumption), the speaker aims at delivering the meaning of <-ED>. In contrast, the meaning of <-s> is activated to some extent, when it should not be activated at all. Therefore, the meaning of <-s> should be attenuated. This reasoning can be expressed mathematically as below:

$$\begin{aligned} \mathbf{s}_{t=2} &= (\mathbf{s}_{\text{paid}} - \hat{\mathbf{s}}_{\text{paid},t=1}) = \\ &= \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \\ \left[\; 1 \right. & 0 & \left. 1 \;\right] \end{array} - \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \\ \left[\; 0.50 \right. & 0.25 & \left. 0.25 \;\right] \end{array} \\ &= \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \\ \left[\; 0.50 \right. & -0.25 & \left. 0.75 \;\right] \end{array} \end{aligned} \tag{78}$$

This is the meaning in the speaker's head, which drives the speaker to decide what to produce next. At the same time, physiologically plausible sublexical forms are ensured by $\mathbf{G}_{t=2}$.

$$\mathbf{v}_{t=1} = \mathbf{v}_{\texttt{\#p1}} = \begin{array}{c|ccccc|} & \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \\ \texttt{\#p1} & 0 & 1 & 0 & 1 & 0 \end{array} \tag{79}$$

$$\begin{aligned} \mathbf{G}_{t=2} &= \mathbf{G} \cdot \text{Diag}(\mathbf{v}_{t=1}) \\ &= \begin{array}{c|ccccc|} & \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \\ \text{<PAY>} & 0.000 & 0.333 & 0.000 & 0.333 & 0.000 \\ \text{<-S>} & 0.000 & 0.667 & 0.000 & -0.333 & 0.000 \\ \text{<-ED>} & 0.000 & -0.333 & 0.000 & 0.667 & 0.000 \end{array} \end{aligned} \tag{80}$$

Based on $\mathbf{s}_{t=2}$ and $\mathbf{G}_{t=2}$, sublexical forms are predicted as follows:

$$\hat{\mathbf{c}}_{\text{paid},t=2} = \mathbf{s}_{t=2} \cdot \mathbf{G}_{t=2} = \begin{array}{ccccc} \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \\ \left[0.00 \right. & -0.25 & 0.00 & 0.75 & \left. 0.00\right] \end{array} \tag{81}$$

As indicated in Equation (81), p1d is selected. Conceptually saying, the speaker decided to produce the diphthong [eɪ̯] (i.e., 1), while initiating the articulation of [d] (i.e., d), which appears as anticipatory coarticulation.

At the next time step (i.e., $t = 3$), the speaker updates his assumption about the listener's understanding again. While the sublexical form p1d was activated only by 0.75, the word form is either perceived (i.e., 1) or not (i.e., 0) by the listener, once the speaker produces the sublexical form. Therefore, the speaker's assumption about the listener's understanding contains only 1 and 0 as below

$$\mathbf{c}_{\text{paid},t=2} = \begin{array}{c} \begin{array}{ccccc} \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & 1 & 0 \end{array}\right] \end{array} \tag{82}$$

The speaker's assumption about the listener's understanding for $t = 3$ is updated as follows:

$$\hat{\mathbf{s}}_{\text{paid},t=2} = \mathbf{c}_{\text{paid},t=2} \cdot \mathbf{F} = \begin{array}{c} \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \end{array} \\ \left[\begin{array}{ccc} 0.750 & 0.125 & 0.625 \end{array}\right] \end{array} \tag{83}$$

Based on this assumption about the listener's understanding, the speaker would feel the necessity to emphasize each semantic dimension as below:

$$\begin{aligned} \mathbf{s}_{t=3} &= \left(\mathbf{s}_{\text{paid}} - \hat{\mathbf{s}}_{\text{paid},t=2}\right) = \\ &= \begin{array}{c} \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \end{array} \\ \left[\begin{array}{ccc} 1 & 0 & 1 \end{array}\right] \end{array} - \begin{array}{c} \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \end{array} \\ \left[\begin{array}{ccc} 0.750 & 0.125 & 0.625 \end{array}\right] \end{array} \\ &= \begin{array}{c} \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \end{array} \\ \left[\begin{array}{ccc} 0.250 & -0.125 & 0.375 \end{array}\right] \end{array} \end{aligned} \tag{84}$$

Based on this 'necessity', the speaker decides the next sublexical form to produce:

$$\hat{\mathbf{c}}_{\text{paid},t=3} = \mathbf{s}_{t=3} \cdot \mathbf{G}_{t=3} = \begin{array}{c} \begin{array}{ccccc} \texttt{\#p1} & \texttt{p1z} & \texttt{1z\#} & \texttt{p1d} & \texttt{1d\#} \end{array} \\ \left[\begin{array}{ccccc} 0.000 & 0.000 & 0.000 & 0.000 & 0.375 \end{array}\right] \end{array} \tag{85}$$

The sublexical form 1d# is activated the most strongly, thus being selected by the speaker to produce at $t = 3$. Assuming the listener has heard what the speaker has produced correctly, the set of sublexical forms the listener has received is expressed as below:

$$\mathbf{c}_{\text{paid},t=3} = \begin{array}{c} \\ \end{array}\begin{array}{ccccc} \texttt{\#pl} & \texttt{plz} & \texttt{lz\#} & \texttt{pld} & \texttt{ld\#} \\ \left[\ 1\right. & 0 & 0 & 1 & \left.1\ \right] \end{array} \tag{86}$$

With this set of sublexical forms, the listener should have understood the following meaning:

$$\hat{\mathbf{s}}_{\text{paid},t=3} = \mathbf{c}_{\text{paid},t=3} \cdot \mathbf{F} = \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \\ \left[\ 1\right. & 0 & \left.1\ \right] \end{array} \tag{87}$$

In other words, the listener should have understood the meaning of *paid* correctly, as intended by the speaker, assuming that the listener has received all the sublexical forms produced by the speaker correctly. This assumption makes the speaker feel that they do not have to produce anything any more, at least in order to deliver the meaning of *paid* to the listener, as expressed as the zero vector between the goal semantic vector $\mathbf{s}_{\text{paid}}$ and the meaning the listener has understood so far in the speaker's assumption, namely $\mathbf{s}_{\text{paid},t=3}$.

$$\begin{aligned} \mathbf{s}_{t=4} &= (\mathbf{s}_{\text{paid}} - \hat{\mathbf{s}}_{\text{paid},t=3}) = \\ &= \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \\ \left[\ 1\right. & 0 & \left.1\ \right] \end{array} - \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \\ \left[\ 1\right. & 0 & \left.1\ \right] \end{array} \\ &= \begin{array}{ccc} \text{<PAY>} & \text{<-S>} & \text{<-ED>} \\ \left[\ 0\right. & 0 & \left.0\ \right] \end{array} \end{aligned} \tag{88}$$

Because there is no semantics that drives the speaker, no sublexical forms will be activated any more, as shown in Equation 89 below.

$$\hat{\mathbf{c}}_{\text{paid},t=4} = \mathbf{s}_{t=4} \cdot \mathbf{G}_{t=4} = \begin{array}{ccccc} \texttt{\#pl} & \texttt{plz} & \texttt{lz\#} & \texttt{pld} & \texttt{ld\#} \\ \left[\ 0\right. & 0 & 0 & 0 & \left.0\ \right] \end{array} \tag{89}$$

In summary, in the current toy lexicon with only two words *pays* and *paid*, the speaker is predicted to produce `#pl` first, `pld` second, and `ld#` last. Conditional semantic support values predicted at each time step for each sublexical form are summarized in Table 4 below.

Table 4. The conditional semantic support values at each time step for the example lexicon.

Selected	#pl	plz	lz#	pld	ld#
#pl	1.000	0.000	0.000	0.000	0.000
pld	0.000	-0.250	0.000	0.750	0.000
ld#	0.000	0.000	0.000	0.000	0.375

C. Simulated word and mora duration

In the current study, word duration was predicted more accurately by unconditional semantic support, and mora duration was predicted more accurately by conditional semantic support. One explanation of these selective effects is that each word has its own durational target, while moras toward the end of a word tend to be shorter than the moras preceding them, because of accumulation of contextual information provided by those preceding moras. Metaphorically, the word-specific durational target can be understood as the intercept of a regression line. Similarly, the decreasing trend in mora duration from one mora to the next within a word can be understood as the slope of a regression line. There are two possibilities with respect to whether words have their own durational target. There are also two possibilities with respect to whether mora durations decrease from the beginning to the end of each word. All the four combinations of these two factors were listed below and tested by simulating word and mora durations in the subsequent sections.

- Possibility 1: Word-durational intercept with mora-duration decrease
  - Mora duration decreases from the beginning to the end of each word. Word duration is the sum of the duration of the moras of the word and word-specific duration.
- Possibility 2: Word-duration intercept without mora-duration decrease
  - Each mora in a word has approximately the same duration. Word duration is the sum of the duration of the moras of the word and word-specific duration.
- Possibility 3: Mora-duration decrease without word-duration intercept
  - Mora duration decreases from the beginning to the end of each word. Word duration is the sum of these mora durations without any addition of word-specific duration.
- Possibility 4: No word-duration intercept and no mora-duration decrease

- Mora duration stays relatively the same from mora to mora within each word. Word duration is the sum of these mora durations without any addition of word-specific duration.

C.1. Possibility 1: Word-duration intercept with mora-duration decrease

In this possibility, mora duration is hypothesized to decrease from the beginning to the end of each word. Word duration is hypothesized to be the sum of the duration of the moras of the word and word-specific duration.

For this simulation, a total of 100 words were simulated with each word being randomly assigned with a variable number of moras between 2 and 5. Conditional semantic support values were subsequently simulated for constituent moras of each word. The degree, namely the slope, of decrease in simulated conditional semantic support was randomly chosen from the uniform distribution between -0.05 and 0. The slope was set to be always negative, because conditional semantic support should decrease overall from the beginning to the end of a word due to accumulated context from the preceding moras, unless the word has a suffix or other sublexical forms that have clear relationships to certain meanings.

In contrast, unconditional semantic support does not necessarily decrease toward the end of a word, because it does not take intra-word positions into account. Unconditional and conditional semantic supports, however, predict the same value for the word-initial sublexical form, because no context is available for the first sublexical form of the word. Therefore, unconditional semantic support values were simulated to have a similar value to the conditional semantic support value for the first mora of the word, and they were simulated to retain similar unconditional semantic support values throughout the word.

In the current assumption (Possibility 1), mora durations are hypothesized to decrease throughout a word and the decreasing trend should be captured by conditional semantic support. Accordingly, mora durations were simulated to decrease from the beginning to the end of each word. The degree of the decreasing trend in mora durations was randomly determined to be somewhere between a half and twice of the decreasing degree of the simulated conditional semantic support. The slope of the simulated mora durations was defined to be relative to the slope of the simulated conditional semantic support values to reflect the current assumption that conditional semantic support captures the decreasing trend in mora durations. To simulate a word-specific durational target, the simulated mora durations were shifted upwards by adding the mean of the simulated unconditional semantic support values. The mean of the simulated semantic support values was added to reflect the current hypothesis that unconditional semantic support captures word-specific durational targets. Lastly, the simulated mora durations were added up to define word duration.

Figure 5 shows the simulated mora durations, word durations, unconditional semantic support, and conditional semantic support for the first 6 words. Each panel in Figure 5 represents a word. The x-axis of each panel represents the moras of the word. For example, the first word ‘w0001’ has three moras, while the second word ‘w0002’ has four moras. Bars in each panel represent simulated mora duration. Mora durations decrease throughout each word for most of the words. It is, however, also possible that mora durations slightly increase due to randomness in simulation, as can be seen for the word ‘w0002’. The black horizontal lines indicate word duration normalized by the number of moras. Red and green dots and lines indicate simulated conditional and unconditional semantic support values respectively. Since the current hypothesis is that both of conditional and unconditional semantic support contribute to word and mora durations, red dots and lines (i.e., conditional semantic support) decrease in a similar way as mora durations, and green dots and lines (i.e., unconditional semantic support) approximate well the (normalized) word duration.

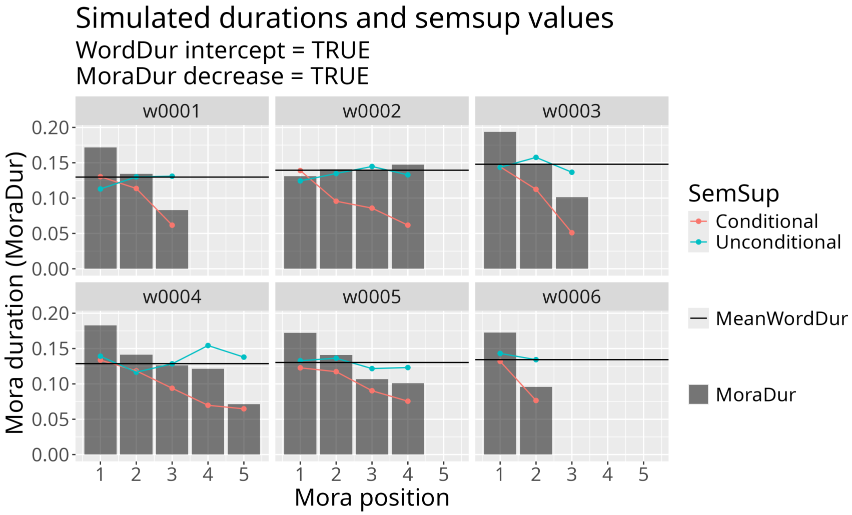
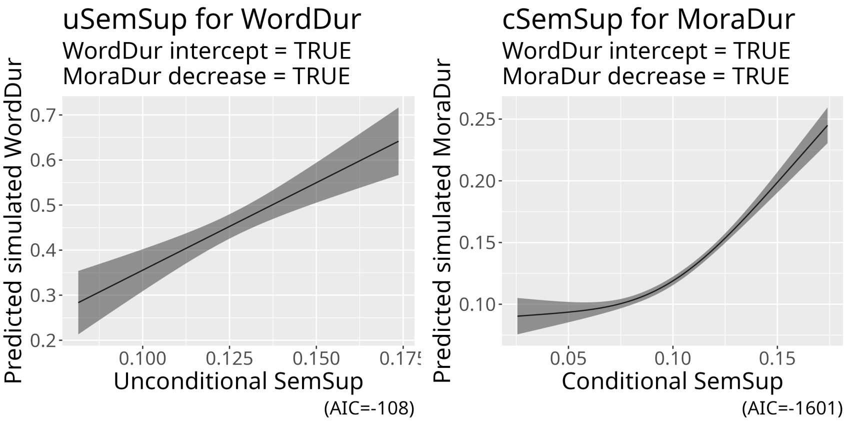


Figure 5. Simulated mora and word durations with word-duration intercepts and mora-duration decrease.

Based on these simulated variables, four GAMs were fitted with the dependent variable being either word duration or mora duration and with the predictor being either unconditional or conditional semantic support. Regarding word duration, the model with unconditional semantic support outperformed the model with conditional semantic support (ΔAIC = 21). Regarding mora duration, the model with conditional semantic support outperformed the model with unconditional semantic support (ΔAIC = 182). Unconditional and conditional semantic support were estimated to significantly contribute to word duration and mora duration respectively. These results of the model comparison are in line with those observed in the Japanese durational data (see section 4.2). Estimated effects of unconditional semantic support for word duration and effects of conditional semantic support for mora duration are visualized in Figure 6. Word duration and mora duration are both correlated positively with unconditional semantic support (Figure 6a) and conditional semantic support (Figure 6b) respectively. These estimated effects are also similar and in line with the actual observations for the real Japanese duration data (see Figures 1 and 3), including the leveling-off of the effects of conditional semantic support for mora duration around smaller conditional semantic support values.



(a) hoge

(b) hoge

Figure 6. hoge

C.2. Possibility 2: Word-duration intercept without mora-duration decrease

For this possibility, word duration and mora duration were simulated in such a way that each word had its own durational intercept but mora duration does not decrease throughout a word. The procedure of the simulation is identical to Possibility 1 in the previous section, except for mora duration. For the current simulation, mora duration was defined to be approximately constant throughout each word (with normally-distributed random noise), independently from the decreasing conditional semantic support values (Figure 7).

Consequently, the models with unconditional semantic support outperformed those with conditional semantic support to predict word duration and mora duration both (ΔAIC = 21 for word duration and ΔAIC = 226 for mora duration). Effects of unconditional semantic support were estimated to be positively correlated with word and mora duration in a linear manner (Figure 8). These observations on the current simulation data do not align with the actual observations for the Japanese durational data, in which mora duration was more accurately predicted by conditional semantic support.

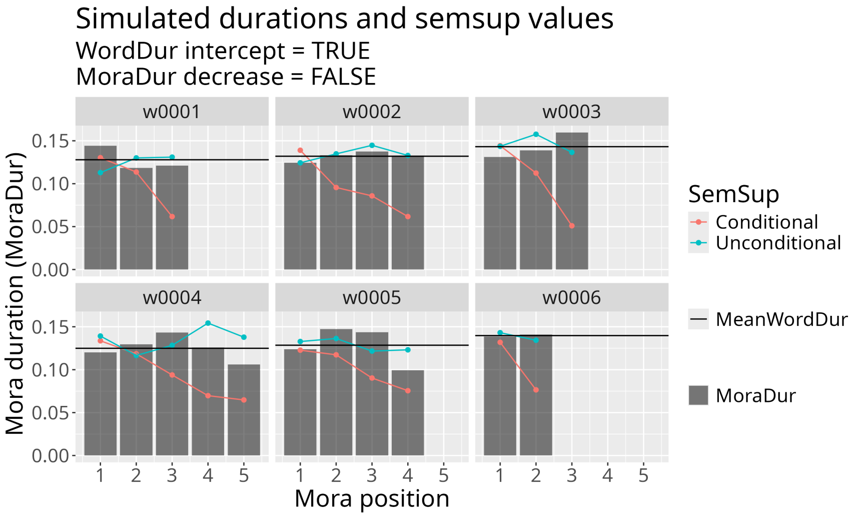
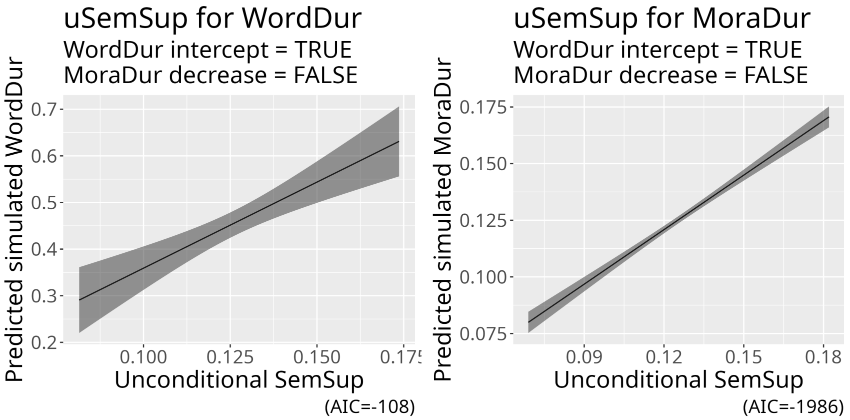


Figure 7. Simulated mora and word durations with word-duration intercepts and mora-duration decrease.



(a) hoge

(b) hoge

Figure 8. hoge

C.3. Possibility 3: Mora-duration decrease without word-duration intercept

For this possibility, mora duration was simulated to decrease throughout a word as conditional semantic support also decreased. No word-specific durational intercepts correlated with unconditional semantic support were included in the simulation. Figure 9 shows simulated durations and semantic support values for the first six words, where mora durations decrease (i.e., bars) as conditional semantic support (i.e., red dots and lines). Unconditional semantic support values (i.e., blue dots and lines), however, do not approximate (normalized) word duration (i.e., the black horizontal lines).

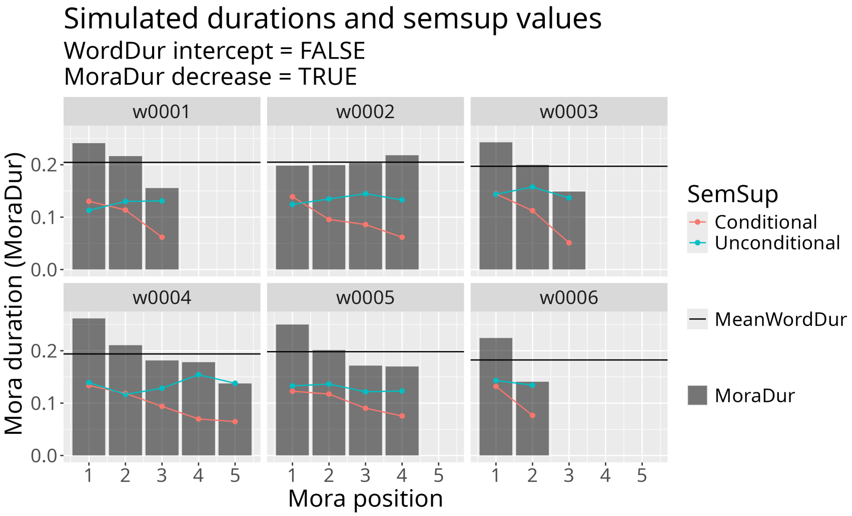
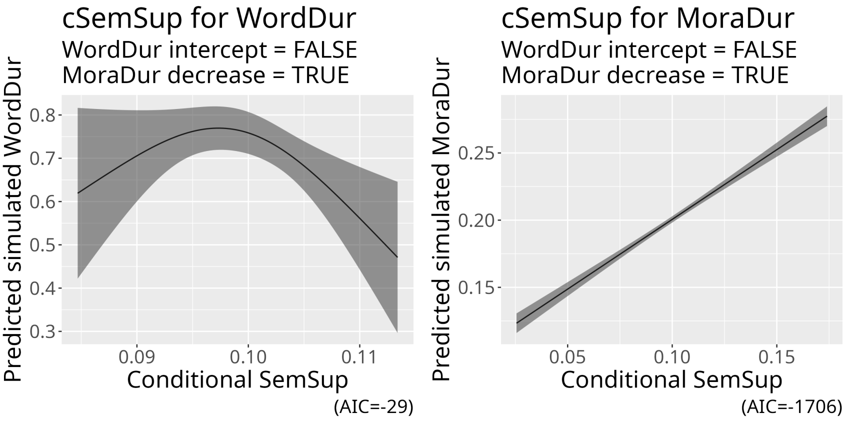


Figure 9. Simulated mora and word durations with word-duration intercepts and mora-duration decrease.

For this simulation, the models with conditional semantic support outperformed those with unconditional semantic support for word duration and mora duration both (ΔAIC = 7 for word duration and ΔAIC = 306 for mora duration). Predicted effects of conditional semantic support for word duration appeared in a non-linear manner, while those for mora duration was linear and positive (Figure 10). Superior performance of conditional semantic support to predict word duration over unconditional semantic support contradicts the actual observations for the real Japanese durational data. The estimated effects of conditional semantic support for simulated word duration (Figure 10a) are also qualitatively different. For the actual data, semantic support was positively correlated with duration. In addition, for the actual data, the effects of conditional semantic support was estimated to be somewhat non-linear with no effect for smaller values of conditional semantic support. The non-linearity was not captured in the current simulation as well (Figure 10b).

C.4. Possibility 4: No word-duration intercept and no mora-duration decrease

In this simulation, neither word-specific durational intercept nor decrease in mora duration within a word was taken into account. As a consequent, mora durations do



(a) hoge

(b) hoge

Figure 10. hoge

not decrease with conditional semantic support. Word duration was not associated with unconditional semantic support, either (Figure 11).

For this simulated data, the models with conditional semantic support outperformed those with unconditional semantic support to predict word duration and mora duration both, although the difference in AIC was quite small for mora duration (ΔAIC = 6 for word duration and ΔAIC = 1 for mora duration). The predicted effects of conditional semantic support for the simulated word and mora duration were qualitatively different from the actual observations for the real durational data (Figure 12).

C.5. Summary for the selective effects of (un)conditional semantic support

For the actual Japanese durational data, unconditional semantic support predicted word duration more accurately, while conditional semantic support predicted mora duration more accurately. These selective effects of unconditional and conditional semantic support were successfully simulated only by the simulated mora and word durations that took into account both of the word-specific durational target and the decrease in mora duration with a word. Logically, the same observations can be produced by different underlying processes. In this sense, this simulation cannot be decisive evidence to explain the selective effects of unconditional and conditional semantic support observed in the current study for the actual durational data. However, this simulation demonstrated the validity of the explanation for the current observations about the actual Japanese durational data that unconditional semantic support captured the word-specific durational aspect while conditional semantic support captured mora-level durational aspect.

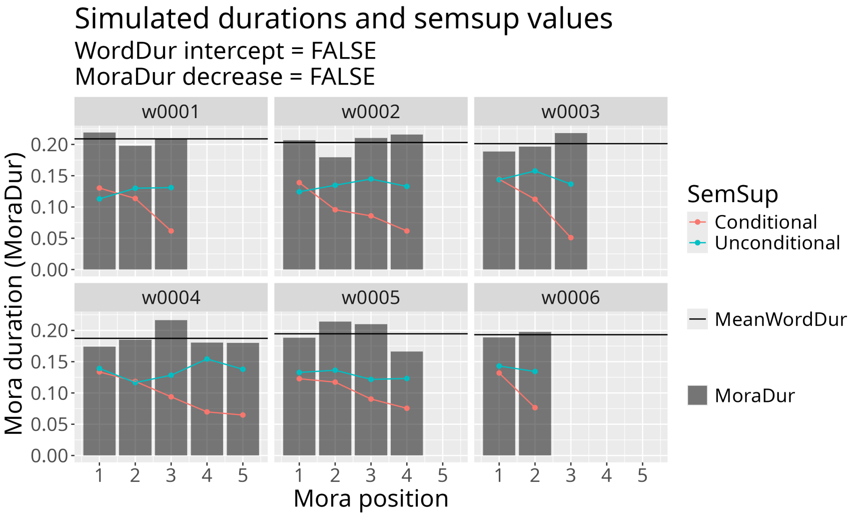
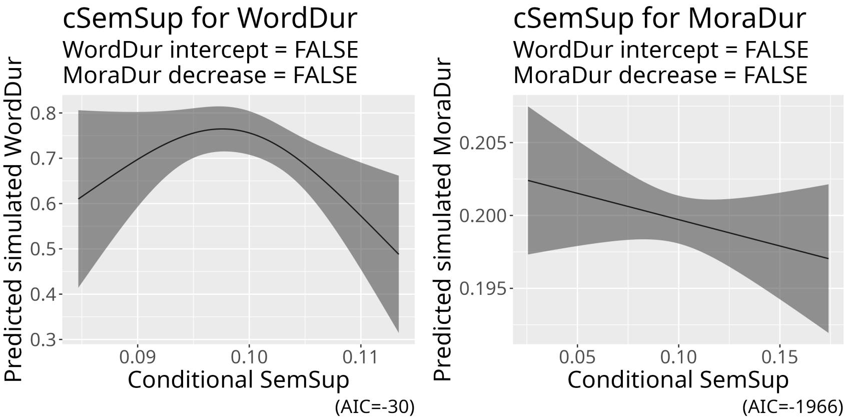


Figure 11. Simulated mora and word durations with word-duration intercepts and mora-duration decrease.



(a) hoge

(b) hoge

Figure 12. hoge

Authors' addresses: *(Saito)*
Eberhard-Karls-Universität Tübingen
Seminar für Sprachwissenschaft
motoki.saito@uni-tuebingen.de, Keplerstraße 2, 72074 Tübingen, Germany
motoki.saito@uni-tuebingen.de

(van de Vijver)
Heinrich-Heine-Universität
Institut für Linguistik
ruben.vijver@hhu.de, Universittätsstraße 1, 40225, Düsseldorf, Germany
ruben.vijver@hhu.de