LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Bilingual and Multilingual Mental Lexicon: A Modeling Study With Linear Discriminative Learning

Yu-Ying Chuang,[a] Melanie J. Bell,[b] Isabelle Banke,[a] and R. Harald Baayen[a]

[a]Eberhard-Karls University of Tübingen, Germany [b]Anglia Ruskin University, UK

Abstract: This study addresses whether there is anything special about learning a third language, as compared to learning a second language, that results solely from the order of acquisition. We use a computational model based on the mathematical framework of Linear Discriminative Learning to explore this question for the acquisition of a small trilingual vocabulary, with English as L1, German or Mandarin as L2, and Mandarin or Dutch as L3. Our simulations reveal that when qualitative differences emerge between the learning of a first, second, and third language, these differences emerge from distributional properties of the particular languages involved rather than the order of acquisition per se, or any difference in learning mechanism. One such property is the number of homophones in each language, since within-language homophones give rise to errors in production. Our simulations also show the importance of suprasegmental information in determining the kinds of production errors made.

Keywords bilingualism; multilingualism; mental lexicon; linear discriminative learning; homophony

Introduction

Is learning a third language qualitatively different from learning a second language? Does transfer to a third language take place only from the first language, or also from the second language (Hermas, 2015)? How is ultimate attainment affected by the point in time at which learning a new language begins? Starting early may be advantageous for mastery of a new language, but

Correspondence concerning this article should be addressed to Yu-Ying Chuang Eberhard-Karls University of Tübingen, Germany. E-mail: yu-ying.chuang@uni-tuebingen.de

DOI: 10.1111/lang.12435

are there any consequences for mastery of the first language? Furthermore, are developmental trends different for comprehension and production?

In this study, we address these very general questions about the global system properties of bilingualism and multilingualism, using a specific computational model of lexical acquisition. Our computational framework is that of Naive Discriminative Learning (NDL; Baayen, Milin, Filipović Durdević, Hendrix, & Marelli, 2011) and its twin, Linear Discriminative Learning (LDL; Baayen, Chuang, Shafaei-Bajestan, & Blevins, 2019). Both NDL and LDL implement discrimination learning, which has a long history in physics (Kalman, 1960; Widrow & Hoff, 1960), statistics (formally, LDL implements multivariate multiple regression), and psychology (Ellis, 2006b; Ramscar, Dye, & McCauley, 2013; Ramscar & Yarlett, 2007; Rescorla & Wagner, 1972; Rescorla, 1988; Siegel & Allan, 1996). In discrimination learning, a learning system—which could be an animal, a human, or a computer—establishes associations between different input stimuli and corresponding outputs or behaviors. For the present study, the inputs for comprehension are sublexical units of form, and for production, dimensions of semantic similarity. The outputs are meanings or forms, respectively.

In the context of second language acquisition, discrimination learning, as formalized by the learning rule of Rescorla and Wagner (1972), has been discussed by Ellis (2006a, 2013) and Ellis and Larsen-Freeman (2009). Ellis (2006a) found that the one-way dependency statistic ΔP (Allan, 1980) was useful for the quantitative evaluation of the consequences of discrimination learning for L2 acquisition. The ΔP statistic assesses the probability of a particular output class given a particular input feature, minus the probability of the same output class in the absence of that input feature. The present study seeks to move this line of research forward by using a more fine-grained quantification of learning. Instead of ΔP, we use simple two-layer neural networks, one for lexical comprehension and another for lexical production. These networks are part of a more comprehensive theory of the mental lexicon that integrates auditory comprehension, visual comprehension, and speech production: namely, the "Discriminative Lexicon" theory proposed by Baayen et al. (2019).

For auditory comprehension, computational models of the Discriminative Lexicon (i.e., NDL and LDL) take real speech as input (for empirical results see Arnold, Tomaschek, Lopez, Sering, & Baayen, 2017; Shafaei Bajestan & Baayen, 2018). For visual comprehension, the input can be either low-level visual features (Linke, Broeker, Ramscar, & Baayen, 2017; Serre, Wolf, & Poggio, 2005) or orthographic features, typically letter n-grams (with small *n*, i.e., strings of letters). For production, semantic input drives the selection of

triphone units that are in turn the input for articulation. Research on implementing speech production using a physical model of the vocal tract is ongoing (Sering, Stehwien, & Gao, 2019). In the present simulation studies, we make use of triphones both as input features for comprehension (simplifying the complexities of actual auditory word recognition), and as targets for speech production (following the modeling of production in Baayen et al., 2019). Triphones can be seen as representations of sounds that take into account that the articulation and comprehension of phones is highly context dependent. For instance, the place of articulation of stops is reflected in different formant transitions in adjacent vowels, and it is these formant transitions that play an important role during comprehension for distinguishing between [p], [t], and [k]. Furthermore, triphones implicitly encode sublexical order information, which is exploited by our model for modeling speech production. For further discussion, see Baayen et al. (2019).

Because the Discriminative Lexicon theory is computationally implemented, it offers novel opportunities to explore, by means of simulation experiments, various aspects of the acquisition of multiple languages. For example, we can investigate how acquisition of a second and a third language is affected by the degree of similarity between the first and subsequent language(s) (cf. Bardel & Falk, 2012; Hawkins & Lozano, 2006). We can also explore how proficiency in production relates to proficiency in comprehension (Mosca & de Bot, 2017). In addition, we can vary the extent to which different languages are used in order to model balanced and asymmetric bilingualism and multilingualism. This enables us to study the influence of usage on acquisition of a new language, and also its consequences for the existing language(s), which, when not used on a regular basis, run the risk of undergoing attrition. Finally, we can begin to model aspects of the day-to-day problems that come with being multilingual, such as language intrusion, that is, unintentionally using a different language from the one intended (cf. Jarema, 2017; Tytus, 2018).

Simulation studies, such as those presented in this article, have the advantage of enabling a researcher to manipulate one factor while holding all others strictly constant. This is seldom possible for experiments carried out with actual speakers. On the other hand, computational models, by their very nature, provide simplified windows on the complex phenomena they seek to illuminate. Aspects of language learning that are ignored by our simulations include various strategic effects (Mosca, 2019), the many social factors influencing which language is most appropriate for communication (Davydova, Tytus, & Schleef, 2017), and the role of meta-linguistic knowledge (Falk, Lindqvist, & Bardel, 2015).

The only computational models we are aware of that have previously been used to address bilingual language processing are the Bilingual Interactive Activation Plus (BIA+) model (Dijkstra & van Heuven, 2002; van Heuven & Dijkstra, 2010) and the MULTILINK model (Dijkstra et al., 2019). Like computational models based on the Discriminative Lexicon, BIA+ and MULTILINK address lexical processing. However, they differ from Discriminative Lexicon models in terms of their underlying architecture. BIA+ and MULTILINK build on the Interactive Activation model of McClelland and Rumelhart (1981), whereas the Discriminative Lexicon finds its roots in learning theory (Rescorla, 1988; Rescorla & Wagner, 1972; Widrow & Hoff, 1960) and multivariate linear regression (Baayen, Chuang, & Blevins, 2018; Sering, Milin, & Baayen, 2018). Computational implementations of the Discriminative Lexicon therefore differ in several important respects from BIA+ and MULTILINK.

The first difference between Discriminative Lexicon models and interactive activation models is that the latter are much more computationally costly. The mechanism of interactive activation is in fact so costly as to be implausible from the perspective of neural computing, as discussed in detail by Gurney, Prescott, and Redgrave (2001a), Gurney, Prescott, and Redgrave (2001b), and Redgrave, Prescott, and Gurney (1999). One particularly problematic aspect of the interactive activation framework is that the number of inhibitory connections between words increases quadratically with the number of words, such that for a lexicon with 50,000 words, no fewer than 2.5 billion inhibitory between-word connections are required. For the modeling of lexical processing with realistically sized lexicons, this renders the interactive activation architecture computationally intractable. It is even more cognitively unattractive because the same high-cost algorithm is supposedly also in place for the many other domains in cognition in which classification problems have to be solved (e.g., vision, audition, olfaction, and sensori-motor discrimination).[1]

A second difference is that, unlike models based on the Discriminative Lexicon, those based on interactive activation cannot model the time course of learning. MULTILINK and BIA+ have various parameters that can be manipulated to adjust the performance of the model. These parameters include a "resting activation level" for each word form, which is related to the frequency of that word form in the language. Dijkstra et al. (2019) suggest that the differences in processing between, for instance, early and late bilinguals could be modeled in MULTILINK by varying the resting activations in the network: The assumption is that late bilinguals will have used L2 words less frequently than early bilinguals, resulting in differences in the resting activation for L2 words. However, modeling the progress of learning over time (see, e.g.,

Ramscar et al., 2013) remains out of reach for such models, because the resting activations and other parameters represent only the final state of the system once learning is complete. In contrast, incremental learning is an inherent feature of models based on the Discriminative Lexicon. This makes it possible to model the time-course of learning and to compare how learning progresses as new languages are encountered.

The third difference is that interactive activation models require more storage than Discriminative Lexicon models. BIA+ and MULTILINK are representationally greedy models that adopt the basic functional architecture of classical paper dictionaries. Both models work with form representations and semantic representations that are stored in the computer's memory. These representations are localist, meaning that each node represents a single word or concept, analogous to the entries in a dictionary. Looking up the meaning of a word in a dictionary involves first finding its form entry, the key to the form's possible meanings. Similarly, in BIA+, word form units are activated first. These, in turn, activate semantic units, and subsequently also get activated by semantic units. In contrast, the Discriminative Lexicon is lean in representation. In reading, for instance, the visual input constitutes an external stimulus that produces a pattern of activation over lower-level orthographic features, for example, trigraphs, rather than whole word forms. This pattern of activation leads to another pattern of activation in a pool of semantic features. These patterns are created dynamically, instead of being retrieved as a whole from memory. This means that Discriminative Lexicon models require much less storage. For instance, whereas adding an extra word to an interactive activation model would always involve creating extra nodes for the word form and its meaning, this is not necessary to the same extent in the Discriminative Lexicon approach. Firstly, the same set of form features are used across the whole lexicon, so no additional form representations are necessary when new words are added. Secondly, a finite set of features are used to represent grammatical properties, and therefore morphologically complex words can be added to the lexicon without requiring additional semantic representations. For example, instead of representing the meanings of *go* and *went* with two unique nodes, these two words are each represented by more than one piece of semantic information, such as tense in addition to the lexical base: The two words share the base meaning of GO, but differ in the meaning domain of TENSE (i.e., PRESENT *vs*. PAST). Thus, when a new complex word is added to the lexicon, provided the base is already present, only the association strengths on the connections between form and meaning units require updating, which is part and parcel of the process of incremental learning (Sering et al., 2018).

A fourth difference is that, because BIA+ and MULTILINK implement a localist approach to semantics, as described in the previous paragraph, they cannot model relationships between words, either within or between languages. Dijkstra et al. (2019) acknowledge that their localist approach simplifies the true complexity of lexical semantics (see De Groot, 2011; Pavlenko, 2009). By representing words' meanings as discrete units, these models are not only glossing over the intricacies of cross-language differences in semantics and conceptualization, but they are also positing that, within a given language, all words have meanings that are completely unrelated to one another. In contrast, Discriminative Lexicon models represent words' meanings as vectors in a high-dimensional semantic space.[2] One approach to creating such a semantic space is to use one of the many methods developed within the framework of distributional semantics (e.g., Landauer & Dumais, 1997; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Turney & Pantel, 2010). In the present study, we construct a semantic space differently, building on the ontology of the lexical database WordNet (Fellbaum, 1998). Importantly, once words' meanings are represented as numerical vectors, it becomes possible to study various aspects of their interrelatedness using mathematical techniques. For example, the emotional content of words can be modeled straightforwardly (Westbury, 2014; Westbury et al., 2015), which in turn can shed light on differences in the emotional connotations of comparable concepts across different languages (Čavar & Tytus, 2018).

A final difference between Discriminative Lexicon models and interactive activation models is that BIA+ incorporates a mechanism for modeling task effects, whereas this is currently not implemented for the Discriminative Lexicon approach. We return to this point in the General discussion.

One design property our model shares with the BIA+ and MULTILINK models is an assumption that bilingual and multilingual processing uses a single system for all languages. In other words, we build on the hypothesis that comprehension is language nonselective, such that encountering a linguistic form in any known language will cause activation within a single shared pool of form and meaning representations (cf. Brysbaert, Verreyt, & Duyck, 2010; Dijkstra, Moscoso del Prado Martín, Schulpen, Schreuder, & Baayen, 2005; Kroll, Van Hell, Tokowicz, & Green, 2010; Mulder, Dijkstra, Schreuder, & Baayen, 2014). We will assume a similar situation for speech production, such that a single set of semantic representations can lead directly to articulation of forms in any known language, without involving translation between languages. However, in the general discussion, we will return to this assumption and reflect on how possible asymmetries between languages in speech

production, as discussed by, for example, Kroll and Stewart (1994) and Kroll et al. (2010), might be accounted for within our framework.

The remainder of this article is structured as follows. In the next section, we introduce further details about the representations of form and meaning that we used in our simulations, as well as the algorithms that predict meaning from form (for comprehension), and form from meaning (production). We then describe the English, German, Mandarin, and Dutch data that we used, before presenting in turn the simulations of monolingual, bilingual, and trilingual lexical learning. Finally, we discuss the implications of our results in the General Discussion.

Representations and Algorithms

Representing Meaning

A central question for a computational model of bilingual and multilingual lexical processing is how to design representations of word meanings. The BIA+ and MULTILINK models make use of localist semantic representations, which are assumed to be shared across languages. Thus, English *raspberry* and Dutch *framboos* are assumed to link up to the same semantic unit, known under the botanic name *Rubus idaeus*. But although localist representations of meaning have been widely used in computational models of the bilingual lexicon, they have two serious drawbacks. Firstly, they cannot represent lexical semantic relations within a single language and, secondly, they cannot represent differences in usage across two or more languages.

The first disadvantage of localist representations is that they entail that the meaning of a given word is completely unrelated to the meaning of every other word in the lexicon. In other words, when word meanings are represented using one-hot encoding (i.e., with one unique node per word), then all words' meanings are orthogonal. For the bilingual lexicon, localist meaning representations make it possible for *dog* and *Hund* to share exactly the same meaning, while at the same time *dog* and *cat* are taken to be semantically completely unrelated.

The second disadvantage of localist semantic representations is that, as Dijkstra et al. (2019) acknowledge, full translation equivalence almost never exists for actual word pairs, and hence localist representations involve a simplification that is motivated by implementational convenience. For example, consider the English-Dutch word pair *raspberry*/*framboos*. Dutch speakers only associate *framboos* with the species *Rubus idaeus*, either the plant or its fruit, or perhaps with various drinks made from the fruit. In English, however, *raspberry* enjoys wider use. In addition to the three senses available for *framboos*, the Oxford English Dictionary lists an additional meaning for *raspberry*: "A

sound made by blowing with the tongue between the lips, suggestive of breaking wind." The etymology of this sense is thought to involve Cockney rhyming slang (*raspberry tart* for *fart*) and is therefore highly language-specific. A single semantic entry for the pair *raspberry/framboos* would not adequately capture this difference in usage. An overview of the great many ways in which the semantics of words can mismatch across languages is provided by Pavlenko (2009).

Dijkstra et al. (2019) suggest that distributional semantics might provide a means for setting up more realistic semantic representations. However, when considering words' meanings in the context of bilingualism and multilingualism, this is far less straightforward than it might seem. The central problem is that constructions and collocations differ between languages. As a consequence, when semantic vectors are constructed from corpora, for a given pair of translation equivalents e_1 and e_2, the words that tend to co-occur with e_1 will not always be translation equivalents of the words that tend to co-occur with e_2. This makes it difficult to directly compare the distribution of a word in one language with the distribution of its counterpart in another language, since it is unclear how the two sets of reference words should be mapped onto one another.

To illustrate the divergence of semantic vectors for translation equivalents, we considered 21 English-German translation pairs (*walk laufen, apple Apfel, mind Geist, dog Hund, beard Bart, bone Knochen, bottle Flasche, castle Schloss, ceiling Decke, ditch Graben, eye Auge, feather Feder, fox Fuchs, food Essen, fun Spass, gift Geschenk, guest Gast, heaven Himmel, kite Drache, leaf Blatt, cat Katze*), and extracted their semantic vectors from those provided at https://www.spinningbytes.com/resources/wordembeddings/ (Deriu et al., 2017). Correlations between translation equivalents ranged from $r = 0.27$ to $\boldsymbol{r} = 0.50$, of which only 14 were significant under Bonferroni correction at $\alpha = 0.05$. Importantly, between English *dog* and *cat*, and German *Hund* and *Katze*, correlations were much higher ($r = 0.83$ and 0.98, respectively) than those of the cross-language correlations for *dog/Hund* and *cat/Katze* ($r = 0.44$ and 0.43, respectively). In other words, semantic vectors constructed for individual languages separately can lead to lower estimations for cross-language similarities between translation equivalents than for within-language similarities between co-hyponyms. Thus, although the distributional method overcomes the localist problem of failing to represent lexical semantic relations within a language, it risks creating a different problem of underestimating semantic similarities across languages.

Within computational linguistics, a wide range of methods have been developed to work around the problem described in the previous paragraph (for a comprehensive review, see Ruder, Vulić, & Søgaard, 2019). One such method, developed for translating between multiple languages, takes one language as a pivot. In order to translate between any pair of source and goal languages, one first maps from the source language onto the pivot language, and subsequently from the pivot language to the target language. The language chosen as pivot is typically English (e.g., Smith, Turban, Hamblin, & Hammerla, 2017), as English is the language for which most computational resources are available.

As a basis for a computational model of multilingual cognition, it might perhaps be argued that a speaker's L1 is actually a pivotal language. However, adoption of the pivot method from computational linguistics as a model of human cognition would imply that a multilingual speaker has distinct lexical semantic representations for each language known. Speaking in one's third language, for instance, would involve an initial conceptualization in L1, followed by a mapping of the resulting semantic vector in L1 onto a semantic vector in L3, followed in turn by producing the corresponding word form in L3. Interestingly, in such a model, conceptualization would be a language-specific process. Thus, this approach is compatible with the perspective developed by Whorf (1953) that languages each have their own way of thinking, albeit allowing that mappings can be set up between language-specific thoughts.

A very different computer science approach to multilingual translation seeks to design a unified semantic space that is shared by all pertinent languages. To do so, one has to change the input for the algorithms that create semantic vectors from corpora, such that information from multiple languages is available at the same time for training. For instance, Duong, Kanayama, Ma, Bird, and Cohn (2017) trained a computational model to predict, from the words in a target word's contexts, not only the target word itself, but also its translation equivalents in the other languages under consideration. Alternatively, the model can predict the words in a target word's sentence, and at the same time also predict all the words in the corresponding sentences in the other languages (Ruder, Vulić, & Søgaard, 2019).

The method described in the previous paragraph presupposes that parallel multilingual corpora are available for training. Importantly, in this approach, words across different languages share exactly the same semantic vectors. Such models are designed to maximally exploit patterns of similarity in word use between languages, while minimizing reliance on language-specific knowledge. Typically, these models do not attempt word sense disambiguation prior to lexical learning, they are blind to idioms and multiword expressions, and

Table 1 Hypernym chains for the two senses of *palm*

	sense: HAND		sense: TREE
S1	palm:hand	S6	palm:tree
S2	area, region	S7	tree
S3	body part	S8	woody plant, ligneous plant
S4	part, piece	S9	vascular plant, tracheophyte
S5	thing	S10	plant, flora, plant life
		S11	organism, being
		S12	living thing, animate thing
		S13	whole, unit
		S14	object, physical object
S15	physical entity	S15	physical entity
S16	entity	S16	entity

can deal with word-internal morphological structure only in a crude way, by constructing semantic vectors for substrings of word forms (cf. Bojanowski et al., 2017).[3] Furthermore, since models are trained on all pertinent languages simultaneously, this approach lends itself only to the modeling of fully balanced multilinguals.

In the present study, we sought to avoid working with localist representations of word meaning, and we also sought to avoid some of the problems that come with current distributional models of multilingual semantics (cf. Ruder, Vulić, & Søgaard, 2019). We therefore took as our point of departure the strong semantic similarities perceived by bilingual or multilingual speakers for translation pairs, and started out by constructing semantic vectors that were identical across languages. However, as will become apparent below, we also included mechanisms by which semantic vectors in one language can be made to be highly similar, but not identical, to the corresponding semantic vectors in other languages.

Our implementation of semantic vectors builds on the ontology underlying the lexical database WordNet (Fellbaum, 1998), using the online version 3.1 available at http://wordnetweb.princeton.edu/perl/webwn. In WordNet, words are grouped into sets of "cognitive synonyms," referred to as "synsets." Each synset is said to represent a distinct concept, and comes with a brief definition ("gloss"). Word forms with multiple senses are represented in a corresponding number of synsets. The hierarchical organization of WordNet makes it possible to extract successive hypernyms for any sense of any word in the database. Take the word *palm*, for example. Table 1 presents the hypernym chains for two

senses of *palm*. For the HAND sense, we start with the relevant synset for *palm*, glossed as "the inner surface of the hand from the wrist to the base of the fingers" (palm: HAND). The direct hypernym of this synset is "area, region," from which it inherits the hypernyms "body part," "part, piece," "thing," "physical entity," and "entity." For the TREE sense, we start with the synset glossed as "any plant of the family Palmae having… palmate leaves" (palm: TREE), which has the direct hypernym "tree," and consequently all the other hypernyms listed in the right-hand column of Table 1. For each word sense in our data, we looked up its hypernyms in WordNet. For conciseness and ease of processing, each synset in the hypernym chain, including the synset for the word-sense itself, was assigned a unique identifier. For example, as can be seen in Table 1, the two senses of *palm* share two hypernyms, S15 ("physical entity") and S16 ("entity"). We used these synset identifiers as the basis for the semantic representations in our model.

Although there are undoubtedly differences in lexical organization across languages, we assume that there are also some basic commonalities in the way people think and experience the world, irrespective of the language or languages they speak. Our use of the WordNet ontology is an attempt to approximate to this common core by using semantic representations that are relatively language-independent compared, for example, to standard distributional vectors. For the HAND sense of *palm*, we believe it is likely to be shared knowledge that, for example, hands are body parts and that they are physical entities. Similarly, for the TREE sense of *palm*, we believe that speakers of different languages are likely to share the knowledge that, for example, trees are plants and that plants are living organisms. For other computational modeling studies that use WordNet to represent word meaning, see Harm and Seidenberg (2004) and Monaghan, Chang, Welbourne, and Brysbaert (2017); see also the BabelNet project for the further extension of WordNet to the multilingual domain (Navigli & Ponzetto, 2012, http://live.babelnet.org/).

The inclusion of the lexical meaning of the word itself in our semantic vectors does mean that, in this particular implementation of LDL, we would have to add an additional element of representation for any new word sense added to our model's lexicon. This is because, in this implementation, we are working exclusively with monomorphemic words. Just as a human learner has to expand their lexicon when they learn a new lexical base, so too does any computational model of the lexicon. Nevertheless, our representations are not localist, because we do not represent the entire meaning of a word with a single node.

By representing each word sense as a vector of ontological features, we overcame the first disadvantage of localist representations, namely, the assumption that all word senses are unrelated to one another. In our system, the degree of ontological relatedness of two senses is directly reflected in the degree of similarity of their semantic vectors. In the case of *apple* and *pear*, the semantic vectors are identical except for the unique identifiers "apple" and "pear," reflecting the strong similarities between these two types of fruit. Both vectors include the digit "1" in the cells corresponding to the features "edible fruit," "produce," "food," "solid," "matter," "physical entity," and "entity." Similarly, the vectors for *dog* and *cat* are identical except for four cells: The vector for *dog* includes the features "dog" and "canine," whereas the vector for *cat* includes the features "cat" and "feline." Our semantic vectors therefore allow us to represent the similarity in meaning between word pairs such as *apple* and *pear* or *dog* and *cat*, while at the same time reflecting ontological knowledge about the nature of apples and pears, dogs and cats, that is likely to be shared by speakers of different languages.

We needed to take steps to address the second disadvantage of localist representations, namely, the assumption of translation equivalence. Since we were extracting synset chains from WordNet and using them to represent translation pairs such as *dog* and *Hund*, the semantic representations of such pairs were initially identical. Although this identity does justice to the strong semantic similarity perceived by bilingual or multilingual speakers for translation pairs, complete identity is cognitively and linguistically unrealistic. Translation equivalents seldom are truly equivalent—"traduire c'est trahir" (Du Bellay, 2013). As a first step away from complete identity, we enriched the semantic vector of each word with an identifier for the language it belongs to. These language identifiers reflect our assumption that the language in which a word is used is one aspect of what speakers know about it. In our model, the language identifiers also act as a proxy for variation in usage between translation equivalents.

We created a semantic matrix in which we specified which synsets from the WordNet hierarchy describe the semantics of each word-sense in our data (i.e., the synset for the word-sense itself, and all synsets in its hypernym chain). The columns of the semantic matrix, henceforth ***S***, list all relevant synsets (cf. Table 1). The rows specify for the senses which of these features are present (1) or absent (0). Homophones have a row for each sense, so there are two rows for the word *palm*, one for the HAND sense, and one for the TREE sense. The last two columns provide language identifiers, coding which language a word belongs to. Thus, the English word *palm* (HAND) and its German counterpart

Table 2 A toy example of the bilingual mental lexicon

Word	Language	Phonological form	Meaning
palm	English	p,m	sense: HAND
palm	English	p,m	sense: TREE
Handfläche	German	h&ntflEx@	sense: HAND
Palme	German	p&lm@	sense: TREE

Note. Phonological forms are encoded using the DISC notation of the CELEX database, adapted to our data

Handfläche have semantic representations that are identical except for the language identifiers. Equation 1 shows the semantic matrix for the dataset shown in Table 2. This is a toy example of a bilingual mental lexicon which contains only the English homophone *palm*, and the German translation equivalents of its two senses, *Handfläche* and *Palme*.

$$\boldsymbol{S} = \begin{array}{c} \\ \textit{palm}\text{:HAND} \\ \textit{palm}\text{:TREE} \\ \textit{Handfläche} \\ \textit{Palme} \end{array} \begin{array}{c} \begin{array}{cccccccccccccccccc} \text{S1} & \text{S2} & \text{S3} & \text{S4} & \text{S5} & \text{S6} & \text{S7} & \text{S8} & \text{S9} & \text{S10} & \text{S11} & \text{S12} & \text{S13} & \text{S14} & \text{S15} & \text{S16} & \text{ENGLISH} & \text{GERMAN} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}. \end{array} \tag{1}$$

Importantly, the addition of language identifiers serves a double purpose. On the one hand, for comprehension, the model can now learn which language is being used, and hence, in principle, retrieve words' appropriate social meanings. On the other hand, the language identifiers are essential for enabling the model to select the proper word form for production given a word's semantic vector. Without a language identifier, the model would not have any information that could guide it to articulate *dog* versus *Hund*.

Representing Form

A second key question for any computational model of lexical processing concerns how to represent words' forms. This question has typically been addressed by breaking the problem down into two sub-questions: What are the relevant features of form, and where in a word's form are these features located? By way of example, consider the English orthographic word form *none*. The letters *n, o, e* are orthographic features present in this form, with *n* appearing in positions 1 and 3, *o* appearing in position 2, and *e* in position 4. In

this vein, most connectionist models construct their form vectors in two steps. First, a numerical (typically binary) representational format is established for each letter (or phone), irrespective of where in a word it occurs. Next, a fixed number of position-slots is defined. Each slot is then filled with the numerical representation of the letter or phone that occurs in that position. Thus, for the orthographic word form *none*, a total of four position-specific slots is set up, with the first and third slot receiving exactly the same numeric vector, namely, the vector specifying the letter *n*. This coding scheme is used both by the Interactive Activation model of McClelland and Rumelhart (1981) and by its bilingual extension, the BIA+ model (Dijkstra & van Heuven, 2002). These two models employ localist representations, implemented with one-hot encoding. Letters, for instance, are defined by a binary vector of length 26, with "1" in the cell corresponding to the relevant letter and "0" in every other cell. Other coding schemes are also possible, with a general constraint that the vectors for words' forms are unique and for all practical purposes orthogonal (i.e. vectors for letters or phones are uncorrelated). See, for example, the form representations used by the triangle model of Harm and Seidenberg (2004).

Unfortunately, slot + filler coding is beset with problems. Words have different numbers of phones or letters, and this raises the question of how to allocate letters to positions. One can, of course, define *n* slots for words of length 1 up to *n*, and assign the first letter (or phone) to the first slot, the second letter (or phone) to the second slot, and so on, adding a vector for the space character to final slots that are not used. But how to proceed with words such as *kind* and *unkind*? If *k* and *u* are assigned to the first slot of each of these words respectively, *i* and *n* to the second slot, and so on, then the two words have no slot-filler combination in common and are effectively represented as unrelated forms. We miss out completely on the similarity of *kind*, positions 1–4, with *unkind*, positions 3–6. When multiword compounds are taken into consideration, slot coding breaks down completely, so alternative solutions for representing words' forms are required. The MULTILINK model (Dijkstra et al., 2019) implements a radical move away from slot coding by using the Levenshtein edit distance to evaluate the similarity between a visual input (a sequence of letters) and orthographic word representations stored in the model's memory. Since Levenshtein distance quantifies the number of changes required to transform one form into another form, adoption of this measure allows the MULTILINK model to take into account the degree of similarity or dissimilarity between different word forms. However, adoption of the Levenshtein distance measure also means that Dijkstra et al. (2019) have given up on the usefulness of the interactive activation framework for modeling the initial

stages of word recognition. Instead of having activation flow between letters and words, we now have an abstract mathematical evaluation metric that does not have a straightforward neural interpretation.

In the present study, we adopted a radically different approach, first explored by Baayen et al. (2011). In this approach, the units representing aspects of a word's form are themselves context-sensitive. This context-sensitivity obviates the need for assigning form units to specific slots. More specifically, in order to represent a word's form, we first extract the letter or phone n-grams ($n > 1$) from that word. Setting n to 3, for the orthographic word form *simulation* we obtain the trigraphs `#si, sim, imu, mul, ula, lat, ati, tio, ion, on#`. Here, the `#` symbol represents the space character. We repeat this process for all word forms to which the model is exposed, and create a vector listing all k unique n-grams. The representation for a specific word form is defined as a binary vector of length k. Each position in this vector is associated with a particular n-gram. The values in the vector are set to 1 in the cells that correspond to the n-grams present in the word, and 0 everywhere else. In other words, a form vector specifies which of the language's possible phone or letter n-grams are present in a given word. In such a representation, order is implicit, since only certain linear sequences of n-grams are possible. From the pool of ten trigraphs in *simulation*, `#si` has to come first, because of the initial space character, `sim` must come second to achieve the required overlap of `si`, and so on.

In the approach of Baayen et al. (2011), a form vector is presented as input to a two-layer network that is trained to predict which word meaning is represented by the n-grams in that input. Wieling, Nerbonne, Bloem, Gooskens, Heeringa, and Baayen (2014) used such a network to calculate strengths of association between different dialectal form variants of a word and its meaning. They showed that the difference in the network's prediction strengths for two input forms was strongly correlated with the value of a Levenshtein edit distance measure applied to those forms (Wieling, Margaretha, & Nerbonne, 2012). In other words, these n-gram form features, which we also use in the present study, in combination with the algorithm used for training the network, provide the same functionality as the Levenshtein edit distance used by the MULTILINK model. However, unlike Levenshtein distance per se, the use of n-gram features is grounded in considerations of biological plausibility. The logic underlying form encoding with n-gram features is that both in the visual and auditory cortex, receptive fields specialized in detecting the presence of specific form features are known to modulate how sensory information is processed (Aertsen & Johannesma, 1981; DeAngelis, Ohzawa, & Freeman, 1995;

Eggermont, Aertsen, Hermes, & Johannesma, 1981; Hubel & Wiesel, 1962). Letter and phone n-grams are high-level proxies for such receptive fields in the respective sensory systems.[4]

In the present study, we based our form representations on triphones and created a matrix specifying the phonological properties of the words in our lexicon. This matrix has a column for each triphone that occurs in the lexicon, and in each row, binary coding specifies whether a triphone is present (1) or absent (0) in a given word form. This form matrix is henceforth denoted by $\boldsymbol{C}$ (which stands for "cues"). Equation 2 shows the form matrix for the toy dataset shown in Table 2.

$$\boldsymbol{C} = \begin{array}{c} \\ \textit{palm} \\ \textit{palm} \\ \textit{Handfläche} \\ \textit{Palme} \end{array} \begin{array}{c} \begin{array}{cccccccccccccccccc} \text{\#p,} & \text{p,m} & \text{,m\#} & \text{\#h\&} & \text{h\&n} & \text{\&nt} & \text{ntf} & \text{tfl} & \text{flE} & \text{lEx} & \text{Ex@} & \text{x@\#} & \text{\#p\&} & \text{p\&l} & \text{\&lm} & \text{lm@} & \text{m@\#} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}. \tag{2}$$

We represented phones using the computer-readable DISC notation, but replaced some symbols with novel symbols. For example, the vowel of *palm* is represented in DISC by the symbol "#," but since "#" is a boundary marker in our model, we replaced it with ",".

Algorithm: Linear Discriminative Learning

The third key question for a model of lexical processing concerns the algorithm that connects form to meaning, and meaning to form. Here, we make use of Linear Discriminative Learning (LDL; Baayen et al., 2018, 2019). LDL uses two-layer networks directly connecting form and meaning representations, without any hidden layers. Word forms and meanings are represented as numeric vectors. A network for comprehension and a network for production can be obtained to generate meanings from forms and forms from meanings, respectively.

To understand how LDL works, consider the toy bilingual lexicon shown in Table 2. Given the semantic matrix $\boldsymbol{S}$ (Equation 1) and the form matrix $\boldsymbol{C}$ (Equation 2), it is possible to obtain two networks, one for comprehension and the other for production. Both networks are fully connected, that is, every triphone in $\boldsymbol{C}$ is connected to every semantic feature in $\boldsymbol{S}$, as shown in Figure 1. The comprehension network uses triphones to predict semantic features. The production network goes in the other direction, using semantic

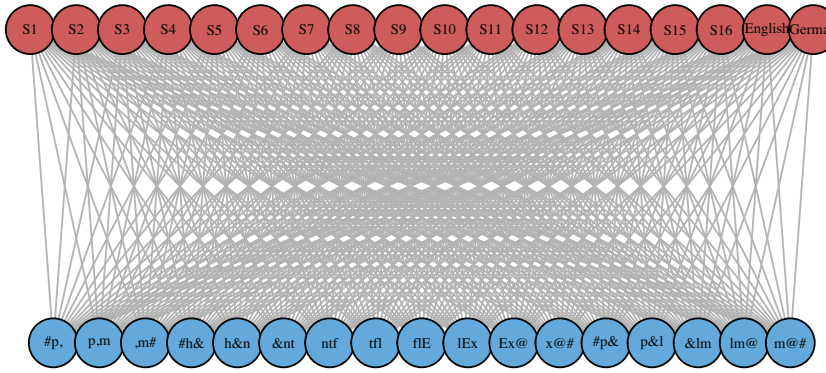

Figure 1 The fully-connected network between triphones and semantic features obtained with LDL. For comprehension, triphones are used to predict semantic features, whereas for production, semantic features are used to predict triphones. [Color figure can be viewed at wileyonlinelibrary.com]

features to predict triphones. Each connection between a triphone and a semantic feature has a particular strength of association, its connection weight. The predictions obtained when mapping form to meaning, or meaning to form, are determined by these connection weights.

There are two methods for estimating the weights on the connections between triphones and semantic features. One can either update the weights incrementally using a learning rule, or set up a system of equations that can be solved using matrix algebra. The first method records learning at different stages, enabling us to trace the trajectory of development and to observe how two or more languages interact during learning. The second method assumes that learning has reached a theoretical end-state, and thus the resulting networks represent fully-developed systems. In this study we used these two methods to explore, by simulation, both the time course of multilingual language acquisition and the theoretical end-point of learning under a variety of conditions.

To implement the first method of estimation, for incremental learning, we applied the Widrow-Hoff learning rule (Widrow & Hoff, 1960), which is related to the Rescorla-Wagner learning rule that figures prominently in the acquisition framework of Ellis (2013, 2006b). This is a form of supervised learning, which means that the model learns by being presented with successive pairings of an input and the corresponding desired output (for detailed discussion and optimized implementation, see Milin, Madabushi, Croucher, &

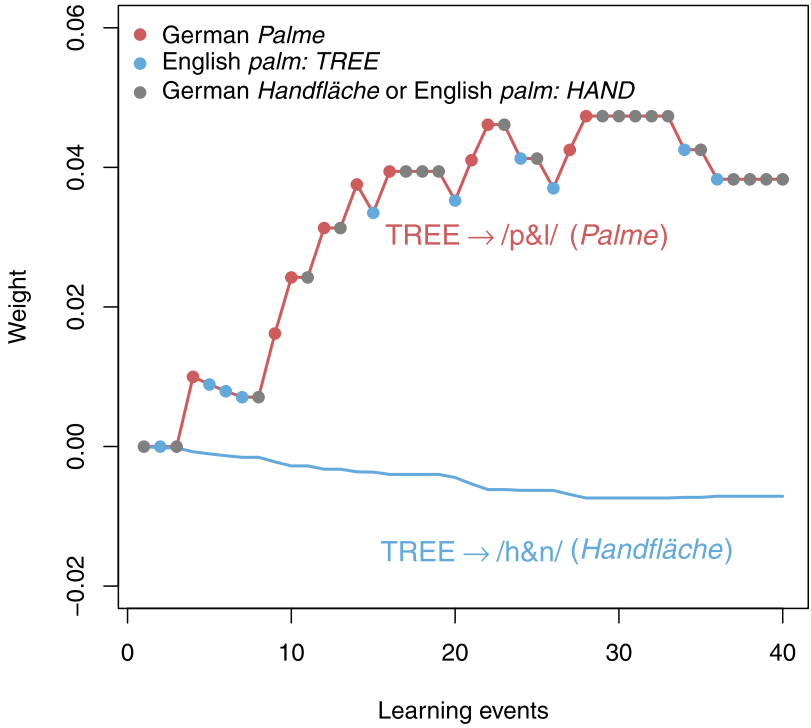

Figure 2 Changes in connection weights from the semantic feature TREE to the triphones /p&l/ and /h&n/ as learning progresses in the production network for the toy lexicon shown in Table 2. [Color figure can be viewed at wileyonlinelibrary.com]

Divjak, 2020). In the production network, for example, a learning event would involve presenting a word's semantic vector as the input and its form vector as the desired output. In the beginning, all weights are zero, but as learning progresses, they are gradually calibrated. At each learning event, the model predicts an output using the given input in conjunction with the current connection weights in the pertinent network. It then compares its prediction to the desired output for that learning event, and updates the weights accordingly. In general, the weight between a given input feature and a given output feature will increase when the two occur in the same input-output pairing, but decrease when the input feature occurs in the absence of the output feature.[5]

Figure 2 plots the changes in two connection weights as learning progresses. These weights are taken from the production network trained on the toy lexicon in Table 2. The red (upper) line in Figure 2 shows development

of the weight from the semantic feature TREE (S7, cf. Table 1) to the triphone /p&l/, while the blue (lower) line shows development of the weight from the same semantic feature to the triphone /h&n/. Each learning event is a discrete point in time at which the model is presented with a word (i.e., a pairing of meaning to form), and the weights are updated accordingly. Figure 2 shows 40 learning events in total, with 10 repetitions for each word. The presentation order of the words to the model was randomized. The dots on the red line indicate the targeted form of each learning event. The German *Palme* and English *palm* of the TREE sense are marked by red and blue, respectively, whereas the gray dots indicate learning events in which the TREE semantic feature was not involved, that is, the German *Handfläche* or the English *palm* of the HAND sense. It can be seen that whenever the model is presented with the German word *Palme* /p&lm@/ (sense: TREE), the connection weight between the semantic feature TREE and the triphone /p&l/ increases. In contrast, whenever the model is presented with the English word *palm* /p,m/ (sense: TREE), the connection weight between the semantic feature TREE and the triphone /p&l/ decreases slightly, because /p&l/ is not part of the targeted English form. When either of the other words is presented, that is, when the semantic feature is not present in the input, its weights are unchanged. As the association between the semantic feature TREE and the triphone /p&l/ becomes stronger, the connection weight between TREE and the other triphone, /h&n/, part of the German word *Handfläche* /h&ntflEx@/ (sense: HAND), gradually decreases. While learning to associate TREE with /p&l/, the model simultaneously learns to dissociate TREE from /h&n/, resulting in a negative weight on its connection to this triphone.

Turning now to the second method of estimation: In addition to modeling the learning process incrementally, LDL can also be used to model the theoretical end-state of learning, where learning is assumed to have continued indefinitely with an infinite number of learning events sampled from a given dataset. In this theoretical end-state, the system is in equilibrium, in the sense that any further learning events would lead to only insignificant changes in connection weights (see Danks, 2003, for detailed discussion). In other words, in this end-state, the system has the best possible weights to accurately map meanings to sounds and sounds to meanings in any of the languages it has learned. We can estimate the connection weights in this end-state by solving a system of equations with matrix algebra.

Given the representations of words' forms and meanings as the row vectors of the matrices $\boldsymbol{C}$ and $\boldsymbol{S}$, respectively, we can think of the comprehension

network, denoted by $\boldsymbol{F}$, as a transformation matrix that maps $\boldsymbol{C}$ onto $\boldsymbol{S}$. The pertinent mathematical equation is Equation 3:

$$\boldsymbol{CF} = \boldsymbol{S}. \tag{3}$$

Details of how to obtain $\boldsymbol{F}$ given Equation 3 are available in Baayen et al. (2018) and Baayen et al. (2019). Likewise, the production network is formally equivalent to a second transformation matrix, denoted by $\boldsymbol{G}$, which now maps $\boldsymbol{S}$ onto $\boldsymbol{C}$, as shown in Equation 4:

$$\boldsymbol{SG} = \boldsymbol{C}. \tag{4}$$

In general, there is no exact solution for Equations 3 and 4. Just as in linear regression it is impossible to draw a straight line through all the points in a data cloud, so it is impossible to exactly predict form vectors from meaning vectors, or meaning vectors from form vectors. We will denote best approximations of $\boldsymbol{F}$ and $\boldsymbol{G}$, that are optimal in the least squares sense, by $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{G}}$, respectively. Given $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{G}}$, the predicted semantic vectors (for comprehension) and form vectors (for production) are brought together in the prediction matrices $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{C}}$, given by Equations 5 and 6:

$$\boldsymbol{C}\hat{\boldsymbol{F}} = \hat{\boldsymbol{S}}, \tag{5}$$

$$\boldsymbol{S}\hat{\boldsymbol{G}} = \hat{\boldsymbol{C}}. \tag{6}$$

The rows of $\hat{\boldsymbol{S}}$ constitute the comprehension model's predicted meaning (i.e. semantic vector) for each word-form in the data, while the rows of $\hat{\boldsymbol{C}}$ constitute the production model's predicted form vector for each word-sense in the data. Model predictions can be obtained not only from the weight matrices estimated for the end-state of learning, but also from any intermediate stage of learning when weights are updated incrementally. In this case, $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{G}}$ are given by the weight matrices at that stage of learning.

To evaluate how well the model has learned the mapping of form to meaning, or meaning to form, it is necessary to compare the model's predicted vectors with the actual vectors in, respectively, the semantic matrix $\boldsymbol{S}$ or the form matrix $\boldsymbol{C}$. The first step is to obtain the predicted vectors. For comprehension, using the matrix method, we can take a word's triphone vector $\boldsymbol{c}$ and multiply it by the transformation matrix $\hat{\boldsymbol{F}}$ to produce the predicted semantic vector $\hat{\boldsymbol{s}}$ (Equation 5). Alternatively, with the incremental learning method, we can use the connection weights established in the relevant network at any given time. By way of example, Table 3 shows the connection weights from the three

Table 3 The connection weights for the triphones of *palm* to all the semantic features

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	ENG	GER
#p,	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.33	0.33	0.33	0
p,m	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.33	0.33	0.33	0
,m#	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.33	0.33	0.33	0
$\hat{\mathbf{s}}_{palm}$	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	0

Note. The sum of weights for every triphone in the word, for each semantic feature, constitutes the word's predicted semantic vector ($\hat{\mathbf{s}}_{palm}$).

triphones of the English word *palm* to all the semantic features, as part of the comprehension network for the toy example in Table 2. When given the triphones of *palm*, the model predicts the corresponding semantic vector by summing up the weights on the connections from each triphone in the word-form to each of the semantic features. The resulting predicted semantic vector is presented in the bottom row of Table 3.

The second step in evaluating the comprehension model is to establish how closely its predicted vectors correspond to the relevant target vectors. In our model, a word is assumed to be successfully recognized if its predicted semantic vector $\hat{\boldsymbol{s}}$ is more highly correlated with the actual semantic vector $\boldsymbol{s}$ of the target word than with the semantic vector of any other word in the lexicon. Given two row vectors, it is possible to calculate the degree of correlation between them in the same way that one might calculate the correlation between two paired variables in a scatterplot. Table 4 shows the predicted semantic vector for *palm* ($\hat{\boldsymbol{s}}_{palm}$) as well as the actual semantic vectors for both senses of *palm* and their German translation equivalents in our toy lexicon. The final column gives the correlation coefficient $\boldsymbol{r}$ for the correlation between $\hat{\boldsymbol{s}}_{palm}$ and each of the other vectors. However, the evaluation of homophone comprehension requires some extra consideration, since when a homophone is encountered without any context, it is impossible to know which meaning is intended. We therefore consider a homophone to have been correctly recognized if either of its possible senses is selected by the model. In the present example, it can be seen that the predicted semantic vector for *palm* is more highly correlated with the vector for the HAND sense of *palm* than with the vector for any other word. The model therefore selects *palm* (sense: HAND) as the output, and the input is considered to have been correctly understood.

For production, using a similar method to that described above for comprehension, we can calculate the predicted triphone vector and could then compare it with all the triphone vectors in the lexicon to find the closest match. However, in the case of production, identifying a target vector provides only part of the information needed to produce a word. The triphone vector tells the system which triphones are likely to be present in a word, but not how they should be ordered. Fortunately, as discussed above in the section on form representation, order is already implicit in the triphones themselves. Thus, for the word *Palme*, the word-initial triphone /#p&/ can be followed by /p&l/ but not, for example, by /h&n/ or /p,m/ due to the mismatch of the first and second phones, respectively. We make use of algorithms from graph theory to search for possible paths among highly activated triphones (i.e., triphones that receive strong semantic support), where a path is any possible sequence of overlapping

Table 4 Correlation coefficients (***r***) of the predicted semantic vector for *palm* ($\hat{\mathbf{s}}_{palm}$) with the semantic vectors (***s***) for the two senses of *palm* and their translation equivalents

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	ENG	GER	***r***
$\hat{\mathbf{s}}_{palm}$	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	0	
$\boldsymbol{s}_{palm:\text{HAND}}$	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	**0.52**
$\boldsymbol{s}_{palm:\text{TREE}}$	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	**0.43**
$\boldsymbol{s}_{\text{Handfläche}}$	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1	**0.03**
$\boldsymbol{s}_{Palme}$	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	**−0.09**

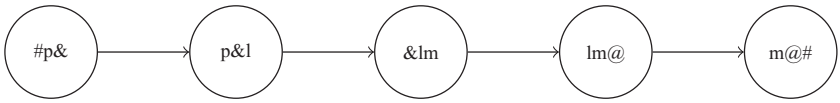

Figure 3 The triphone path for the phonological form of the word *Palme*.

triphones and "high activation" is defined by a threshold parameter in the model. The path for linking all the triphones in the word *Palme* is presented in Figure 3. Based on this path, the model ultimately outputs the predicted form of the word, which in this case is identical to the targeted form /p&lm@/.

For the toy lexicon in Table 2, only one path is found for each word, unsurprisingly given the small number of triphones in that lexicon. In reality, usually more than one path can be found, using different subsets, orders, or repetitions of the set of highly activated triphones, and hence more than one pronunciation is considered by the model. Under such circumstances, the model selects the form whose component triphones best predict the meaning that was input to the production system. Specifically, for each path found, the corresponding form vector $\boldsymbol{c}$ is constructed, and this is used to generate a predicted semantic vector $\hat{\boldsymbol{s}}$ using the comprehension network $\hat{\boldsymbol{F}}$. The form selected for production is the one whose predicted meaning best approximates to the meaning originally input to the production system. Formally, this is again accomplished by calculating the correlation between each candidate's predicted meaning and the original meaning. For example, assume that there are two candidate forms for the word *Palme*, /p&lm@/, and /p,m/. In the comprehension system, the corresponding triphone vectors predict the semantic vectors $\hat{\boldsymbol{s}}_{/p\&lm@/}$ and $\hat{\boldsymbol{s}}_{/p,m/,}$ respectively. These predicted semantic vectors are then found to be correlated with the actual semantic vector of *Palme* ($\boldsymbol{s}_{Palme}$) with correlation coefficients ($\boldsymbol{r}$) of 1 and −0.1, respectively. Therefore, the form /p&lm@/ is favored over /p,m/ and is selected as the target for articulation. Baayen et al. (2018) refer to this selection mechanism, which is effectively a process of production through internal comprehension, as "synthesis-by-analysis."[6]

Data

The multilingual lexicon constructed for the present study was built around an initial set of English and German translation equivalents, which included both homophonous and non-homophonous words in each language. The motivation for including homophones was that they pose a challenge for learning algorithms in a manner that is complementary to translation equivalents. Whereas

translation equivalents associate a single meaning with more than one form, homophones associate a single form with more than one meaning.

For English, homophones were taken from the norming study conducted by Maciejewski and Klepousniotou (2016). For German, homophones were selected on the basis of dictionary and web-based searches. In total, we included 102 English and 118 German homophone pairs. Among them, 27 pairs were shared across the two languages. For example, *summit* in English and *Gipfel* in German both have two senses: the top of a mountain and an important formal meeting. The dataset also included one homophone triplet in each language. In English, this was the word *skin*, which has the senses of body covering, the outer surface of an object, and to peel. In German it was the word *Platte*, which can refer to a record, a disc, or a slab. All senses of a homophone in either language were translated into the other language.

In addition to the homophones, we also included 27 words that were not treated as homophones in either English or German. However, the number of "non-homophones" in the dataset is actually much higher than this, for both languages.[7] This is because a homophone pair in one language often has translation equivalents in another language that are not themselves homophones. For example, *pupil* is translated into *Pupille* and *Schüler* in German, referring to the central part of an eye and a child at school, respectively. Similarly, the two senses of the German word *Decke*, when translated into English, are *blanket* and *ceiling*.

Starting from the English-German lexicon described above, we added two other languages, namely, Mandarin and Dutch. From the perspective of language typology, the former is a language distant from English and German, while the latter is a closely related one. All the Mandarin and Dutch words in the dataset are translation equivalents of their English and German counterparts, sharing exactly the same semantic vectors in the model, apart from their language identifiers. Table 5 shows the distribution of homophonous and non-homophonous words in the full dataset used for our simulations. Since English and German homophones were deliberately designed into this dataset, it is unsurprising that there are far fewer homophones in Mandarin and Dutch.

The phone representations of the English, German, and Dutch words, in DISC notation, were extracted from the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995). The phonological forms of the Mandarin words were transcribed also using DISC notation but with additional symbols for those Mandarin phones that are not included in the standard DISC phone set (e.g., retroflex sounds).

Table 5 The number of homophones and non-homophones for the four languages considered in this study. For Mandarin, the number of homophones and non-homophones is calculated either with (left) or without (right) lexical tones

	Homophone	Non-homophone	Total
English	207	198	405
German	239	166	405
Mandarin	40/36	365/369	405/405
Dutch	71	334	405

Because the shades of meaning of translation equivalents tend to diverge substantially from one another, it is unlikely that the actual relative frequencies of the various word senses in our data would be exactly correlated across all four languages. However, because of the already complex nature of our model, in which more than one form maps to a single meaning and vice versa, we wanted to avoid introducing additional variance, since this would have made it more difficult to understand the model's behavior. We therefore decided to control the frequency of each distinct sense so that, in our dataset, the same sense would have the same frequency in all four languages.

Given the well-established negative correlation of word length with frequency, we decided to base the frequency of each sense on the average length of the corresponding forms in the languages under consideration. To achieve this objective in a principled way, we simulated word frequencies from a lognormal-Poisson distribution. A total of 405 rates (λ) for the Poisson process were sampled from a lognormal ($\mu = 4, \sigma = 1$) distribution. For each word sense *i*, we then sampled a random number from a Poisson distribution (with λ_i), resulting in 405 integer-valued simulated frequencies of occurrences. These frequency values were assigned to the meanings, such that frequency was maximally inversely correlated with mean word length (averaged over English, German, and Mandarin[8]). The meaning with the highest simulated frequency in the current dataset (882) is "tea" (of the DRINK sense), and that with the lowest simulated frequency (1) is "installments."

Simulations of Monolingual Lexical Learning

All simulations in this and subsequent sections were carried out using the `R` package `WpmWithLdl` (Baayen, Chuang, & Heitmeier, 2019). The `R` code for running the simulations is provided in the supplementary material of this paper, which is available at https://osf.io/xq92s/.

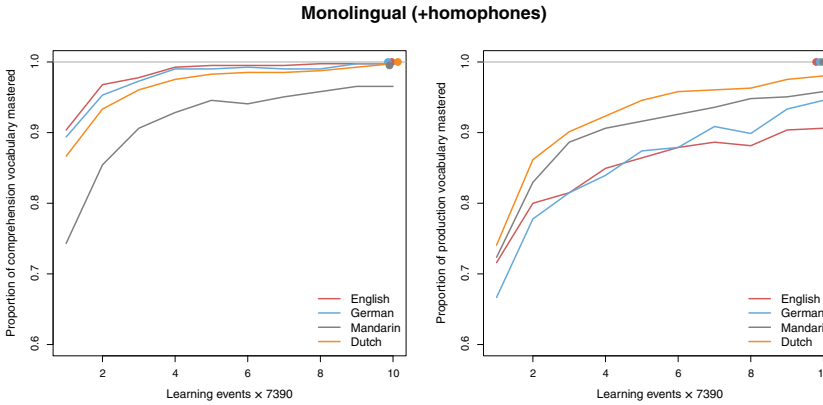

Figure 4 Vocabulary size as a function of exposure for comprehension (left) and production (right), for monolingual learning. The dots to the right of each plot indicate the model's performance at the end-state of learning. [Color figure can be viewed at wileyonlinelibrary.com]

In order to provide a baseline for the comparison of bilingual and trilingual lexical learning, we first carried out simulations of monolingual learning for all four languages under consideration. In each of these simulations, the model was trained on 73,900 word tokens, which is equal to twice the summed frequency of all word senses in the lexicon. Each word token constituted a learning event. The order of the learning events was randomized, and the same random order of senses was used for each of the four languages. Model performance was evaluated every 7,390 learning events.[9] For the monolingual simulations, each semantic vector contained 1,133 digits (0 or 1), which was the number of semantic features used. This number was the same for every language and excluded the language identifiers: In a monolingual lexicon, the language is already selected by default. In contrast, the number of triphones required varied between languages. The Dutch lexicon used the largest number of triphones (1216), followed by German (1002), then Mandarin (958) and finally English (908).

The left and right-hand panels of Figure 4 show, respectively, the proportions of words successfully recognized and produced at each evaluation during the simulation period. In general, comprehension develops faster than production, as can be seen from the fact that the lines in the left-hand panel are always higher than those in the right-hand panel. Furthermore, by the end of the simulation period, comprehension accuracy is higher than production accuracy in

all four languages. The general pattern of results fits well with the well-known asymmetry for comprehension and production (Blair & Harris, 1981; Clark, 1993; Ingram, 1974): Comprehension skills are usually acquired faster and earlier than the corresponding production skills.

The dots in the top right-hand corner of each graph indicate accuracy as estimated for the end-state of learning. In the limit of experience, when the model has reached equilibrium, comprehension accuracy reaches 100% for English, German, and Dutch, and 99.5% for Mandarin. Possibly the slightly lower accuracy for Mandarin is due to lexical tones not being represented in this simulation. In later sections, we will show how suprasegmental information can be added, and how it influences model performance. Production accuracy reached 100% for all languages.

The learning trajectories for the four languages show some differentiation. The left-hand panel of Figure 4 shows that English and German quickly approach error-free comprehension, Dutch takes a little more time and Mandarin is much slower, having not yet achieved optimal performance by the end of the simulation period. The right-hand panel shows a very different pattern for production. Here, English and German lag behind Mandarin and Dutch all the way through the simulation period. Since Dutch is typologically close to English and German, it is puzzling that, in terms of production, it patterns along with Mandarin.

Upon closer inspection, it turns out that the differences in our monolingual simulations, for both comprehension and production, are due to the different numbers of homophones in our four monolingual lexicons. Recall that for evaluating comprehension accuracy, a predicted meaning of a homophone is accepted as correct as long as it is one of the possible meanings of that homophone. This means that less precision is required in homophone comprehension, hence comprehension accuracy develops more quickly in the languages with more homophones. In production, however, homophones are especially susceptible to error. At the end of the simulation period, 37 out of the 38 English production errors involve homophones, and 20 out of the 21 German errors likewise involve homophones. The vulnerability of production to the presence of homophones actually originates in the comprehension system. This is because, as described above in the section on the LDL algorithm, the production model makes use of synthesis-by-analysis: The production system generates several candidate forms, which are fed back into the comprehension system to find the one that most closely matches the input meaning. However, for homophones, the mapping from form to meaning suffers from frailty, as explained in the following paragraph.

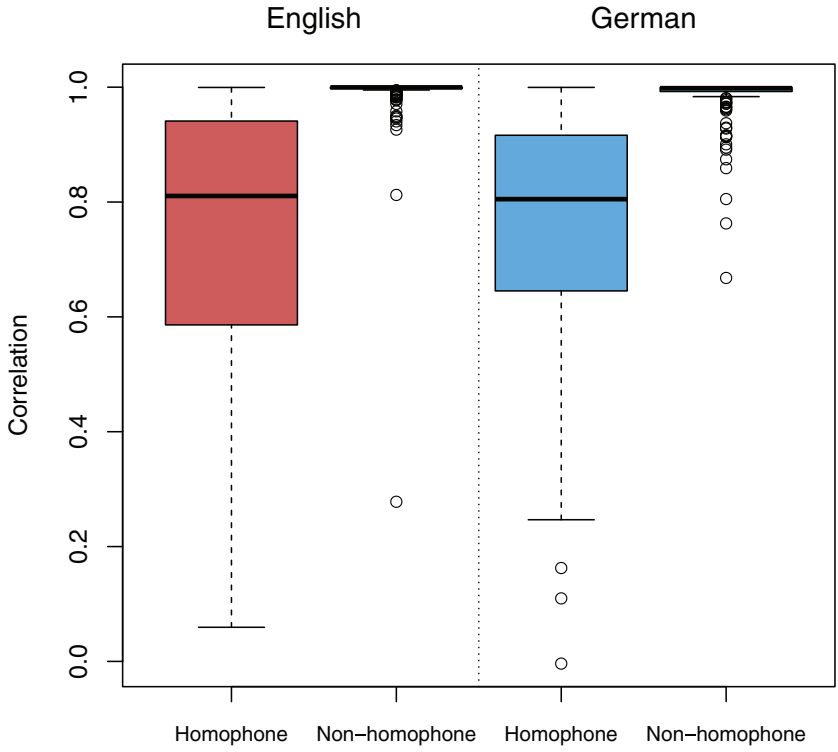

Figure 5 Correlations between predicted and targeted semantic vectors for homophones and non-homophones for English (red) and German (blue). [Color figure can be viewed at wileyonlinelibrary.com]

Consider Figure 5. The boxplots summarize the distributions of the correlation coefficients (***r***) between predicted and targeted semantic vectors for homophones as opposed to non-homophones at the end of the simulation period. (Every homophone contributes two or three correlation coefficients to the distribution, one for each sense.) The weaker correlations between predicted and targeted semantic vectors for homophones are an inevitable consequence of discrimination learning. Because a single homophonous form is associated with more than one sense, the comprehension system has to learn to associate the form's triphones with a wider range of semantic features, leading to lower weights on the relevant connections in the network. This means that homophones suffer from less precision and increased semantic ambiguity compared to non-homophones. It is noteworthy that in our production simulations,

whenever the presentation of a homophone leads to an error, the correct form is always listed among the candidates. This implies that the targeted forms have obtained sufficient support from the semantics to be present in the candidate set generated by the production system, but not enough to be selected during synthesis-by-analysis. If the list of candidate forms includes a competitor that predicts the input semantic vector more closely than the target form does, then this alternative form will be selected. Competitive alternative forms are typically closely related semantic neighbors of the target form. For example, in our simulations, *organ* (the INSTRUMENT sense) is produced incorrectly as *piano*, *almond* is produced as *walnut*, and *gold* is produced as *silver*. Other semantic errors include *marker* for *pen*, *lemon* for *orange*, and *hood* (the CAR sense) for *shield*. Because the predicted semantic vector for a homophonous target form is likely to be relatively weakly correlated with the target semantic vector, the probability of error increases, and wrong selections are more often found for homophones than for non-homophones. However, it is worth noting that such semantic errors do not persist through learning time. As shown in the right-hand panel of Figure 4, production becomes increasingly accurate as learning proceeds, and reaches error-free performance when learning reaches equilibrium.

To verify that the homophones are indeed the cause of the different learning patterns between languages in our production model, we created a second, smaller dataset in which we randomly selected just one meaning for each of the homophone pairs. This smaller set of 190 meanings therefore included no homophones in any language, with the exception of one homophone pair in Mandarin.[10] All words inherited their frequencies from the complete dataset. The summed frequency across all words for the homophone-free dataset was 16,482. We ran a simulation with a total of $2 \times 16{,}482 = 32{,}964$ learning events and evaluated comprehension and production every 2,747 learning events. As shown in Figure 6, the learning curves for the four languages are now much more similar, and they converge as learning progresses. Comparing Figure 6 with Figure 4 we see that, for comprehension, without the homophone advantage, English and German are now learned a little more slowly. In contrast, for production, performance in these languages is now substantially improved.

In summary, for monolingual learning, our model reproduces the comprehension-production asymmetry. Furthermore, our simulations reveal that homophones cause frailty in learning the mapping from form to meaning, and that this frailty considerably slows down accurate word learning in production.

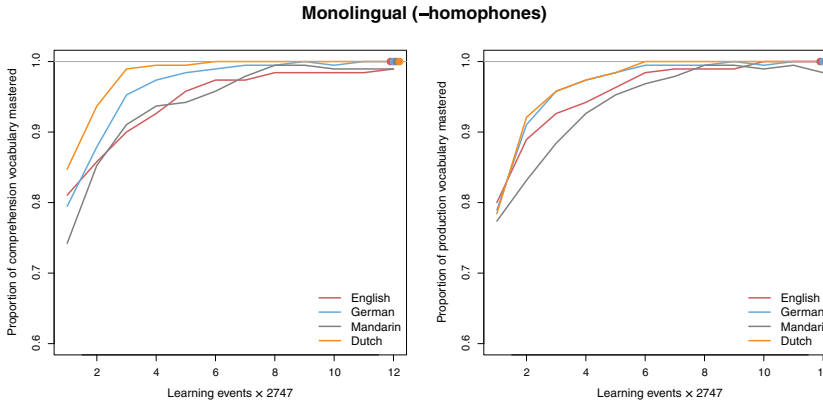

Figure 6 Vocabulary size as a function of exposure for comprehension (left) and production (right), for monolingual learning of a smaller dataset without homophones. The dots to the right of each plot indicate the model's performance at the end-state of learning. [Color figure can be viewed at wileyonlinelibrary.com]

Simulations of Bilingual Lexical Learning

Simultaneous English-German Bilinguals

In this section, we focus on lexical learning of English-German bilinguals. We first present the simulation results for simultaneous balanced bilinguals, in which the model has to learn two languages with an equal amount of input in each language right from the beginning. The model set-up is similar to that for the monolingual models, except that two language identifiers, one for English (ENG) and one for German (GER), are added to the semantic vectors (cf. Equation 1). There were a total of 73,900 learning events, with evaluations of comprehension and production at every 7,390 learning events, the same as for the monolingual simulations. Within each of these learning periods, half of the 7,390 learning events contained English words, and the other half German words, so overall the model only received half as many exposures to each language as the monolingual models did.[11]

Results are summarized in Figure 7. The lines represent the proportion of the total vocabulary of each language that the model correctly understood (left panel) or correctly produced (right panel). Bilingual comprehension resembles monolingual comprehension insofar as the number of words recognized gradually increases for both languages as learning progresses. However, due to the reduced amount of input in each language, bilingual learning progresses more slowly than monolingual learning. For production, the difference between

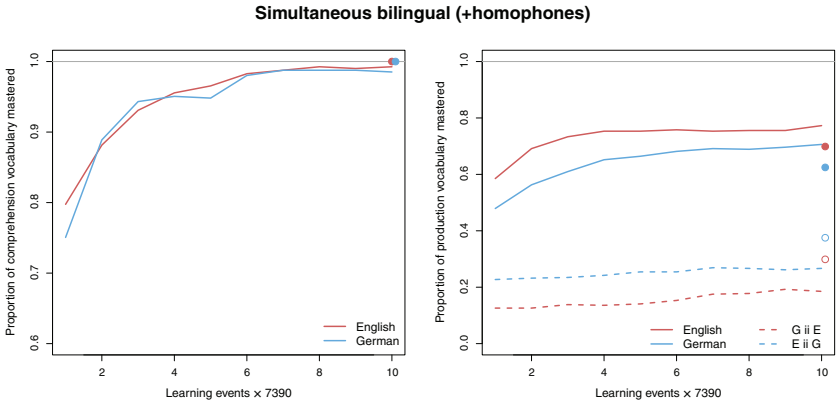

Figure 7 Vocabulary size as a function of exposure for comprehension (left) and production (right), for simultaneous balanced English-German bilinguals. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

monolingual and bilingual learning is even greater than for comprehension. In the monolingual simulations, production lags somewhat behind comprehension, but the learning curves show steady growth (right panel of Figure 4). However, in the bilingual simulation, the learning curves for production not only grow much more slowly, but they also start to plateau sooner. Furthermore, the estimates of the end-state of learning (indicated by the solid dots in the right-hand panel), indicate that ultimate achievement is even lower than that observed at the end of the simulation period, so that the curves would eventually show a downturn if the simulation period were extended long enough. A closer inspection of the production errors reveals that a great number of errors are due to "language intrusion," that is, the model produces the translation equivalent of the target form: For example, if the targeted form is English *palm*, the model selects the German form *Palme* instead. The dashed lines in Figure 7 denote the proportion of language intrusions that develop over learning. The higher dotted blue line indicates that German suffers more intrusions than English does. At the end of the simulation period (i.e., at the tenth evaluation), intrusion errors constitute about 80% and 90% of the production errors for English and German, respectively.

We observed that, in our monolingual production models, homophones are more error-prone than non-homophones, and that production errors usually

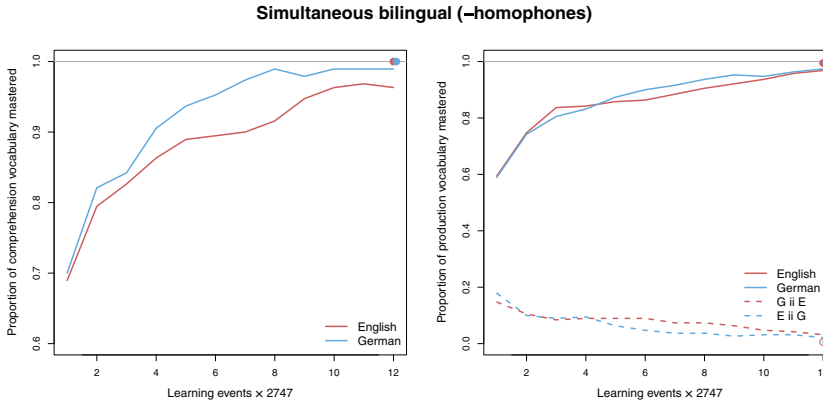

Figure 8 Vocabulary size as a function of exposure for comprehension (left) and production (right), for English-German bilinguals. The training data for this simulation is the smaller dataset without homophones. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

involve replacement of the target form by a semantically similar word. In the present bilingual situation, language intrusion is effectively a semantic error as well, given that the semantic vectors of translation equivalents differ only with respect to their language identifiers. These considerations suggest that most intrusion errors will occur for homophones, and this turns out indeed to be the case. Inspection of the tenth evaluation of production reveals that 97% of the intrusion errors for English targets involve homophones, and that all intrusion errors for German targets involve homophones. Homophones clearly render the mappings in our model more fragile. Note that the greater number of intrusions for German can now be understood as a straightforward consequence of the larger number of homophones in our model's German lexicon. To verify that this explanation is on the right track, we also carried out bilingual learning with the smaller dataset without homophones. Results are presented in Figure 8. As expected, the learning curves for production are now much more similar to those for comprehension. Intrusion errors occur mainly at the beginning, and are almost completely absent by the end of the simulation period.

The difficulty posed by homophones for our production model is perhaps unexpected given the great prevalence of homophones in natural languages. For example, a count of homophones using the CELEX lexical database (Baayen et al., 1995), restricted to monomorphemic words with three or more

phones, suggests that nearly one in five (17.5%) of English lemmas is a homophone. So although homophones may be slightly overrepresented in our bilingual dataset, there is no doubt that they are part of everyday language experience. We therefore considered whether we could make our model more robust against language intrusions for homophones. To do this, we took account of the fact that the semantics of supposed translation equivalents will actually differ in various ways. Such differences are often more subtle than a simple language-identifying feature can account for, as exemplified above, in the section on representing meaning, for the English/Dutch translation equivalents *raspberry*/*framboos* (see also Pavlenko, 2009, for a wealth of examples). In order to model such differences in usage we needed to create semantic vectors for translation equivalents that were very similar but not completely identical to one another, over and above the differences in language identifiers. This was achieved by adding a small amount of Gaussian noise to our semantic vectors: To each binary digit (0 or 1) in a word's semantic vector, we added a random number drawn from a normal distribution with a mean of 0 and standard deviation of 0.1.[12] Since this was done for each word sense independently of its translation equivalent in the other languages, the desired result was achieved. Conceptually, this implements the idea that in addition to contextual differences of language use (as represented by the language identifiers ENG and GER in the semantic vectors), words have some fine semantic nuances that are truly both language and word specific. Thus, the TREE sense of English *palm*, thanks to the addition of a tiny bit of Gaussian noise, is now modeled as very similar to the TREE sense of German *Palme*, but not completely identical.[13] Figure 9 shows the results of simulations using the semantic vectors with added noise. It can be seen that with this amendment, production learning is very smooth and highly effective, with hardly any intrusion from the other language. In our model, small cross-language differences in meaning between translation equivalents turn out to be beneficial for the acquisition of bilingual production.

Nonbalanced Bilinguals: English as L1, German as L2

We now turn to simulations in which, during the first phase of learning, the model was exposed only to English (L1). To examine the effect of L2 onset time, we ran two different simulations in which German (L2) was introduced at different stages. For the simulation of late bilinguals, the onset of L2 learning was after the fifth evaluation, that is, after 36,950 learning events of only English words. For the simulation of early bilinguals, L2 learning started after the second evaluation, that is, after 14,780 learning events of only English words.[14] For these two simulations, during the bilingual learning phase,

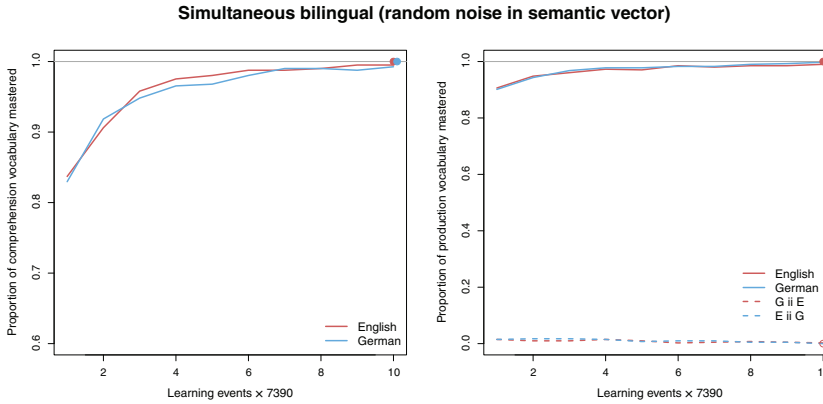

Figure 9 Vocabulary size as a function of exposure for comprehension (left) and production (right), for English-German bilinguals. In this simulation, a small amount of random noise was added to the semantic vector of each individual word. The dots to the right of each plot indicate model performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

learning events were evenly distributed across languages, such that equal numbers of German and English word tokens were encountered after L2 onset.

The first four panels of Figure 10 present the proportion of vocabulary mastered as learning unfolds. Left panels show the development of comprehension, while right panels show the development of production. The upper panels show the learning curves when L2 is encountered after the fifth evaluation, whereas the center panels show development when L2 learning begins after the second evaluation. Since the learning of German starts later than the learning of English in these simulations, the model unsurprisingly understands and produces fewer German words than English words. Also as expected, by the end of the simulation period, the late bilingual model has mastered fewer German words than the early bilingual model. For L2 comprehension, the early onset model actually approaches L1-like accuracy by the end of the simulation period. However, the situation is different for production. Not only is production accuracy lower than comprehension accuracy for L2, but the entry of the second language into the system also leads to a slight reduction in production accuracy for L1. This loss of production accuracy is counterbalanced by the intrusion errors, of which we find more for the L2, German, than for the L1, English. By summing the y coordinates for the unbroken and dotted lines of each color in the right-hand panels of Figure 10, it can be seen that, by the

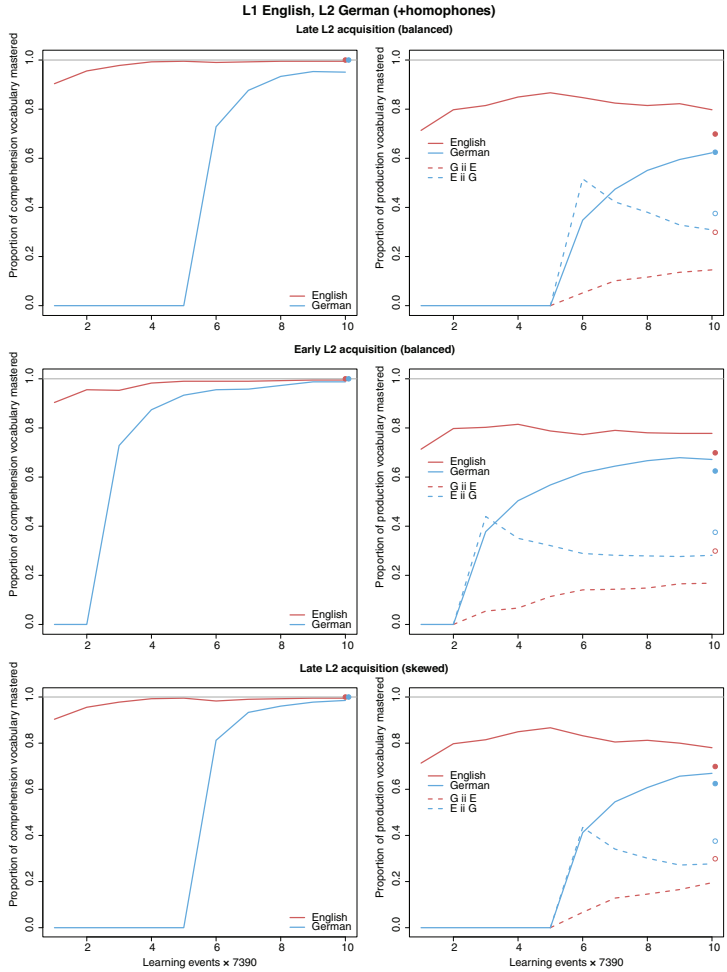

Figure 10 Vocabulary size as a function of exposure for comprehension (left) and production (right), for L1-English L2-German bilinguals. For the upper panels, L2 learning starts after the fifth evaluation, whereas for the middle panels, L2 learning starts earlier, after the second evaluation. The lower panels show results with unequal amounts of L1 and L2 input after L2 onset: One quarter English and three quarters German. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

end of the simulation period, the model's problem is not so much finding a proper word for a given meaning, but rather selecting the word form from the proper language.

The bottom panels of Figure 10 show how learning proceeds when, after L2 onset, 75% of word tokens are German and only 25% of the words are English.[15] More intensive use boosts L2 learning for comprehension, as can be seen by comparing the top and bottom left panels. For production, the main advantage appears to be a small reduction in intrusion errors from English into German, as can be seen by comparing the blue dashed lines in the top and bottom right panels. Most intrusion errors are again found when the target is a homophone. When the model is trained on the dataset without homophones (Figure 11), English rarely suffers from intrusion, regardless of the onset time of L2 and the amount of L2 input. This is because, in the absence of homophones, strong mappings of meanings to L1 forms are established before L2 learning starts. When L2 is introduced, the absence of L2 homophones means that strong mappings can also be established from meanings to L2 forms. Thus, in the absences of homophones, intrusion errors into L2 occur primarily at the beginning of learning, after which the amount of intrusion tapers off.

Simulations of Trilingual Lexical Learning

In the preceding section, we have shown that a computational model of the Discriminative Lexicon, previously only applied to monolingual learning, can be extended to bilingual learning while maintaining high levels of accuracy for both comprehension and production. In this section, we extend the model to include vocabulary learning in three languages. We believe this is the first attempt to computationally model trilingual lexical acquisition, since the most prominent current models of L3 acquisition focus on syntax and are in any case not yet computationally implemented, for example, the Typological Primacy Model (TPM; Rothman, 2015), the Cumulative Enhancement Model (CEM; Berkes & Flynn, 2012; Flynn, Foley, & Vinnitskaya, 2004), and the Linguistic Proximity Model (LPM; Westergaard, Mitrofanova, Mykhaylyk, & Rodina, 2017).

One question of interest is whether learning a third language is qualitatively different from learning a second language. Possibly, if the system has already been stressed by learning a second language, then learning a third language will be substantially more difficult. On the other hand, how the system adapts to a third language might depend on the typological properties of that language. To address these questions, we present simulations in which either Mandarin or Dutch is added as L3 to a model trained on English as L1 and

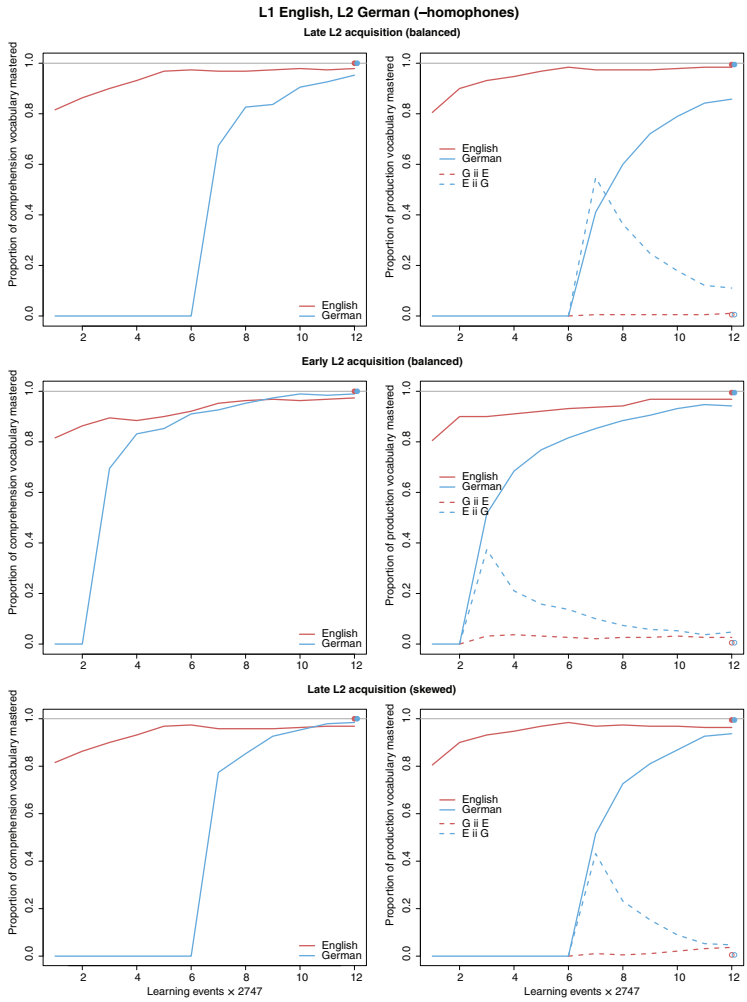

Figure 11 Vocabulary size as a function of exposure for comprehension (left) and production (right), for L1-English L2-German bilinguals. The simulations were run with the smaller dataset without homophones. For the upper panels, L2 learning starts after the fifth evaluation, whereas for the middle panels, L2 learning starts earlier, after the second evaluation. The lower panels show results with unequal amounts of L1 and L2 input: One quarter of English and three quarters of German. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

German as L2. Given that Mandarin is a tone language, the subsequent section introduces how the model incorporates suprasegmental information into lexical learning. Finally, we switch the learning order of German and Mandarin, forming a trilingual situation with L1-English, L2-Mandarin and L3-German. This simulation helps clarify to what extent our results can be generalized to the learning of different language combinations.

Late Trilingual Learning: Mandarin and Dutch as L3

The phonotactics of English, German and Dutch all allow complex syllable structures, including consonant clusters in both onset and coda, for example, English *splash, adjuncts*; German *Sprache, Herbst*; Dutch *spraak, herfst*. Unsurprisingly, therefore, the Dutch words in our dataset share 127 and 263 triphones with the English and German words, respectively. Fifty five of these occur in all three language samples. In contrast, Mandarin has much stronger restrictions on its syllable structure, allowing only single consonants and affricates in the onset, and only nasal consonants in the coda. As a result, the Mandarin words in our dataset share only 29 triphones with the English words and 37 triphones with the German words, and the three sets of words have only two triphones in common. Because Dutch is phonologically similar to English and German, whereas Mandarin is phonologically dissimilar, we expected to observe differences in their interactions with the other two languages.

In the following simulations, we build on the model of late-onset bilingual learning described in the previous section. Each simulation started with five learning periods of 100% L1-English, followed by five learning periods of 50% L1-English and 50% L2-German. After this, either L3-Mandarin or L3-Dutch was added, and learning continued for a further 10 learning periods, with each language contributing one third of the word tokens.[16]

Figure 12 presents the development of lexical learning when the third language is Mandarin. Figure 13 presents the corresponding developmental curves for L3-Dutch. The highly similar patterns of acquisition for the two third languages, despite their being so phonologically different, was contrary to our expectations. For comprehension, the learning curves of Mandarin and Dutch rise steadily, and gradually approximate those of English and German, with Dutch acquired slightly faster than Mandarin. Comprehension accuracy at the end-state of learning (indicated by the dots to the right of the plots) is virtually identical for all languages. With respect to production, both Mandarin and Dutch are learned effectively. By the end of the simulation period, the proportion of L3 vocabulary produced correctly, exceeds the accuracy levels for both L2-German and L1-English. Recall that for bilingual learning, the

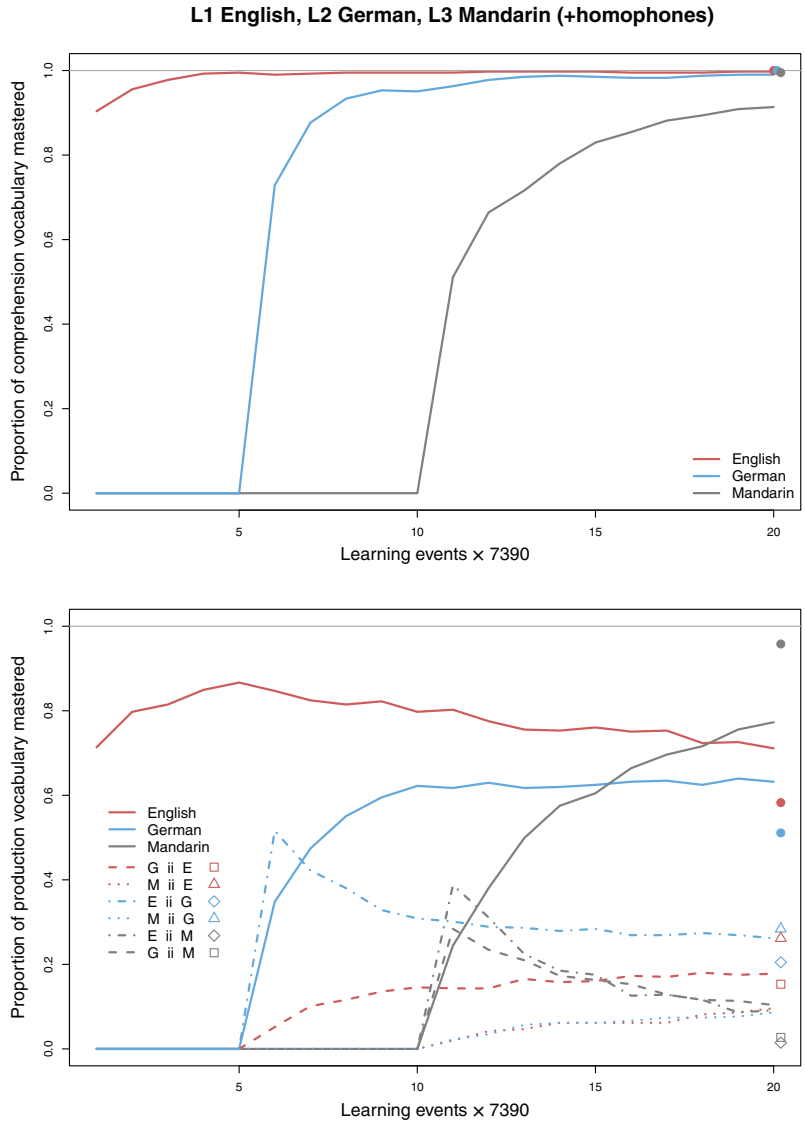

Figure 12 Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Mandarin trilinguals. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the bottom panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

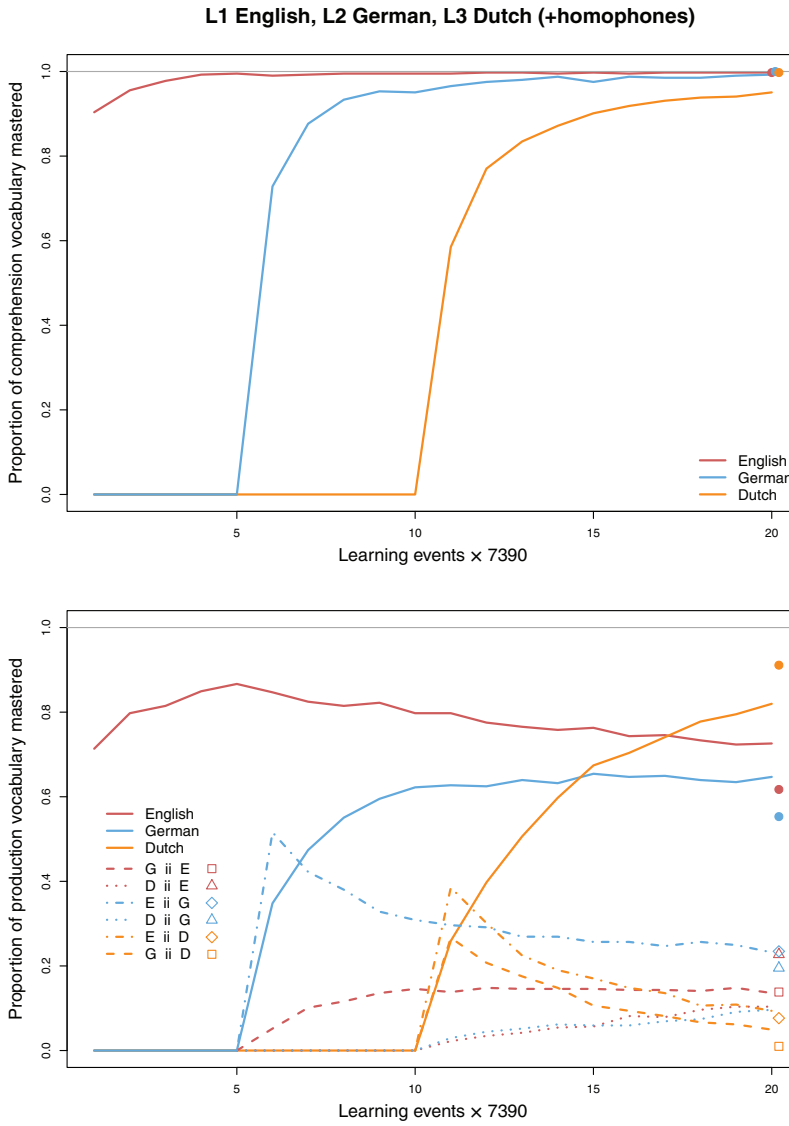

Figure 13 Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Dutch trilinguals. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

presence of homophones introduces frailty into the system and renders it vulnerable to language intrusion. The same holds for trilingual learning. In our dataset, Dutch and especially Mandarin have fewer homophones than English or German (cf. Table 5). Consequently, the two L3 languages suffer less from intrusion and therefore attain higher levels of production accuracy. Conversely, the vulnerability of English and German, already reflected in high intrusion rates by the end of the bilingual phase, continues during trilingual learning. However, if homophones are excluded by using the reduced, homophone-free dataset, intrusion into English and German virtually disappears by the end of the simulation period, as shown in the lower panels of Figure 14 for Mandarin and Figure 15 for Dutch.

If learning were to continue indefinitely, production accuracy would ultimately be higher for Mandarin (96%) than for Dutch (91%), as indicated by the gray and orange dots in the lower panels of Figure 12 and Figure 13, respectively, even though Dutch appears to approach its final accuracy level slightly more quickly than Mandarin. These subtle differences are likely to be due to the many triphones that are unique to Mandarin in our dataset, which enable the system to find, in the limit of experience, a solution that is more accurate than is possible for Dutch.

Interestingly, when homophones are included in the dataset, the balance of intrusions into L3 differs according to which L3 is being learned. L3-Mandarin suffers roughly equal numbers of intrusions from L1-English and L2-German. Dutch, on the other hand, suffers more intrusions from L1-English than from L2-German, that is, more Dutch target words are pronounced as their English equivalents than are pronounced as their German equivalents. To see why this is the case, consider Figure 16: The boxplots represent the distributions of the amount of support from the semantic system to the triphones of the words output by the production model at the end of the simulation period. These distributions are presented for both L3 Mandarin (left) and L3 Dutch (right). For each language, the distributions are visualized separately for words produced in the correct language (Mandarin or Dutch) and for words incorrectly produced in either English or German. It can be seen that, on average, with the exception of English intrusions into Dutch, the triphones of intruding words always receive less semantic support than the triphones of L3 words produced correctly. Since, for intrusion to occur, the triphones of the intruder must receive greater support than the triphones of the target form, it follows that in cases of intrusion the target forms must normally receive exceptionally low support. What is unclear, is why English intrusions into Dutch should be an exception in this respect, although contributing factors may be the numbers of

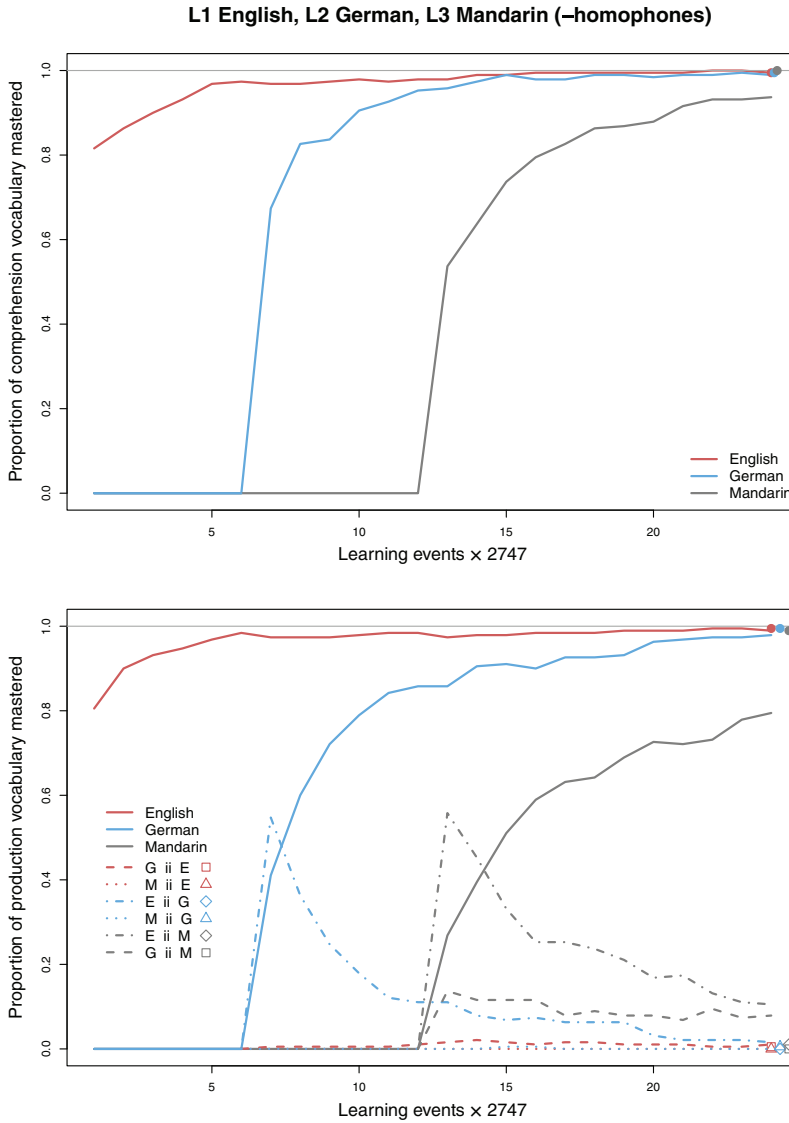

Figure 14 Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Mandarin trilinguals. The simulations were run with the smaller dataset without homophones. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the bottom panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

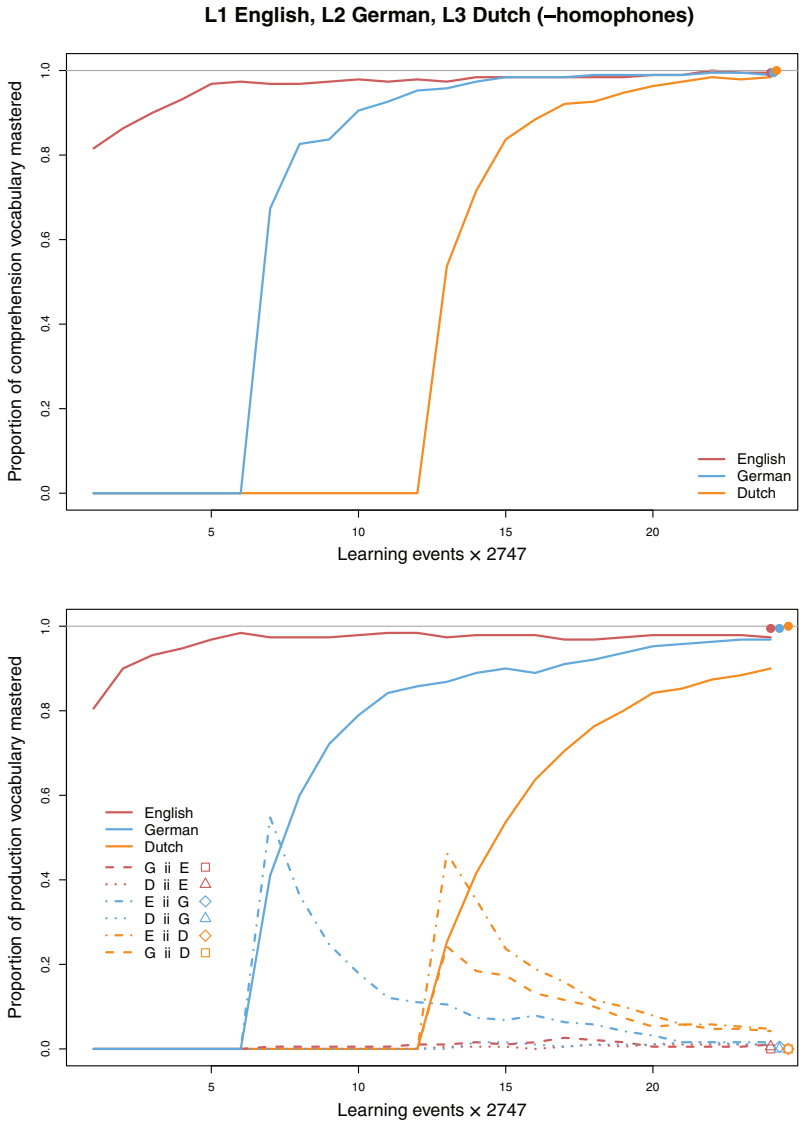

Figure 15 Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Dutch trilinguals. The simulations were run with the smaller dataset without homophones. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

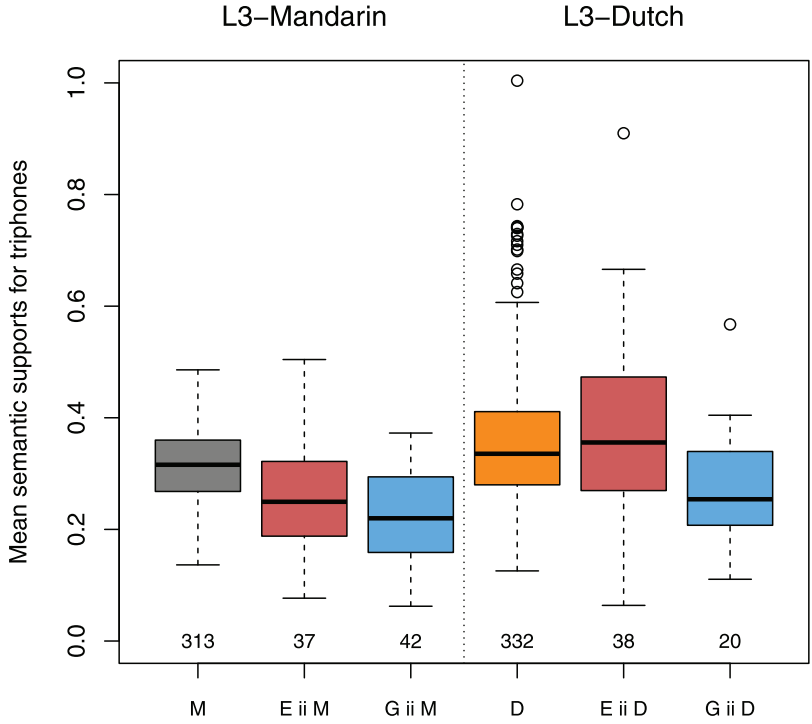

Figure 16 Boxplots for the distributions of the amount of support received by words' triphones from the semantics, for Mandarin (left) and Dutch (right) calculated at the end of the simulation period. The first boxplots of the two panels are for words without language intrusion, which are correctly produced (M and D). The second and third boxplots are for words with language intrusion from English (E ii M/D) and from German (G ii M/D), respectively. At the bottom the numbers of words in each condition are indicated. [Color figure can be viewed at wileyonlinelibrary.com]

triphones shared between the different languages and the different numbers of homophones in each language.

Learning Suprasegmental Features

In the trilingual simulations thus far, we have ignored one important difference between English, German, and Dutch on the one hand, and Mandarin on the other, namely, that Mandarin is a tone language. Mandarin is far from exceptional in this respect; in fact, 60%–70% of the world's languages are thought to be tonal, that is, to use pitch to signify lexical contrasts (Yip, 2002). The ability to handle such suprasegmental information is therefore an

Table 6 Tonal representations of Mandarin

Tone	Description	Representations	Examples
Tone 1	High-level	H	mā *'mother'*
Tone 2	High-rising	LH	má *'hemp'*
Tone 3	Low dipping	L	mǎ *'horse'*
Tone 4	High-falling	HL	mà *'scorn'*

essential prerequisite for any model of the lexicon that aspires to be capable of generalizing across languages. In what follows, we present a first proposal about how tone can be incorporated into the framework of LDL, and use this to explore the extent to which the tone system of Mandarin, as opposed to the intonational systems of the Germanic languages, influences lexical learning.

Mandarin has four lexical tones, termed high-level, high-rising, low-dipping, and high-falling (Chao, 1968). Each of these lexical tones has a distinct pitch contour pattern that can be described in terms of movement, or lack of movement, between high (H) and low (L) pitch. In the simulation experiment reported below, we therefore represented tones 1, 2, and 4 as H, LH, and HL, respectively. Tone 3, though prescriptively defined as a dipping tone, has a free variant, low-falling (Chao, 1968), that is often taken to be a low tone (Shih, 1997). We therefore chose to represent it with a single L (see Table 6).

English and German are not tone languages and therefore do not have lexical tones. However, these languages do use intonation to express different syntactic or pragmatic meanings, such as signalling a question, or surprise. In both languages, the neutral declarative intonation is characterized by a falling contour (Bolinger, 1989; Grice, Baumann, & Benzmüller, 2005), which can be formalized using ToBI notation as a high pitch accent (H*), followed by a low boundary tone (L%) (Beckman & Ayers, 1997; Grice et al., 2005; Pierrehumbert & Hirschberg, 1990). In the present study, which is limited to simulating the processing of single words without contexts, we assigned this neutral statement intonation to all the English and German words. For ease of implementation, we omit the non-alphabet characters from the ToBI notations, representing the pitch accent as "H" and the boundary tone as "L."

Although we assigned the same declarative intonation pattern to all the English and German words in our dataset, the details of its realization can actually be much more variable than the falling tone in Mandarin, depending on the position of the stressed syllable in a Germanic word. This is because, in Mandarin, every syllable has its own tone, whereas in the Germanic languages,

Table 7 Tonal representations of the falling tonal pattern for English and German

Representations	English examples	German examples
HL	***bark***, ***or****gan*	***Tee*** 'tea,' ***A****ffe* 'monkey'
H–L	***cu****stomer*	***an****rufen* 'to call'
–HL	*pi****a****no, bou****quet***	*Kar****to****ffel* 'potato,' *Vio****lett*** 'violet'
–H–L	*elec****tri****city*	***Oli****venbaum* 'olive tree'

Note. Stressed syllables are in bold.

the contour of a single accent can extend across several syllables. Consider the English words *bark*, *organ,* and *customer*, uttered with declarative intonation. Because *bark* is monosyllablic, both the pitch accent (H) and the boundary tone (L) must occur on the same syllable. Bisyllabic *organ*, has the pitch accent (H) on the stressed first syllable, immediately followed by the boundary tone (L) on the second syllable. Trisyllabic *customer*, on the other hand, also starts with the pitch accent (H), but this needs to extend across the second syllable before getting to the boundary tone (L) on the third syllable. Now consider the word *piano*. Similar to *organ*, the pitch accent (H) of *piano* is immediately followed by the boundary tone (L). But unlike *organ*, *piano* has an unstressed syllable before the pitch accent (H), which now falls on the second syllable. To take these pitch patterns into account, we used the annotation "–" to indicate the presence of one or more unstressed syllables either before the pitch accent or between the pitch accent and the boundary tone (see Table 7).

Given these representations for the pitch patterns found in English, German and Mandarin, we next added the pertinent suprasegmental features to our model. Similar to the system for phones, we used "tritones," sequences of three tonal targets, such as #HL and H–L, as inputs, and we added these tritones to the word form vectors. For the Mandarin word *ji*^{4}*hua*^{4} "plan," which has two falling tones, the tritones are #HL, HLH, LHL, and HL#. Returning to our example of English *palm* (sense: HAND) and its German counterpart *Handfläche*, we have as pitch contour patterns HL and H–L, respectively. With tritones included, the form vectors of these words are now as shown in Equation 7:

$$\boldsymbol{C} = \begin{array}{c} \\ palm \\ Handfläche \end{array} \begin{array}{c} \begin{array}{cccccccccccccccccc} \#p, & p,m & ,m\# & \#h\& & h\&n & \&nt & ntf & tfl & flE & lEx & Ex@ & \ldots & \#HL & HL\# & \#H– & H–L & –L\# \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \ldots & 0 & 0 & 1 & 1 & 1 \end{pmatrix}. \end{array} \tag{7}$$

The number of different tonal patterns for Mandarin was 41, whereas that for English and German was 4. The much larger number for Mandarin is a direct consequence of every syllable having its own tone in that language.

In the simulation experiment, we used the same semantic vectors as in the preceding simulations. To assess production accuracy, we had to modify the algorithm that constructs a legal triphone sequence from the unordered set of semantically well-supported triphones. As it would not make sense to include tritones as part of a triphone path, we applied the path-searching algorithm separately to the triphones and to the tritones. In this way, we obtained two lists of partial candidate forms, one for phones and the other for tones. A list of all complete candidate forms was obtained by considering all possible pairs of a phone candidate on the one hand and a tone candidate on the other hand. For example, for the English word *palm* (sense: HAND), the candidate phone forms could be /p,m/ and /h&ntflEx@/, and the candidate tone forms "HL" and "H–L." In this case, the set of full candidate forms, comprising both phones and tones, has the elements /p,m/$_{\mathrm{HL}}$, /p,m/$_{\mathrm{H-L}}$, /h&ntflEx@/$_{HL}$, and /h&ntflEx@/$_{H-L}$. The predicted form was selected from this set, using synthesis-by-analysis.

Comprehension and production accuracies for the simulation including suprasegmental features, using the full dataset with homophones, are shown in Figure 17. When phones and suprasegmental features are learned together, the patterns of vocabulary growth for comprehension (upper panel) are very similar to those when only phones are taken into account (cf. Figure 12). In production (lower panel), the overall patterns for English and German are also little changed, although there is a slight overall drop in accuracy of about 5% and 4%, respectively. In contrast, Mandarin production suffers to a much larger extent from the requirement to learn suprasegmental information. Although productive vocabulary in Mandarin gradually and steadily increases, it is apparent that the rate of learning is slowed down compared to the other two languages, even when we take into account that exposure to Mandarin is less than initial exposure to German (1/3 and 1/2 of tokens, respectively).

The difficulty of learning Mandarin in this simulation is obviously due to the addition of tonal features. The inclusion of more features is, apparently, not harmful to comprehension, as comprehension accuracy develops in a very similar way irrespective of whether suprasegmental information is included in the words' form representations. However, the tonal features render production more demanding, since for a word to be produced correctly, both the phones and the tonal pattern have to be correct. Given that there are more tonal patterns for Mandarin than for English and German (41 *vs*. 4), learning Mandarin

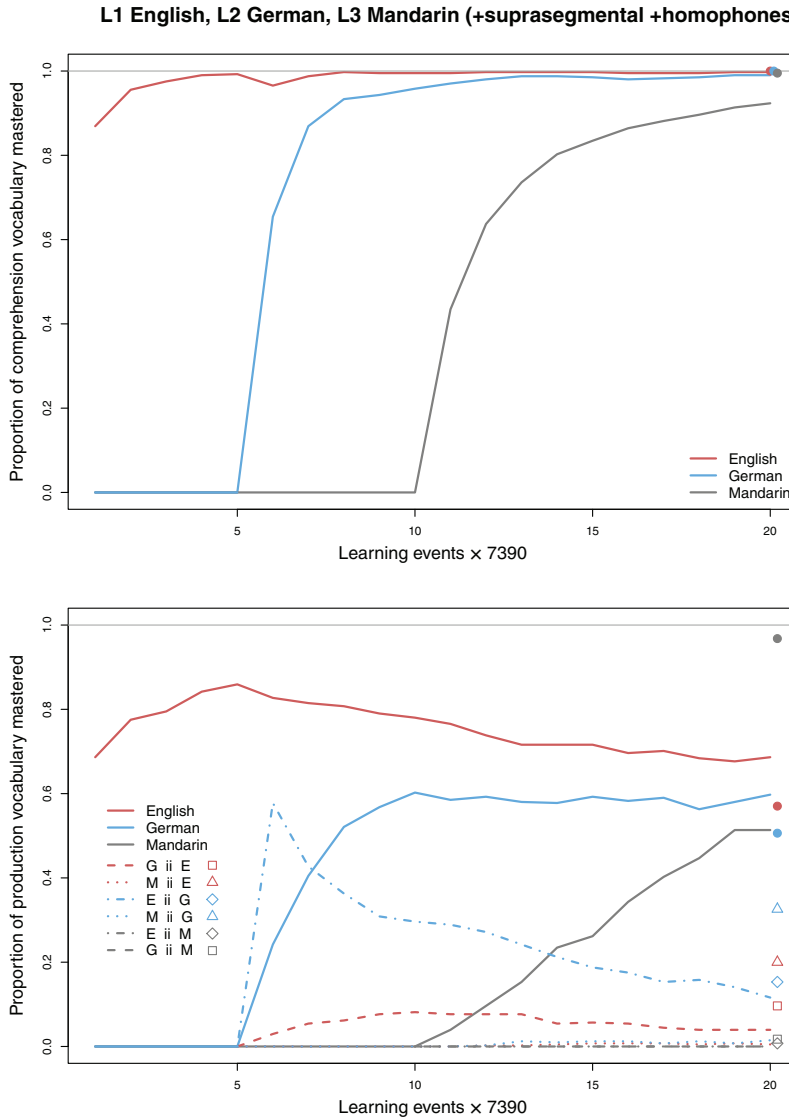

Figure 17 Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Mandarin trilinguals. In this simulation, words' form representations included suprasegmental information. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

is simply more difficult as there are more ways in which things can go wrong. At the end of the simulation period, the model predicts the wrong tonal patterns for 46% of the Mandarin words, whereas less than 2% and 8% of the English and German words respectively still have incorrect prosodic predictions. Interestingly, for Mandarin, the "HL" tone is particularly error-prone (21 out of 28 "HL" words are wrongly predicted). This tone is the equivalent of the prosodic pattern assigned to the majority of English and German words in the dataset, and at the end of the simulation period it has therefore established stronger associations with English and German than with Mandarin. Nevertheless, with more and more Mandarin input as learning continues, Mandarin production will eventually catch up and attain high accuracy at the end-state of learning, as indicated by the gray dot in the right-hand margin of Figure 17.

A striking difference between learning with and without tones is the lower rate of language intrusion when suprasegmental information is included in the form vectors. The lower graph of Figure 17 shows that, when L3-Mandarin is introduced into the simulation with L1-English and L2-German, including tone information in the form representations, there is negligible intrusion either from Mandarin to English and German, or vice versa. L1-English and L2-German suffer on average 17% and 14% less language intrusion in this simulation than in the phone-only simulation. However, the reduction in intrusion is counterbalanced by an increase in errors where the form produced is not a word in any of the three languages. Among these errors, one finds examples where the right phones are combined with the wrong tones, and vice versa. For example, the English word *bat* (the ANIMAL sense) is produced with the German form *Fledermaus*, but the suprasegmental pronunciation remains the English one, that is, "HL" instead of "H–L." Conversely, there are also cases where only tonal features intrude, for example, the German word *Veilchen* adopts the "H–L" prosody of the English equivalent *violet*. Interestingly, Mandarin words do not suffer language intrusion from the tonal patterns of the other two languages at all. Our simulated Mandarin does suffer from many language intrusions for phones, but the tonal features of English and German are rarely adopted.

Following the same procedure, we also simulated the learning of both phones and tones for Dutch trilinguals (Figure 18). Given the similar suprasegmental features of English, German, and Dutch, the learning does not differ substantially from the learning of phones alone (Figure 13). Compared to learning L3-Mandarin with tones, the learning of L3-Dutch with suprasegmental information is initially much more rapid. However, if learning were to continue indefinitely, Mandarin would eventually be learned better than Dutch

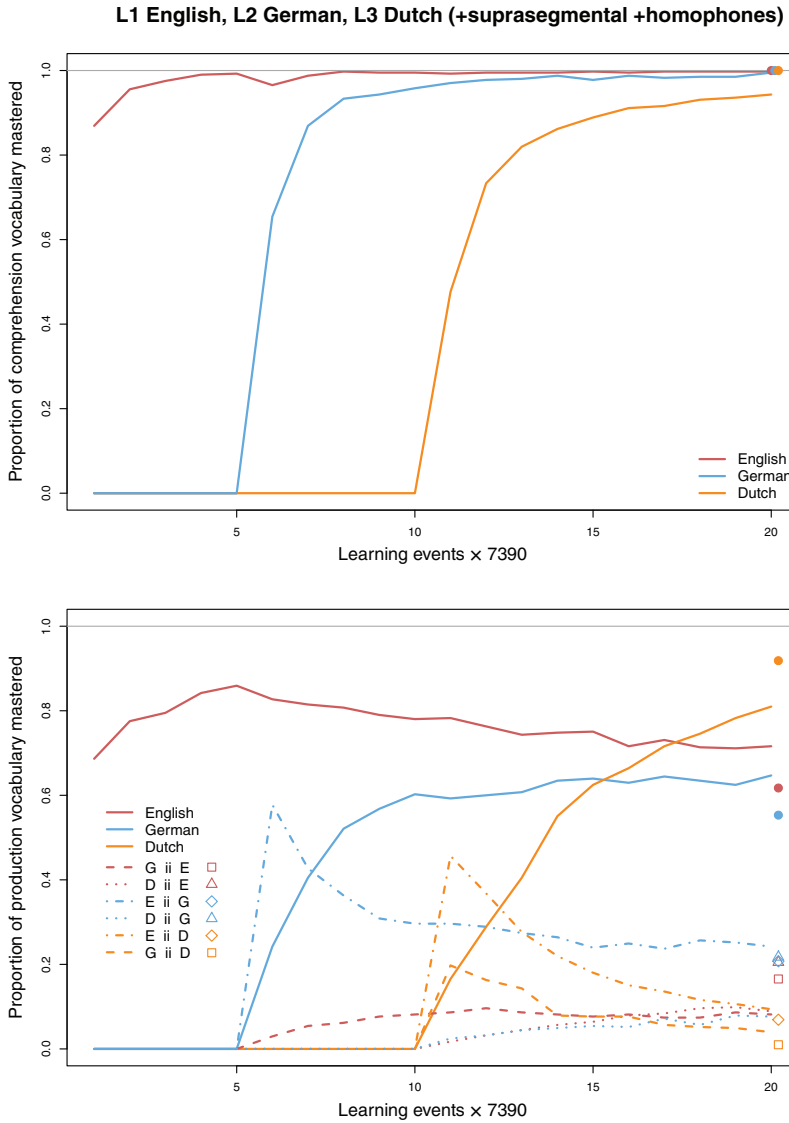

Figure 18 Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Dutch trilinguals. In this simulation, words' form representations included suprasegmental information. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

(97% *vs*. 90%), as indicated by the brown and orange dots to the right of the lower plots in Figures 17 and 18, respectively. The superior end-state production accuracy of L3-Mandarin in simulations using our full dataset results from the fact that Dutch has more homophones in this dataset than Mandarin does. We expect that with homophones excluded from the dataset, the model would ultimately achieve full production accuracy in both Mandarin and Dutch, similar to the results without suprasegmental features (lower panels of Figures 14 and 15).

Mandarin as L2 and German as L3

When homophones are included in the lexicon, irrespective of whether or not the form representations include suprasegmental information, the learning trajectories of L3-Mandarin are very different from those of L2-German, especially for production (cf. Figures 12 and 17). When such a qualitative difference in learning development is observed for real languages, this might suggest that a qualitatively different learning strategy is employed. However, in our simulations, the learning mechanism is kept constant, and we have therefore suggested that this qualitative difference must result from the different proportions of homophones in the two languages in our dataset. To explore this issue further, we ran one more simulation with Mandarin learned as L2 and German learned as L3. Tonal information was included in the form representations, and except for switching the order in which the languages were learned, all other settings remained the same. Simulation results are presented in Figure 19.

For comprehension, L2-Mandarin is learned somewhat more slowly and starts to plateau at a slightly lower level of accuracy than L2-German (upper panel, Figure 17). Changing the learning order does not change the relative comprehension learning rates for the two languages: The learning curve for L3-Mandarin also grows much more slowly than the learning curve for L3-German. With regard to production, comparing the lower panels in Figure 19 and Figure 17, we can observe different interactions between L1-English and the second language, depending on whether this is German or Mandarin. When the second language is Mandarin, by the end of the simulation period, production accuracy for L2 is slightly higher than for L1. However, when the second language is German, production accuracy for L1 slightly exceeds that for L2 at the end of the simulation period. This difference results mainly from the different amounts of intrusion suffered by the two second languages, which in turn results from the relative numbers of homophones they have in our dataset. Whereas L2-German, with a high proportion of homophones, suffers significant intrusion from English, Mandarin, with very few homophones,

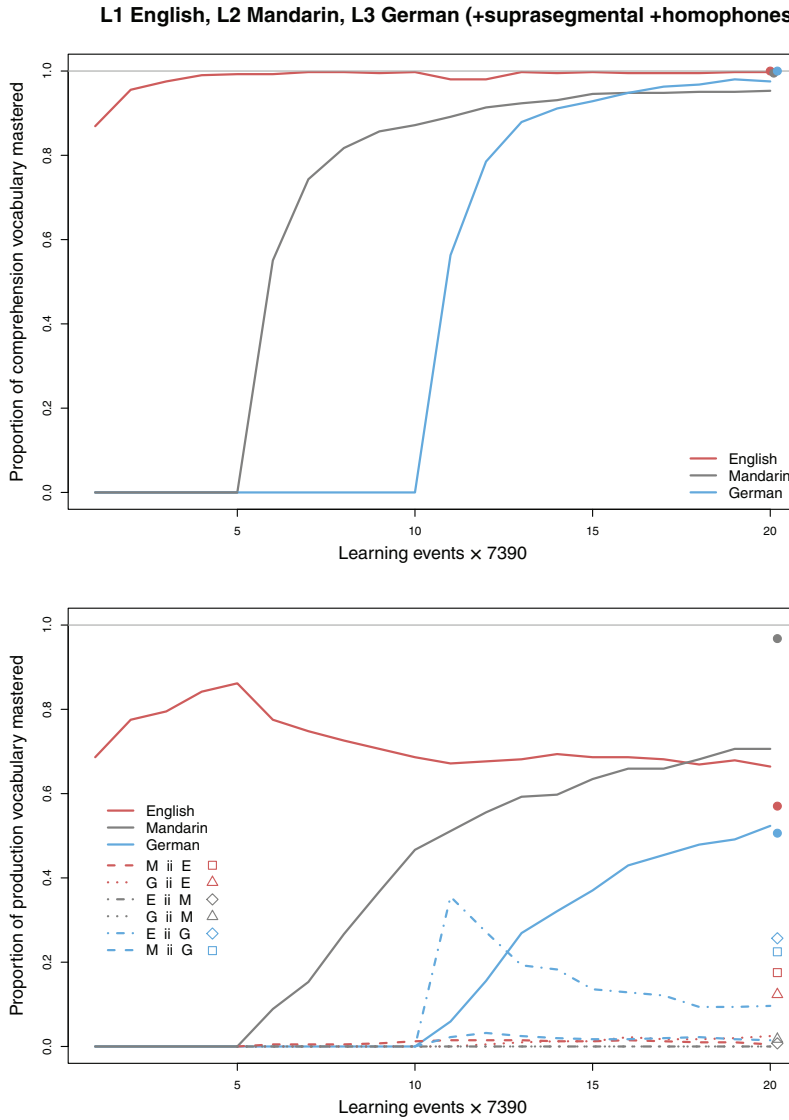

Figure 19 Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-Mandarin, L3-German trilinguals. In this simulation, words' form representations included suprasegmental information. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel "X ii Y" represent the proportion of intrusions from language X into language Y. [Color figure can be viewed at wileyonlinelibrary.com]

suffers almost no intrusion at all. Turning to L3 production, we see that the learning curve of L3-German again differs from that of L3-Mandarin: The former grows faster at the beginning and then gradually slows down, whereas the latter exhibits an almost linear trajectory. Moreover, while L3-German receives a lot of intrusion from L1-English, L3-Mandarin receives almost no intrusion from either of the other languages. Taken together, these results suggest that neither L2 learning nor L3 learning straightforwardly follows any fixed pattern. Instead, much is dependent on which language is learned at what time, and the distributional properties of these languages.

General Discussion

Is multilingualism qualitatively different from bilingualism? In this study, we addressed this question by means of a series of simulation studies implementing central concepts of the Discriminative Lexicon theory. This line of research builds on previous work by Ellis (2013) and Ellis and Larsen-Freeman (2009) on the role of discrimination learning in L2 acquisition, and work by Ramscar, Yarlett, Dye, Denny, and Thorpe (2010) and Ramscar et al. (2013) on discrimination learning in L1 acquisition.

Monolingual Learning

Our first set of simulations addressed monolingual lexical learning for translation equivalents in four languages: English, German, Dutch, and Mandarin. These studies provided a baseline for subsequent simulations with two and three languages. There were two main findings. Firstly, in the monolingual simulations, just as in human learning, production accuracy lagged behind comprehension accuracy. Secondly, within-language homophones gave rise to frailty in the mappings of form to meaning, thereby constituting an inherent weakness in the comprehension system, where this frailty gives rise to poorer approximations of the targeted semantic vectors. Nevertheless, the fact that in a context-free situation any possible meaning of a homophone has to be accepted as correct, meant that the comprehension models were very successful. For production, however, the reduction in quality in the predicted semantic vectors gave rise to semantic errors at the stage where an output is selected from a set of candidate forms. For potential consequences of uncertainty on lexical processing, see, for example, Chuang et al. (2020) and Tomaschek, Plag, Ernestus, and Baayen (2019).

According to Wilhelm von Humboldt's universal (van Marle & Koefoed, 1980), also known as the bi-uniqueness principle, an ideal language system, especially from the perspective of a language learner, should have a

one-to-one correspondence between lexical forms and lexical meanings. Violations of this principle are hypothesized to be suboptimal. For instance, Casenhiser (2005) reports experiments showing that children can have trouble learning homonyms. Although for the present monolingual simulations, learning ends up being highly accurate for both comprehension and production across non-homophones and homophones, for homophones, there is nevertheless more uncertainty in the mappings. In the earlier stages of learning, this uncertainty gives rise to errors in production, with the model producing semantic errors such as *piano* replacing *organ*. The resulting frailty is a straightforward consequence of discrimination learning, and fits well with the hypothesis that violations of the bi-uniqueness principle are suboptimal.

Bilingual Learning

The second set of simulations addressed bilingual lexical learning, with English as first language and German as second language. In these simulations, the homophone-related frailty observed for L1 learning emerged prominently as the primary source of language intrusions. The errors made by the model almost all involved homophones in one language that were inappropriately produced with a form of the other language (e.g., English *palm* being produced as German *Palme*). From the perspective of von Humboldt's one form, one meaning principle, bilingual learning in the presence of within-language homophones is confronted with two problems simultaneously. Not only does the model have to deal with specific forms that map onto two meanings, but at the same time there are also meanings that have to be associated with different forms, one for each language. Under this double stress, learning breaks down, such that even in the limit of learning, intrusions remain unavoidable.

These results were obtained under the assumption that translation equivalents have semantic representations that are identical, except for one feature specifying which language is being used. However, such extreme similarity of meaning is seldom observed for natural language, as translation equivalents participate in different collocations and idiomatic expressions (as exemplified by English *carry off the palm* versus German *auf die Palme bringen*, "to drive someone nuts"). To do justice to the translator's conundrum that "traduire c'est trahir," we added a small amount of Gaussian noise to our words' semantic vectors, so that translation pairs stayed largely similar in meaning, but now also had their own semantic idiosyncracies. With these adjusted semantic vectors, the model no longer produced language intrusions.

This result is of interest from a developmental perspective. Initially, learners of an L2 do not have sufficient experience with the L2 to absorb the

fine details of a word's collocational preferences and idiomatic usages. The semantics of the translation equivalent will be understood as identical to the meaning of the word in L1, and under these circumstances, it follows straightforwardly from learning theory (as implemented by our model) that intrusions are unavoidable. However, as a learner bootstraps into L2, they will tune in to the subtly different lexis of the L2. This in turn will allow their cognitive system to better target words' forms, resulting in a reduction in intrusions. In our simulations, this gradual development is not properly represented, as we only compared a simulation without any subtle variation in meaning against a simulation in which such variation was included from the outset. In future studies, by integrating WordNet-based semantic vectors with vectors derived from corpora using methods of distributional semantics, it may be possible to approximate better to a developmental process in which speakers gradually tune in to the lexis of the L2. In such an approach, the WordNet element of representation would reflect ontological knowledge at the cross-linguistic level, whereas the distributional element would reflect knowledge about language-specific differences in conceptualization and labelling.

Trilingual Learning

The third set of simulations started out with a bilingual lexicon with English as L1 and German as L2, and added in a third language, either Mandarin or Dutch. Comprehension accuracy developed rapidly for L3, irrespective of whether the third language was phonologically similar (Dutch) or dissimilar (Mandarin) to L1 and L2. For production, again irrespective of whether the third language was Dutch or Mandarin, accuracy for the L3 increased rapidly, with a final attainment in the limit of learning that was substantially better than that for the L1 or L2. In other words, a qualitatively different learning pattern emerged for the learning of the third language as compared to the learning of the second language. However, since in our dataset, Dutch and Mandarin have far fewer homophones (71 and 40, respectively) than English and German (more than 200 each), there was a confound between learning order and the proportion of homophones in the different languages. Hence, no firm conclusion could be drawn about the effects of learning order per se. We therefore ran another simulation, where we exchanged the learning order of German and Mandarin. When Mandarin was learned as L2, it received hardly any intrusion from either L1-English or L3-German, which is in sharp contrast to the situation for L2 German. Importantly, however, when homophones were removed from the dataset, production learning for L1-English and L2-German steadily increased, no longer showing the downward sloping trends resulting from intrusion of

L3-Mandarin when homophones were present (cf. lower panel, Figure 14). In other words, in the absence of homophones, L3 learning was very similar to L2 learning, indicating that the characteristics of the different languages in our simulations were more important than the order of learning.

In the presence of homophones, which are widespread in natural languages, the development over time of lexical acquisition may seem to be qualitatively different for an L3 as compared to an L2. However, in our simulations, nothing in the way the model learns has changed. The crucial issue is whether a homophone in L1 has a meaning that is realized by a non-homophone in the L2 or L3. If so, due to the frailty in the mapping of form to meaning for homophones, L1 production suffers. Since in our model, internal comprehension is part of production (synthesis-by-analysis), the frailty in the mapping from form to meaning has the consequence that a form of the L2 or L3 may provide a better match for the meaning input for production, in which case an intruding, "borrowed" form is selected for articulation. An example illustrating this frailty in synthesis-by-analysis is available in Appendix S1.

On Homophones

Given the vulnerability of our model to homophones, one might wonder why they are so ubiquitous in human languages. One possibility is that, just as we saw the benefits to bilingual acquisition of slight differences in the semantic representations of translation equivalents, natural language benefits from slight differences in the form representations of homophones. In fact, there is mounting evidence that homophones differ in phonetic detail (e.g., Gahl, 2008; Lohmann, 2018). Furthermore, in natural language, any ambiguity about the meaning of a homophonous form will be greatly reduced by its position in discourse as well as a wide variety of other contextual factors.

In this study, we based our lexicon on a small number of English and German homophones. However, using a larger lexicon that included a representative number of homophones from all relevant languages, would enable us to approximate even more closely to the pervasive mismatches and various kinds of partial semantic overlap that occur between languages. For instance, English *cut* (with the verbal sense that might involve an axe, knife, or scissors) has two possible translations in German (*hacken* [with an axe] and *schnitten* [with a knife or scissors]), and three in both Dutch (*hakken, knippen, snijden*) and Mandarin (*kan, qie, jian*) (see Berthele, 2012; Cook, Bassetti, Kasai, Sasaki, & Takahashi, 2006; Pavlenko, 2011; Wang & Wei, 2019, for detailed discussion of the consequences of such mismatches for language acquisition). In the approach developed in the present study, the different senses of

English *cut* could be modeled by setting up three correlated semantic vectors, one for each sense. In German, one of these vector would then be used for *hacken*, whereas *schnitten* would be coupled with two vectors. For Dutch and Mandarin, each vector would be associated with one specific word-form. This would mean that English *cut* and German *schnitten* would effectively be treated as homophones, but the different senses would have very similar semantic vectors, more similar to one another than the semantic vectors of the homophones used in the present study. We expect that the difficulties encountered by language learners in handling cross-linguistic differences in lexical overlap would then arise in our model along exactly the same lines as for the homophones in the present simulations.

What's Special About Multilingualism?

In the light of our results, several of the questions raised in the introduction can now be addressed. First, with respect to the question of whether learning a third language is qualitatively different from learning a second language, our simulations reveal that qualitative differences can indeed emerge between the pattern of learning of a second language and that of a third. However, these patterns are not consistent for all second and third languages, indicating that they do not arise from the order of acquisition per se. Rather, the differences emerge from distributional properties of the particular languages involved and how they interact with one another (in our simulations, the relative numbers of homophones). Furthermore, since our model uses exactly the same learning algorithm for all simulations, qualitative differences between patterns of acquisition cannot result from different learning mechanisms. This finding suggests that when qualitatively different patterns of acquisition are observed, it is not necessarily the case that a qualitative change in the cognitive system has taken place. In such cases, computational simulation experiments can help decide whether explanatory parsimony is justified, that is, assuming that there is no qualitative change in the system, but only a change in the input to that system.

A related issue is whether a third language can be "dormant" with respect to L1 and L2, in the sense that it doesn't interfere much with L1 and L2, and seems to be developing independently (Tytus, 2019). In our comprehension simulations, L2 and L3 are consistently dormant, but this does not hold for production. Depending on the assumptions made about the representation of meaning across languages (i.e., whether translation equivalents are represented as semantically identical), and depending on the distributional properties of the pertinent languages (e.g., the relative numbers of homophones), an L2 or L3 can be either dormant or actively interfering.

Does transfer to L3 take place only from L1, or also from L2? In our simulations, intrusions into the L3 are observed from both L1 and L2, but the amount of intrusion depends on the phonological similarity between the relevant languages, with greater similarity resulting in more intrusions and form errors. Thus, within the constraints under which our simulations were set up, there is no reason to suppose that transfer from L1 is privileged compared to transfer from L2.

A third question that can be addressed to some extent on the basis of our simulations is whether ultimate attainment of L2 and L3 is affected by the point in time at which learning the new language begins. When we define ultimate attainment as performance at the end-state of learning, with infinite experience, then the order and amount of exposure no longer matter. What does determine ultimate attainment is the system of contrasts in meaning and form, and the analogies between the two (in our model, the number of homophones in each language). Surprisingly, it appears to be the oppositions and contrasts, at a type level, that matter: In our simulations, ultimate attainment is determined by the relative numbers of L1, L2, and L3 homophones in the lexicon (i.e., types) and not by the time of onset or relative exposure (i.e., tokens). Although token frequencies influence learning in the early stages, when exposure is sent to infinity, tokens give way to types.

A fourth issue is whether there are any consequences of L3 for mastery of L1 and L2. In our simulations, at the onset of L3 learning, intrusions from L3 into L1 and L2 occur, but the rate at which this happens decreases quickly. Whether intrusion errors persist at the end-state of learning, depends on how words' semantics are represented.

Finally, are developmental trends different for comprehension and production? In our simulations, they are. Comprehension is consistently ahead of production, and errors in comprehension rapidly disappear as learning unfolds. In contrast, it is in production that we see errors arise as new languages are learned, resulting in imperfect learning that may persist even at the end-state. That comprehension is ahead of production was also observed by Chuang et al. (2019) for Estonian noun inflection, but their study only considered the end-state of learning. Here, we replicate their result, and at the same time extend it to incremental learning.

Limitations

An issue not addressed by our simulations is the possibility, pointed out by Kroll and Stewart (1994) and Kroll et al. (2010), that L2 (or L3) speakers might not proceed directly from meaning to the L2 form, but rather first retrieve

the form in L1, and then proceed to map this form onto the proper form in L2. Within the framework of discrimination learning, this learner strategy can be implemented by setting up a network $\boldsymbol{O}$ mapping L1 forms to L2 forms, resulting in a system with both a direct route and an indirect route, shown in Equations 8 and 9:

$$\boldsymbol{C} = \boldsymbol{SG}^{(L2)} \text{ (direct route)}, \tag{8}$$

$$\boldsymbol{C} = \boldsymbol{SG}^{(L1)}\boldsymbol{O} \text{ (indirect route)}. \tag{9}$$

The indirect route might be especially useful in cases where the direct route results in only very weak semantic support for words' triphones.

Another issue that we have not addressed is how task-specific effects might be accounted for within the present framework. For instance, interlingual homographs give rise to different effects in the visual lexical decision task depending on whether the task is to decide whether a word belongs to English rather than Dutch, or whether a word can belong to either language (see, e.g., Dijkstra et al., 2005). In single-language lexical decision, interlingual homographs are more difficult to respond to, whereas in generalized lexical decision, they are easier to respond to. Within the framework proposed in the present study, interlingual homographs will be close to a hyperplane in semantic space separating words of one language from the words in the other language. Proximity to the classification boundary will give rise to greater uncertainty and hence to elongated response times. In other words, although our model remains silent on the details of the decision procedures required for these tasks, the representational space is rich enough to provide these procedures with the information required for lexical decision making.

An important limitation of the present simulation experiments is that they are based on a small lexicon, with more intensive use of L2 and L3 than is typically the case for L2 and L3 learning in common learning situations in Western societies, and all this under perfect learning conditions. However, the shortcomings of this pilot study can be addressed. Models based on the Discriminative Lexicon scale up well to large datasets, and such datasets would make it possible to examine the predictions of the model in more detail. For example, does the model correctly predict that nouns are more prone to intrusion than verbs (Marian & Kaushanskaya, 2007)? Different learning situations could be modeled. For instance, the amount of L2 and L3 input could be brought down to that typical for second language learning in high-school settings. It would also be possible to model individual differences between language learners.

For example, the ease with which additional languages are picked up could be brought into the model by varying the learning rate of the Rescorla-Wagner and Widrow-Hoff learning rules. Or the adverse consequences of ADHD for learning could be modeled by injecting some error into the learning process.

Strengths and Outlook

We note here that computationally, our study offers two innovations to the theory of the Discriminative Lexicon, as developed in Baayen et al. (2019) and Chuang et al. (2019). First, whereas in previous work, the focus was on the end-state of learning, in the present study, we have demonstrated the potential of the learning rule of Widrow and Hoff (1960) to study the trajectory of learning (see Milin et al., 2020, for this learning rule, related learning strategies, and efficient implementation). Previous work on discrimination learning in second language acquisition (e.g., Ellis, 2013; Ellis, 2002; Ellis & Larsen-Freeman, 2009) has focused on the learning rule of Rescorla and Wagner (Rescorla & Wagner, 1972). Although their learning rule figures prominently in the naive discriminative learning model (Baayen et al., 2011; Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017), its dependence on one-hot encoded monadic meaning representations renders it unsuitable for exploring the effects of within and between-language similarities in meaning. Here, we have found that the Widrow-Hoff learning rule, which is mathematically related to the Rescorla-Wagner rule, provides us with a promising tool for modeling incremental learning with semantic vectors. A second contribution of the present study to the computational framework of the Discriminative Lexicon is the implementation of algorithms that take suprasegmental information into account. Validation of these algorithms awaits further experimental research in which the predictions of the model are pitted against human behavior.

The present explicit computational model has been developed in the hope that it will turn out to be a useful tool enabling precise clarification of the consequences of theoretical assumptions about lexical representation and lexical learning, and for generating quantitative predictions. Specifically, the observed frailty induced by within-language homophones and its consequences for the model's performance in speech production in L2 and L3 generates predictions about processing times and accuracies that can be pitted against human second- and third-language performance.

Material Exemption Statement

This study reports computer simulations with selections of words taken from common openly accessible resources. Hence, no further statements about the materials and privacy protection can be made.

Final revised version accepted 14 July 2020

Acknowledgment

This research was funded by an ERC advanced Grant (WIDE, no. 742545) to R. Harald Baayen.

Open access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Notes

1 But see van Geffen (2019) for an efficient algorithmic solution implemented in Multilink that applies lateral inhibition only to a small shortlist of most relevant competitors.

2 To understand what we mean by this, image a cube that contains all possible word meanings. The position of each word in this cube could be identified by a vector of the form (x, y, z), where x, y, and z are the coordinates of the relevant point in the cube. This is a metaphor for the way our semantic representations work, except that we use vectors containing many more than three values. Our vectors therefore represent points in a space with more than three dimensions (which cannot easily be visualized).

3 For a different approach to morphology that builds on linguistic domain knowledge see Baayen et al. (2019) and Chuang, Lõo, Blevins, and Baayen (2019).

4 For vision, more fine-grained receptive field features can be provided by Histograms of Oriented Gradients (HOG) features (Dalal & Triggs, 2005; Linke et al., 2017). For auditory comprehension, low-level detectors representing auditory receptive fields can be provided by frequency band features (Arnold et al., 2017). Previous studies have shown for real spontaneous conversational speech that discriminative learning works surprisingly well. On isolated word recognition, our models outperform deep learning models (Shafaei Bajestan & Baayen, 2018). In the present study, we did not make use of these more fine-grained features for reasons of interpretational simplicity.

5 The magnitude of weight changes is also determined by the learning rate, which is held constant at 0.01 in all of our simulations and thus will not be further considered here.

6 Baayen et al. (2018) derived this name from analysis by synthesis, referring to a process of comprehension through internal production (Halle & Stevens, 1962).

7 By "non-homophones" we mean that these words were only assigned one meaning in our dataset, although in actual use they might well have several senses.
8 Since in this study we do not consider the situation of quadrilinguals, mean word length was calculated across three languages. Since mean word length for English, German, and Mandarin is strongly correlated with mean word length for English, German, and Dutch ($r = 0.8$), we maintained the same frequency-sense assignments for both sets of trilingual simulation studies.
9 At the first evaluation, 94% of the words had been encountered, and by the fourth evaluation, the model had been presented with every word in the dataset at least once.
10 This pair of words would not have been homophones if we had also included tone information in our model.
11 At the first evaluation, all but 4.4% of the total vocabulary (18 English and 18 German words) had been encountered. By the fifth evaluation, the model had been presented with all the English words and all but one German word.
12 Thus the semantic vector (0, 0, 1, 1, 0), for example, might become (0.001, 0.0005, 1.002. 0.099, −0.001).
13 To motivate this kind of distinction, we note that German *Palme* is used in expressions such as *Das brachte mich auf die Palme*, meaning "that drove me nuts," whereas *palm* in English is used in expressions such as *to carry off the palm*, meaning "to be judged the best of all."
14 In both simulations, all the English words had been encountered before the onset of L2 learning, and 95.6% of German words had also been encountered by the first evaluation after L2 onset. For early bilinguals, all German words had been encountered by the ninth evaluation, while for late bilinguals, all but one German word had been encountered at least once by the end of the simulation period.
15 At the first evaluation after L2 onset, all but three German words had been presented to the model at least once. All German words had been encountered at the end of learning.
16 At the first evaluation after the introduction of L3, about 91% of the L3 words had been encountered; by the fourth evaluation after the introduction of L3, all L3 words had been encountered at least once.

References

Aertsen, A. M. H. J., & Johannesma, P. I. M. (1981). The spectro-temporal receptive field: A functional characteristic of auditory neurons. *Biological Cybernetics*, *42*, 133–143. https://doi.org/10.1007/bf00336731

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*, 147–149. https://doi.org/10.3758/bf03334492

Arnold, D., Tomaschek, F., Lopez, F., Sering, T., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE*, *12*, e0174623. https://doi.org/10.1371/journal.pone.0174623

Baayen, R. H., Chuang, Y.-Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, *13*, 232–270. https://doi.org/10.1075/ml.18010.baa

Baayen, R. H., Chuang, Y.-Y., & Heitmeier, M. (2019). *WpmWithLdl: Implementation of Word and Paradigm Morphology with Linear Discriminative Learning*. R package version 1.4.6. Retrieved from https://osf.io/xq92s/

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (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*, 1–39. https://doi.org/10.1155/2019/4895891

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–482. https://doi.org/10.1037/a0023851

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Bardel, C., & Falk, Y. (2012). The L2 status factor and the declarative/procedural distinction. In J. Cabrelli, S. Flynn, & J. Rothman (Eds.), *Third language acquisition in adulthood* (pp. 61–78). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/sibil.46.06bar

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. The Ohio State University Research Foundation. Retrieved from http://www.cs.columbia.edu/~agus/tobi/labelling_guide_v3.pdf

Berkes, É., & Flynn, S. (2012). Further evidence in support of the cumulative-enhancement model. In J. Cabrelli, S. Flynn, & J. Rothman (Eds.), *Third language acquisition in adulthood* (pp. 143–164). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/sibil.46.11ber

Berthele, R. (2012). On the use of PUT verbs by multilingual speakers of romansh. In A. Kopecka, & B. Narasimhan (Eds.), *Events of putting and taking: a crosslinguistic perspective* (pp. 145–166). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/tsl.100.11ber

Blair, D., & Harris, R. J. (1981). A test of interlingual interaction in comprehension by bilinguals. *Journal of Psycholinguistic Research*, *10*, 457–467. https://doi.org/10.1007/bf01067169

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1007/bf01067169

Bolinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse*. Stanford, CA: Stanford University Press.

Brysbaert, M., Verreyt, N., & Duyck, W. (2010). Models as hypothesis generators and models as roadmaps. *Bilingualism: Language and Cognition*, *13*, 383–384. https://doi.org/10.1017/s1366728910000167

Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language*, *32*, 319–343. https://doi.org/10.1017/s0305000904006749

Čavar, F., & Tytus, A. E. (2018). Moral judgement and foreign language effect: When the foreign language becomes the second language. *Journal of Multilingual and Multicultural Development*, *39*, 17–28. https://doi.org/10.1080/01434632.2017.1304397

Chao, Y.-R. (1968). *A Grammar of Spoken Chinese*. Los Angeles: University of California Press.

Chuang, Y.-Y., Lõo, K., Blevins, J. P., & Baayen, R. H. (2019). Estonian case inflection made simple. A case study in word and paradigmmorphology with linear discriminative learning. *PsyArXiv*, 1–19. https://doi.org/10.31234/osf.io/hdftz

Chuang, Y.-Y., Vollmer, M.-L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2020). The processing of nonword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*, 1–32. https://doi.org/10.3758/s13428-020-01356-w

Clark, E. V. (1993). *The lexicon in acquisition*, vol. 65. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511554377

Cook, V., Bassetti, B., Kasai, C., Sasaki, M., & Takahashi, J. A. (2006). Do bilinguals have different concepts? The case of shape and material in japanese l2 users of English. *International Journal of Bilingualism*, *10*, 137–152. https://doi.org/10.1177/13670069060100020201

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (vol. 1, pp. 886–893). https://doi.org/10.1109/cvpr.2005.177

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, *47*, 109–121. https://doi.org/10.1016/S0022-2496(02)00016-0

Davydova, J., Tytus, A. E., & Schleef, E. (2017). Acquisition of sociolinguistic awareness by German learners of English: A study in perceptions of quotative be like. *Linguistics*, *55*, 783–812. https://doi.org/10.1515/ling-2017-0011

De Groot, A. M. B. (2011). *Language and Cognition in Bilinguals and Multilinguals: An introduction*. New York, NY: Psychology Press.

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1995). Receptive-field dynamics in the central visual pathways. *Trends in Neurosciences*, *18*, 451–458. https://doi.org/10.1016/0166-2236(95)94496-r

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hoffmann, T., & Jaggi, M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *Proceedings of the 26th International World Wide Web Conference (WWW-2017)*, Perth, Australia. https://doi.org/10.1145/3038912.3052611

Dijkstra, T., Moscoso del Prado Martín, F., Schulpen, B., Schreuder, R., & Baayen, R. H. (2005). A roommate in cream: Morphological family size effects on interlingual homograph recognition. *Language and Cognitive Processes*, *20*, 7–41. https://doi.org/10.1080/01690960444000124

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*, 175–197. https://doi.org/10.1017/s1366728902003012

Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, *22*, 657–679. https://doi.org/10.1017/s1366728918000287

Du Bellay, J. (2013). *Défense et illustration de la langue française*. Paris, France: Presses électroniques de France (Original work published in 1549).

Duong, L., Kanayama, H., Ma, T., Bird, S., & Cohn, T. (2017). Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.: Volume 1, Long Papers* (pp. 894–904). https://doi.org/10.18653/v1/e17-1084

Eggermont, J., Aertsen, A., Hermes, D., & Johannesma, P. (1981). Spectro-temporal characterization of auditory neurons: Redundant or necessary? *Hearing Research*, *5*, 109–121. https://doi.org/10.1016/0378-5955(81)90030-7

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*, 1–24. https://doi.org/10.1093/applin/ami038

Ellis, N. C. (2006b). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*, 164–194. https://doi.org/10.1093/applin/aml015

Ellis, N. C. (2013). Frequency-based accounts of second language acquisition. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (p. 193). Abingdon, UK: Routledge.

Ellis, N. C., & Larsen-Freeman, D. (2009). *Language as a complex adaptive system*, vol. 11. Hoboken, NJ: Wiley.

Falk, Y., Lindqvist, C., & Bardel, C. (2015). The role of L1 explicit metalinguistic knowledge in L3 oral production at the initial state. *Bilingualism: Language and Cognition*, *18*, 227–235. https://doi.org/10.1017/s1366728913000552

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Flynn, S., Foley, C., & Vinnitskaya, I. (2004). The cumulative-enhancement model for language acquisition: Comparing adults' and children's patterns of development in first, second and third language acquisition of relative clauses. *International Journal of Multilingualism*, *1*, 3–16. https://doi.org/10.1080/14790710408668175

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*, 474–496. https://doi.org/10.1353/lan.0.0035

Grice, M., Baumann, S., & Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology. In S.-A. Jun (Ed.), *Prosodic typology* (pp. 55–83). Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199249633.003.0003

Gurney, K., Prescott, T. J., & Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, *84*, 401–410. https://doi.org/10.1007/pl00007984

Gurney, K., Prescott, T. J., & Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological Cybernetics*, *84*, 411–423. https://doi.org/10.1007/pl00007985

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. In *IRE Transactions on information theory* (vol. 8, pp. 155–159). *Institute of Electrical and Electronics Engineers (IEEE)*. https://doi.org/10.1109/tit.1962.1057686

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720. https://doi.org/10.1037/0033-295X.111.3.662

Hawkins, R., & Lozano, C. (2006). Second language acquisition of phonology, morphology and syntax. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (pp. 67–74). London, UK: Elsevier. https://doi.org/10.1016/B0-08-044854-2/00634-9

Hermas, A. (2015). The categorization of the relative complementizer phrase in third-language english: A feature re-assembly account. *International Journal of Bilingualism*, *19*, 587–607. https://doi.org/10.1177/1367006914527019

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*, 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Ingram, D. (1974). The relation between comprehension and production. In R. L. Schiefelbusch & L. L. Lloyd (Eds.), *Language perspectives—Acquisition, retardation, and intervention* (pp. 313–334). Baltimore, MD: University Park Press.

Jarema, G. (2017). Polyglossia as a personal journey. In M. Libben, M. Goral, & G. Libben (Eds.), *Bilingualism. A framework for understanding the mental lexicon* (pp. xiii–xvii). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/bpa.6.002pro

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*, 35–45. https://doi.org/10.1115/1.3662552

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*, 149–174. https://doi.org/10.1006/jmla.1994.1008

Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The revised hierarchical model: A critical review and assessment. *Bilingualism: Language and Cognition*, *13*, 373–381. https://doi.org/10.1017/s136672891000009x

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211–240. https://doi.org/10.1037/0033-295x.104.2.211

Linke, M., Broeker, F., Ramscar, M., & Baayen, R. H. (2017). Are baboons learning "orthographic" representations? Probably not. *PLOS-ONE*, *12*, e0183876. https://doi.org/10.1371/journal.pone.0183876

Lohmann, A. (2018). Cut (n) and cut (v) are not homophones: Lemma frequency affects the duration of noun–verb conversion pairs. *Journal of Linguistics*, *54*, 753–777. https://doi.org/10.1017/s0022226717000378

Maciejewski, G., & Klepousniotou, E. (2016). Relative meaning frequencies for 100 homonyms: British edom norms. *Journal of Open Psychology Data*, *4*, 1–5. https://doi.org/10.5334/jopd.28

Marian, V., & Kaushanskaya, M. (2007). Cross-linguistic transfer and borrowing in bilinguals. *Applied Psycholinguistics*, *28*, 369–390. https://doi.org/10.1017/s014271640707018x

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of the basic findings. *Psychological Review*, *88*, 375–407. https://doi.org/10.1016/b978-1-4832-1446-7.50048-0

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119). Retrieved from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PLOS One*, *12*, e0171935. https://doi.org/10.1371/journal.pone.0171935

Milin, P., Madabushi, H. T., Croucher, M., & Divjak, D. (2020). Keeping it simple: Implementation and performance of the proto-principle of adaptation and learning in the language sciences. *Arxiv Preprint Arxiv:2003.03813*, 1–26. Retrieved from https://arxiv.org/abs/2003.03813

Monaghan, P., Chang, Y.-N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *Journal of Memory and Language*, *93*, 1–21. https://doi.org/10.1016/j.jml.2016.08.003

Mosca, M. (2019). Trilinguals' language switching: A strategic and flexible account. *Quarterly Journal of Experimental Psychology*, *72*, 693–716. https://doi.org/10.1177/1747021818763537

Mosca, M., & de Bot, K. (2017). Bilingual language switching: Production vs. recognition. *Frontiers in Psychology*, *8*, 934. https://doi.org/10.3389/fpsyg.2017.00934

Mulder, K., Dijkstra, T., Schreuder, R., & Baayen, R. H. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language*, *72*, 59–84. https://doi.org/10.1016/j.jml.2013.12.004

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250. https://doi.org/10.1016/j.artint.2012.07.001

Pavlenko, A. (2009). Conceptual representation in the bilingual lexicon and second language vocabulary learning. In A. Pavlenko (Ed.), *The bilingual mental lexicon: Interdisciplinary approaches* (pp. 125–160). Bristol, UK: Multilingual Matters. https://doi.org/10.21832/9781847691262-008

Pavlenko, A. (2011). (Re-)naming the world: Word-to-referent mapping in second language speakers. In A. Pavlenko (Ed.), *Thinking and speaking in two languages* (pp. 198–236). Bristol, UK: Multilingual Matters. https://doi.org/10.21832/9781847693389-009

Pierrehumbert, J., & Hirschberg, J. B. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 271–311). Cambridge, MA: MIT press. https://doi.org/10.7551/mitpress/3839.003.0016

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, *89*, 760–793. https://doi.org/10.1353/lan.2013.0068

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*, 927–960. https://doi.org/10.1080/03640210701703576

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*, 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Redgrave, P., Prescott, T., & Gurney, K. (1999). The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, *89*, 1009–1023. https://doi.org/10.1016/s0306-4522(98)00319-4

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *American Psychologist*, *43*, 151–160. https://doi.org/10.1037/0003-066X.43.3.151

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton Century Crofts.

Rothman, J. (2015). Linguistic and cognitive motivations for the Typological Primacy Model (tpm) of third language (l3) transfer: Timing of acquisition and proficiency considered. *Bilingualism: Language and Cognition*, *18*, 179–190. https://doi.org/10.1017/s136672891300059x

Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, *65*, 569–631. https://doi.org/10.1613/jair.1.11640

Sering, K., Milin, P., & Baayen, R. H. (2018). Language comprehension as a multiple label classification problem. *Statistica Neerlandica*, 1–15. https://doi.org/10.1111/stan.12134

Sering, K., Stehwien, N., & Gao, Y. (2019). create_vtl_corpus: Synthesizing a speech corpus with vocaltractlab (version v1.0.0). https://doi.org/10.5281/zenodo.2548895

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, (vol. 2), (pp. 994–1000). IEEE. https://doi.org/10.1109/cvpr.2005.254

Shafaei Bajestan, E., & Baayen, R. H. (2018). Wide learning for auditory comprehension. In *Proceedings of Interspeech 2018* (pp. 966–970). ISCA. https://doi.org/10.21437/interspeech.2018-2420

Shih, C. (1997). Mandarin third tone sandhi and prosodic structure. In J. Wang & N. Smith (Eds.), *Studies in Chinese Phonology* (pp. 81–123). Berlin, Germany: Mouton de Gruyter. https://doi.org/10.1515/9783110822014.81

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*, 314–321. https://doi.org/10.3758/bf03210755

Smith, S. L., Turban, D. H., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *Arxiv Preprint Arxiv:1702.03859*, 1–10. Retrieved from https://arxiv.org/abs/1702.03859

Tomaschek, F., Plag, I., Ernestus, M., & Baayen, R. H. (2019). Modeling the duration of word-final s in english with naıve discriminative learning. *Journal of Linguistics*, 1–38. https://doi.org/10.1017/s0022226719000203

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Rresearch*, *37*, 141–188. https://doi.org/10.1613/jair.2934

Tytus, A. E. (2018). Rising to the bilingual challenge: Self-reported experiences of managing life with two languages. *International Journal of Bilingual Education and Bilingualism*, *21*, 207–221. https://doi.org/10.1080/13670050.2016.1153598

Tytus, A. E. (2019). Active and dormant languages in the multilingual mental lexicon. *International Journal of Multilingualism*, *16*, 357–374. https://doi.org/10.1080/14790718.2018.1502295

van Geffen, A. (2019). Reducing noise from competing neighbours: Word retrieval with lateral inhibition in multilink. (master's thesis). Radboud University, Nijmegen, The Netherlands.

van Heuven, W. J. B., & Dijkstra, T. (2010). Language comprehension in the bilingual brain: fmri and erp support for psycholinguistic models. *Brain Research Reviews*, *64*, 104–122. https://doi.org/10.1016/j.brainresrev.2010.03.002

van Marle, J., & Koefoed, G. A. T. (1980). Over humboldtiaanse taalveranderingen, morfologie en de creativiteit van taal. *Spektator*, *10*, 111–147. Retrieved from https://www.dbnl.org/tekst/marl002humb01_01/marl002humb01_01.pdf

Wang, Y., & Wei, L. (2019). Cognitive restructuring in the bilingual mind: Motion event construal in early Cantonese–English bilinguals. *Language and Cognition*, *11*, 527–554. https://doi.org/10.1017/langcog.2019.31

Westbury, C. (2014). You can't drink a word: Lexical and individual emotionality affect subjective familiarity judgments. *Journal of Psycholinguistic Research*, *43*, 631–649. https://doi.org/10.1037/e505772014-208

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2015). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *The Quarterly Journal of Experimental Psychology*, *68*, 1599–1622. https://doi.org/10.1080/17470218.2014.970204

Westergaard, M., Mitrofanova, N., Mykhaylyk, R., & Rodina, Y. (2017). Crosslinguistic influence in the acquisition of a third language: The linguistic proximity model. *International Journal of Bilingualism*, *21*, 666–682. https://doi.org/10.1177/1367006916648859

Whorf, B. L. (1953). *Language, thought and reality: Selected writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, 96–104. https://doi.org/10.21236/ad0241531

Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from dialect variation. *Journal of Phonetics*, *40*, 307–314. https://doi.org/10.1016/j.wocn.2011.12.004

Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., & Baayen, R. H. (2014). A cognitively grounded measure of pronunciation distance. *PLOS-ONE*, *9*, e75734. https://doi.org/10.1371/journal.pone.0075734

Yip, M. (2002). *Tone*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139164559

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. The effect of within-language homophones in bilinguals

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Computational Modeling Provides Insights to Linguistic Theories of Bilingual and Multilingual Learning

What This Research Was About and Why It Is Important

Despite the widespread interest in bilingualism and multilingualism, most theories and models that are put forward thus far are abstract and stay at the conceptual level. Without computational implementations, it remains unclear to what extent these theories are generalizable. The goal of this study is to introduce Linear Discriminative Learning, a computational framework grounded in the theory of discriminative learning, as a tool to study lexical learning and processing in bilinguals and trilinguals. To illustrate the application of this model, we addressed some of the common issues in the field. For example, how is the learning of a new language affected by the point at which the learning starts? Furthermore, is L3 learning qualitatively different from L2 learning? We explain how computational simulations help us answer these questions.

What the Researchers Did

- We constructed a small multilingual lexicon that contains English, German, Dutch, and Mandarin words of translation equivalents. We assumed that these translation equivalents have similar meanings and share the same frequency distribution in the respective languages. Some of these words are homophones.

- In each simulation, words from one or more languages were presented to the model, and the model was trained incrementally to comprehend and produce these words. To monitor the learning process, we traced and evaluated model performance throughout the simulation.
- Different learning scenarios were created.

What the Researchers Found

- Model performance improves with learning. The onset of L2 determines the efficiency of L2 learning to a large extent. All else being equal, the earlier L2 is learned, the faster L2 attains high accuracy. In addition, the learning trajectory of a given language differs, depending on whether it is learned as L2 or L3.
- When more than one language is learned, a large number of production errors are due to language intrusion. In general, L2 and L3 suffer a lot from L1 intrusion, but L2 and L3 intrude into L1 as well. Intrusion errors decrease as learning progresses.
- Homophones delay production learning. In multilingual settings, languages with more homophones are prone to intrusion from languages with fewer homophones.

Things to Consider

- The advantage of an early onset for L2 learning is in line with findings in the bilingual literature. However, one confounding factor that is often not considered together with the onset effect is the amount of L2 input. When L2 learning starts earlier, the amount of L2 input that a given L2 speaker receives also increases. With computational simulations, we can easily tease the two factors apart and examine the effects of individual factors.
- Although the learning trajectory of a given language appears to differ when learned as L2 and L3, this does not necessarily entail that different mechanisms are involved in L2 and L3 learning. Given that the learning algorithm in the model never changes, differences are more likely to result from the distributional properties of a given language and of the other languages. Computational modeling can help us clarify the cause of such differences.

How to cite this summary: Chuang, Y.-Y., Bell, M. J., Banke, I., & Baayen, H. R. (2020). Computational modeling provides insights to linguistic theories of bilingual and multilingual learning. *OASIS Summary* of Chuang et al. (2021) in *Language Learning*. https://oasis-database.org