Bilingual and multilingual mental lexicon: a modeling study with Linear Discriminative Learning

Yu-Ying Chuang[1], Melanie J. Bell[2], Isabelle Banke[1], R. Harald Baayen[1]

1: Eberhard-Karls University of Tübingen, Germany
2: Anglia Ruskin University, UK

Abstract

This study addresses the question of whether there is anything special about learning a third language, as compared to learning a second language, just by virtue of the third language being the third language acquired, and independently of the specific properties of the third language. We used computational modeling to explore this question for the learning of a small vocabulary of some 400 words, with English as L1, German or Mandarin as L2, and Mandarin and alternatively Dutch, as L3. For computational modeling, we made use of the mathematical framework of linear discriminative learning, which we extended with the learning rule of Widrow-Hoff to enable the modeling of incremental learning of the mappings between form and meaning when words' meanings are represented by vectors of real numbers (embeddings) rather than by abstract symbolic units. A series of simulation experiments covering single-language learning, bilingual learning, and finally trilingual learning, clarified that within the framework of discrimination learning, within-language homophones give rise to frailty in comprehension that in turn for production gives rise to semantic errors in L1, and language intrusions in L2 and L3. Our model correctly predicts production to lag behind comprehension in learning, and it clarified that, within the boundaries of discrimination learning, the properties of the L3 crucially determine whether L3 learning appears to involve a language that is 'dormant' with respect to L1 and L2. Qualitatively surprisingly different patterns of acquisition of the L3, and its interactions with L1 and L2, can arise in our simulations without any changes in the mathematics driving learning. Our simulations also show that when words' forms incorporate not only segmental but also suprasegmental information, the nature of errors that arise in production changes. In the general discussion, we reflect on the implications of our findings for the question of what is special about multilingualism.

keywords:

Bilingualism, multilingualism, mental lexicon, word comprehension, word production, Linear Discriminative Learning, Widrow-Hoff learning rule, homophony

1 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 are there any

Author note: Each author was responsible for the selection of translation equivalents and phonological transcriptions of the words in our study originating from their L1.

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

In this study, we address these high-level questions about the global system properties of bilingualism and multilingualism by means of low-level computational modeling of lexical acquisition. The computational framework that we are using is that of Naive Discriminative Learning (NDL, Baayen et al., 2011) and its twin, Linear Discriminative Learning (LDL, Baayen et al., 2019b). Both NDL and LDL implement discrimination learning, which has a long history in physics (Widrow and Hoff, 1960; Kalman, 1960), statistics (formally, LDL implements multivariate multiple regression) and psychology (Rescorla and Wagner, 1972; Rescorla, 1988; Siegel and Allan, 1996; Ellis, 2006b; Ramscar and Yarlett, 2007; Ramscar et al., 2013). 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. (2019b).

For auditory comprehension, computational models of the Discriminative Lexicon (i.e. NDL and LDL) take real speech as input (for empirical results see Arnold et al., 2017; Shafaei Bajestan and Baayen, 2018). For visual comprehension, the input can be either low-level visual features (Serre et al., 2005; Linke et al., 2017) 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 et al., 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. (2019b)). 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. (2019b).

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. Hawkins et al., 2006; Bardel and Falk, 2012). We can also explore how proficiency in production relates to proficiency in comprehension (Mosca and 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: unintentionally using a different language from the one intended (cf. Tytus, 2018; Jarema, 2017).

Simulation studies, such as those presented in this paper, 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 et al., 2017), and the role of meta-linguistic knowledge (Falk et al., 2015).

Recent theoretical models for the acquisition of the syntax of multiple languages include the Typological Primacy Model (TPM; Rothman, 2015), the Cumulative Enhancement Model (CEM; Flynn et al., 2004; Berkes and Flynn, 2012), and the Linguistic Proximity Model (LPM; Westergaard et al., 2017). According to the TPM, when a third language is encountered, at the very earliest stages both the L1 and the L2 can function as models for transfer. However, as soon as the cognitive system has identified which prior language the L3 is typologically closest to, this language will be selected as the basis for generating predictions about how the grammar of the L3 works. In contrast, according to the CEM, neither prior language plays a priviliged role during acquisition of a third language. Instead, all currently known languages influence how a new language is acquired. This influence is assumed to be gradual and cumulative, and non-facilitative transfer is excluded. The LPM also postulates that all previously learned languages will influence the acquisition of a new language, but allows for non-facilitative as well as facilitative influences. According to this theory, such influences arise when the new language is similar to a previously learned language in terms of its structural properties. The PTM, CEM, and LPM are all formulated at high levels of theoretical abstraction. Each model provides a functional rationale for a series of experimental results, typically obtained for non-overlapping or only partially overlapping sets of languages. To the extent of our knowledge, there are no computational implementations of these models.

The only computational models we are aware of that address bilingual language processing are the Bilingual Interactive Activation Plus (BIA+) model (Dijkstra and Van Heuven, 2002; van Heuven and 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 (Widrow and Hoff, 1960; Rescorla and Wagner, 1972; Rescorla, 1988) and multivariate linear regression (Sering et al., 2018; Baayen et al., 2018). Computational implementations of the Discriminative Lexicon therefore differ in several important respects from BIA+ and MULTILINK.

The first difference between Discriminative Lexicon (DL) models and Interactive activation (IA) 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 et al. (2001) and Redgrave et al. (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).

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 modelled 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 IA 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, e.g. 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 IA model would involve creating extra nodes for the word form and its meaning, this is not necessary in the Discriminative Lexicon approach, since the same set of form and meaning features are used across the whole lexicon. 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.

A fourth difference is that, because BIA+ and MULTILINK implement a localist approach to semantics, 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.[1] One approach to creating a such a semantic space is to use one of the many methods developed within

[1] 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 visualised).

the approach of distributional semantics (e.g. Landauer and Dumais, 1997; Turney and Pantel, 2010; Mikolov et al., 2013). In the present study, we construct a semantic space differently, building on the ontology of WordNet. Importantly, once words' meanings are represented as numerical vectors, it becomes possible to study various aspects of their inter-relatedness using mathematical techniques. For example, the emotional content of words can be modeled straightforwardly (Westbury, 2014; Westbury et al., 2014), which in turn can shed light on differences in the emotional connotations of comparable concepts across different languages (Čavar and Tytus, 2018).

A final difference is that BIA+ incorporates a mechanism for modeling task effects, whereas this is currently not implemented for Discriminative Lexicon models.

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 non-selective, such that encountering a linguistic form in any known language will cause activation within a single shared pool of form and meaning representations (cf. Kroll et al., 2010; Brysbaert et al., 2010; Mulder et al., 2014; Dijkstra et al., 2005). 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 e.g. Kroll and Stewart (1994) and Kroll et al. (2010), might be accounted for within our framework.

The remainder of this paper is structured as follows. Section 2 introduces further details on the representations for form and meaning that we use in our simulations, as well as the algorithms that predict meaning from form (for comprehension), and form from meaning (production). Section 3 introduces the materials from English, German, Mandarin, and Dutch that we have used in our simulations. The simulations themselves for monolinguals, bilinguals, and trilinguals are reported in Sections 4, 5, and 6 respectively. Finally, we discuss the implications of our results in Section 7.

2 Representations and algorithms

2.1 Representing meaning

A central question for a computational model of bilingual and multilingual lexical processing is how to design the representations for words' 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 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 differences in usage across languages and, secondly, they cannot represent lexical semantic relations within a language.

Dijkstra et al. (2019) are aware that full translation equivalence is hardly ever reached for actual word pairs, and that 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).

The second 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.

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 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 0.27 to 0.50, of which only 14 were significant under Bonferroni correction at $\alpha = 0.05$. Importantly, for English *dog* and *cat*, and German *Hund* and *Katze*, correlations were much higher (0.83 and 0.98 respectively) than those of the cross-language correlations for *dog/Hund* and *cat/Katze*, which were 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 under-estimating 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 et al., 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 et al., 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 et al. (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 et al., 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 multi-word expressions, and 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)).[2] 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 et al., 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`. The hierarchical organisation 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 the two senses of *palm*. For the HAND sense, meaning 'the inner surface of the hand', its direct hypernym is 'area, region', from which it inherits the hypernyms 'body part', 'part, piece', 'thing', 'physical entity', and 'entity'. With respect to the TREE sense, it has the direct hypernym of '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 hypernym 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 hypernym identifiers as the basis for the semantic representations in our model.

The inclusion of the lexical meaning of the word itself in the meaning representation requires further discussion. The reason for doing this is that WordNet sometimes assigns exactly the same hypernyms to different words, as, for example, is the case for *palm* and *birch* (with the TREE sense), or *apple* and *pear* (with the FRUIT sense). In order to make the meanings of *palm* and *birch*, and those of *apple* and *pear*, semantically distinguishable, we added a unique identifier for each word

[2] For a different approach to morphology that builds on linguistic domain knowledge see Baayen et al. (2019c) and Chuang et al. (2019)

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

sense. Hence, the representations of the two meanings of *palm* include the semantic features of 'palm:hand' (S1) and 'palm:tree' (S6) respectively.

For translation equivalents such as *dog* and *Hund*, the representations derived from WordNet are 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). A first step away from complete identity is to enrich words' semantic vectors with an identifier for the language they belong to. The language within which words are used thus becomes one aspect of what we assume speakers know about them, and acts as a proxy for variation in usage between translation equivalents.

We created a semantic matrix in which we specified which hypernyms from the WordNet hierarchy describe the semantics of each word in our data. The columns of the semantic matrix, henceforth $\boldsymbol{S}$, list all relevant hypernyms (cf. Table 1). The rows specify for the words 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 *Handfläche* have semantic representations that are identical except for the language nodes. 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*.

Table 2: A toy example of the bilingual mental lexicon. Phonological forms are encoded using the DISC notation of the CELEX database, adapted to our data.

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

$$\boldsymbol{S} = \begin{array}{r|cccccccccccccccccc|}
 & \text{S1} & \text{S2} & \text{S3} & \text{S4} & \text{S5} & \text{S6} & \text{S7} & \text{S8} & \text{S9} & \text{S10} & \text{S11} & \text{S12} & \text{S13} & \text{S14} & \text{S15} & \text{S16} & \textsc{english} & \textsc{german} \\
\textit{palm}\text{: }\textsc{hand} & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
\textit{palm}\text{: }\textsc{tree} & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\
\textit{Handfläche} & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
\textit{Palme} & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1
\end{array} \tag{1}$$

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 node, the model would not have any information that could guide it to articulate *dog* versus *Hund*.

2.2 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 and 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 et al. (2014) used an NDL 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 (Wieling et al., 2012) applied to those forms. Thus, the n-gram form features that we also use in the present study provide, in combination with the algorithm used for training the network, the same functionality as the Levenshtein edit distance used by the MULTILINK model. 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 (Hubel and Wiesel, 1962; DeAngelis et al., 1995; Eggermont et al., 1981; Aertsen and Johannesma, 1981). Letter and phone n-grams are high-level proxies for such receptive fields in the respective sensory systems.[3]

In the present study, we based our form representations on triphones. In a matrix specifying words' phonological properties, we listed all triphones observed in the lexicon in columns, and in each row, we used binary coding to specify whether a triphone is present (1) or absent (0) in a given word form. This form matrix is henceforth denoted by $\boldsymbol{C}$. Equation (2) shows the form matrix for the toy dataset shown in Table 2.

[3] For vision, more fine-grained receptive field features can be provided by Histograms of Oriented Gradients (HOG) features (Dalal and 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). In the present study, we did not make use of these more fine-grained features for reasons of interpretational simplicity.

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

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 ','.

2.3 Algorithm: Linear Discriminative Learning

2.3.1 Overview

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., 2019b, 2018). 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 form matrix $\boldsymbol{C}$ (Section 2.2) and the semantic matrix $\boldsymbol{S}$ (Section 2.1), 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 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.

2.3.2 Estimating 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 will use 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 and 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 et al., 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 outputs using the input in conjunction with the current connection weights in the pertinent network. It then compares its prediction to the

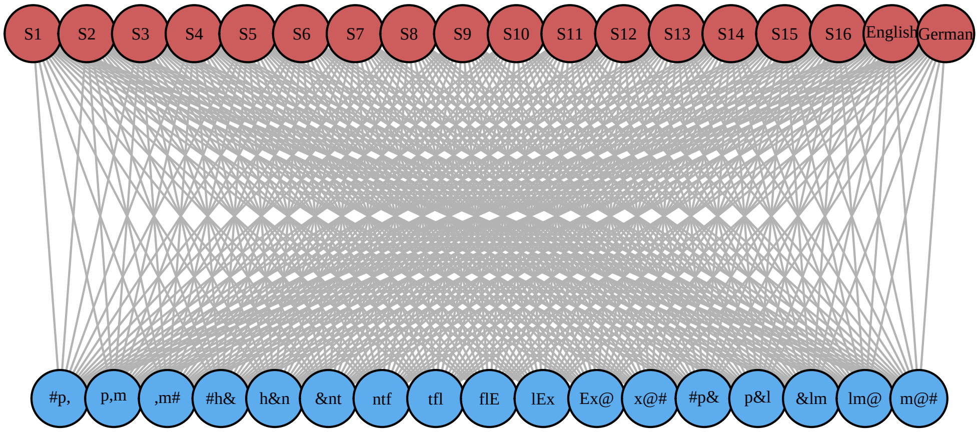


Figure 1: The fully-connected network between triphones and semantic features obtained with LDL. For comprehension, triphones are used to predict semantic features, whereas for production, semantic features are used to predict triphones.

actual output in the 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.[4]

Figure 2 plots the changes in two connection weights as learning progresses. These weights are taken from the production network learned for the toy example 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 green dots indicate learning events in which the TREE semantic feature was not involved, i.e. 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 one of the targeted English triphones. When either of the other words is presented, i.e. 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,

[4] 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.

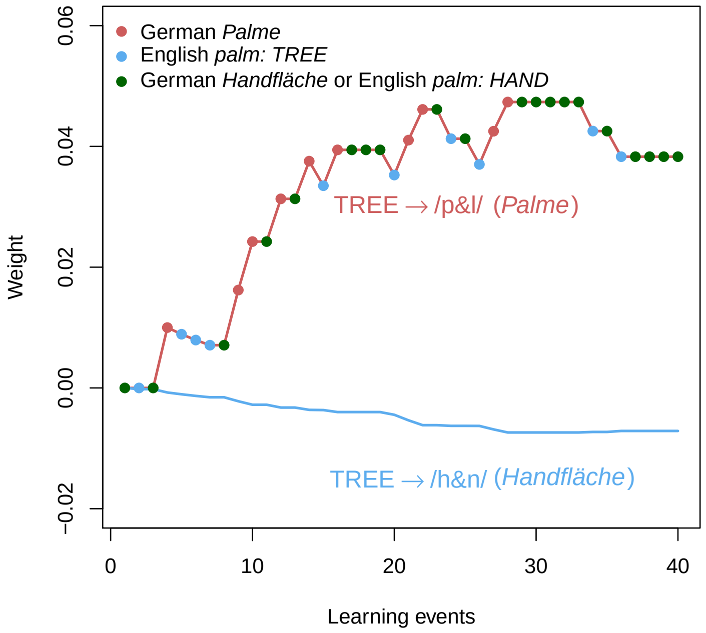


Figure 2: Changes in connection weights from the semantic feature TREE to the triphones /p&l/ and /h&n/as learning progresses in the production network for the toy lexicon shown in Table 2

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 negative weights on its connections to this triphone.

Turning now to the second method of estimation: in addition to modelling 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 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:

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

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

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

Table 3: The connection weights for the triphones of *palm* to all the semantic features. The sum of all triphones' weights for each semantic feature constitutes the word's predicted semantic vector ($\hat{\boldsymbol{s}}_{palm}$).

	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
$\hat{\boldsymbol{s}}_{palm}$	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

Table 4: Correlation coefficients ($\boldsymbol{r}$) of the predicted semantic vector for *palm* ($\hat{\boldsymbol{s}}_{palm}$) with the semantic vectors ($\boldsymbol{s}$) 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	$\boldsymbol{r}$
$\hat{\boldsymbol{s}}_{palm}$	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**
$\boldsymbol{s}_{Handfläche}$	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**

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:

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

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

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.

2.3.3 Evaluation

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 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 scatter plot. 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 present in a word, but not how they should be ordered. Fortunately, as discussed in Section 2.2, 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). 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, 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 $\hat{\boldsymbol{s}}_{/p,lm/}$ respectively. These predicted semantic vectors are then found to be correlated with the actual semantic vector of *Palme* ($\boldsymbol{s}_{Palme}$) with correlation coefficients 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'.[5]

[5] Baayen et al. (2018) derived this name from analysis by synthesis, a term originally coined by Halle (1959) for a process of comprehension through internal production.

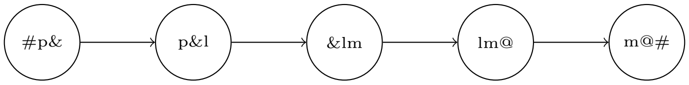


Figure 3: The triphone path for the phonological form of the word *Palme*.

3 Data

3.1 Materials

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 is 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 based on 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.[6] 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 *Schúler* 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 word senses in the model. 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. It should be emphasised, however, that this distribution applies only to our dataset, reflecting the number of senses we assigned to the various words. It is not representative of the actual distribution of homophones and non-homophones in these languages more widely. In reality, across the four languages, homophones are far more abundant in Mandarin than in the three Germanic languages (Duanmu, 1999).

The phone representations of the English, German and Dutch words, in DISC notation, were extracted from the CELEX lexical database (Baayen et al., 1995). The phonological forms of the

[6] 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.

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

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).

3.2 Simulated word frequency

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 behaviour. 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[7]). The meaning with the highest frequency in the current dataset (882) is 'tea' (of the DRINK sense), and that with the lowest frequency (1) is 'installments'.

4 Simulations for monolingual lexical learning

All simulations in this and subsequent sections were carried out using the R package WpmWithLdl (Baayen et al., 2019a).

In order to provide a baseline for the comparison of bilingual and trilingual lexical learning, we first present simulations of monolingual learning for all four languages under consideration. The number of word tokens on which the model was trained was equal to 73,900, which is equal to twice the summed frequency of all word senses. Each word token constituted a learning event. The order of the learning events (word tokens) was randomized. The same random order of senses was used

[7] 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.

for each of the four languages. Thus, for a given word sense, the number of tokens of that word sense presented for learning was twice its frequency, and the locations of the word tokens in the sequence of 73,900 learning events were uniformly distributed. Model performance was evaluated every 7390 learning events. 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.

For these simulations, each semantic vector contained 1133 digits (0 or 1), which was the number of semantic features used. This number was the same for every language and excluded the language nodes: 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). In our dataset, both English and German have many more homophones than Dutch does, so it is unsurprising that the number of triphones in these two languages is reduced compared to Dutch. The relatively small number of triphones for Mandarin is a straightforward consequence of the strong phonotactic restrictions that govern its syllable structure.

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 (Ingram, 1974; Clark, 1993; Blair and Harris, 1981): comprehension skills are usually acquired faster and ahead of the corresponding production skills.

The dots in the top right corner 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 Section 6.2, 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 difference in learning rates reflects the difference in the number of homophones in each language: largest for English and German, intermediate for Dutch, and lowest for Mandarin.

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 these differences 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 in Section 2.3.3, the production model makes use of synthesis-by-analysis: the production system

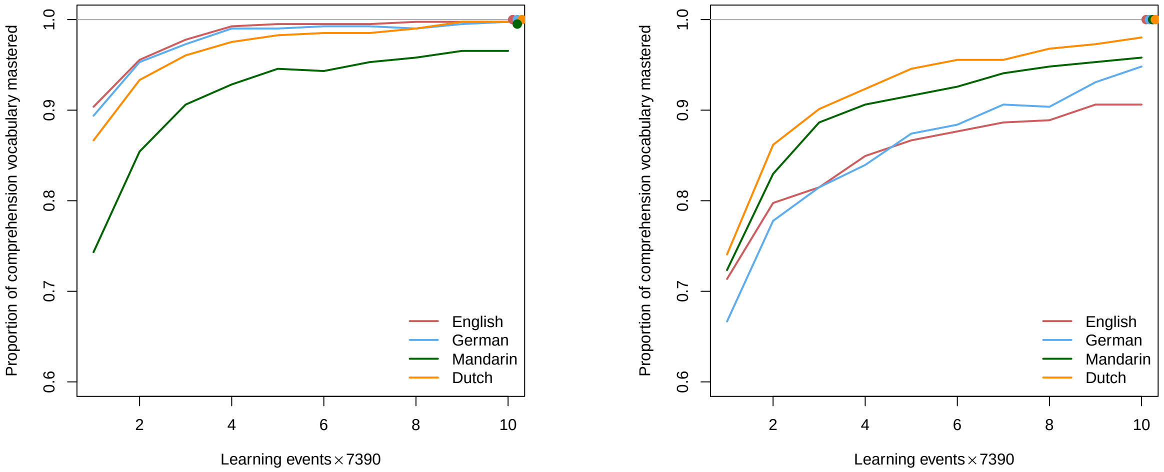


Figure 4: Vocabulary size as a function of exposure for comprehension (left) and production (right), for monolingual learning. The dots to the right of each plot indicate the model's performance at the end-state of learning.

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. Consider Figure 5. The boxplots summarize the distributions of the correlations between predicted and targeted semantic vectors for homophones as opposed to non-homophones at the end of the simulation period. (Every homophone contributes two correlations to the distribution.) The weaker correlations between predicted and targeted semantic vectors for homophones is 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, but not enough to be selected during synthesis-by-analysis. Because the predicted semantic vector for a homophonous target form is likely to be relatively weakly correlated with the target semantic vector, the probability for error increases.

Wrong selections are therefore more often found for homophones than for non-homophones. 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*. It is worth noting that such semantic errors do not persist through learning time. As shown in the right 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 variegated learning pattern in production, we created a second, smaller dataset in which we randomly selected just one meaning for

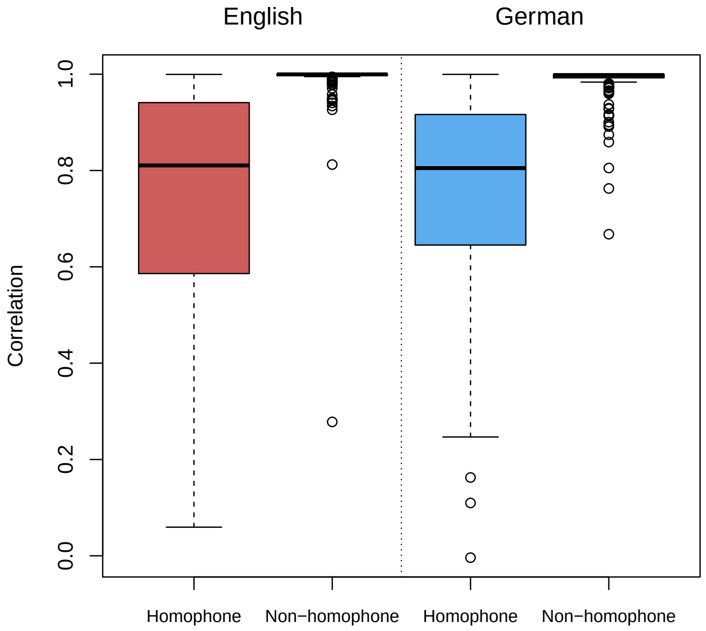


Figure 5: Correlations between predicted and targeted semantic vectors for homophones and non-homophones for English (red) and German (blue).

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.[8] 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 the 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. By 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.

5 Simulations for bilingual lexical learning

5.1 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

[8] This pair of words would not have been homophones if we had also included tone information in our model.

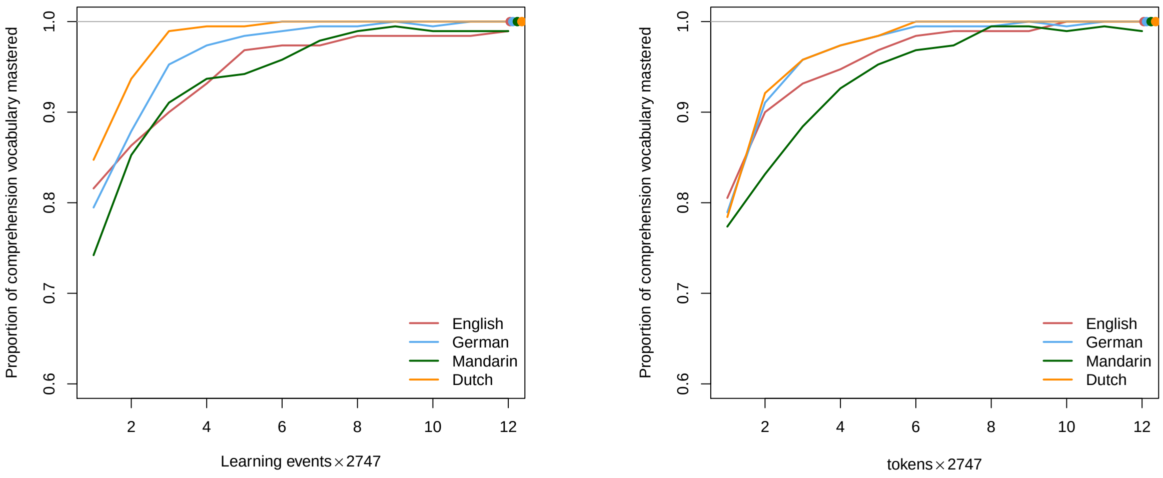


Figure 6: Vocabulary size as a function of exposure for comprehension (left) and production (right), for monolingual learning of a smaller dataset without homophones. The dots to the right of each plot indicate the model's performance at the end-state of learning.

similar to that for the monolingual models, except that two language nodes, one for English (ENG) and one for German (GER), are added to the semantic vectors (cf. Section 2.3). 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. 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.

Results are summarized in Figure 7. The lines represent the proportion of the total vocabulary of each language that the model correctly understands (left panel) or correctly produces (right panel). Bilingual comprehension resembles monolingual comprehension in so far 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 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', i.e. 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.

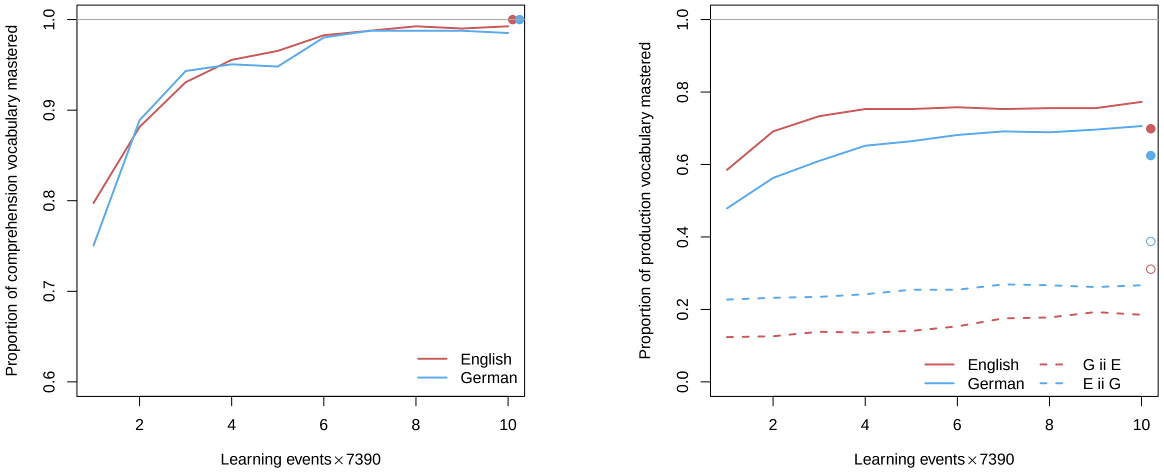


Figure 7: Vocabulary size as a function of exposure for comprehension (left) and production (right), for simultaneous balanced English-German bilinguals. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel 'X ii Y' represent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

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 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 nodes. 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, 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 in

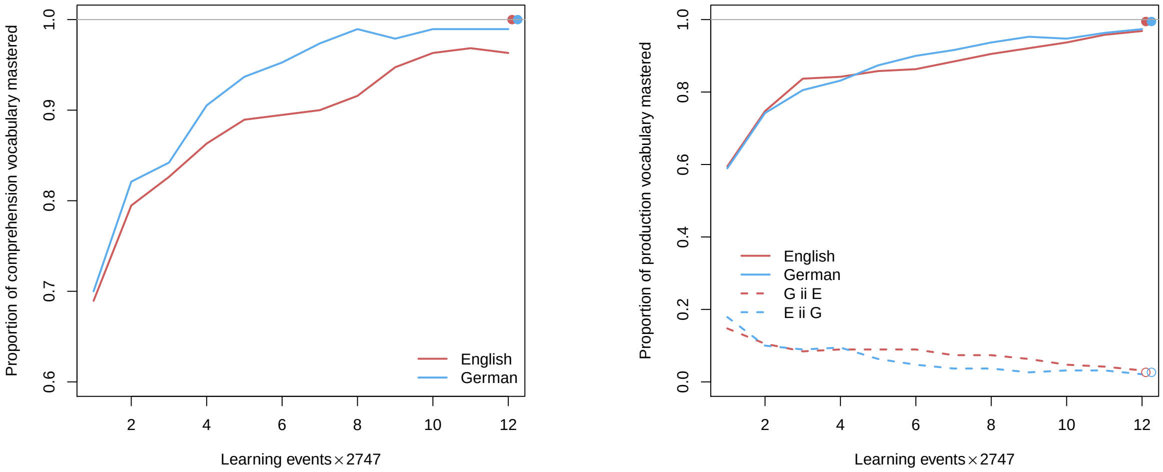


Figure 8: Vocabulary size as a function of exposure for comprehension (left) and production (right), for English-German bilinguals. The training data for this simulation is the smaller dataset without homophones. The dots to the right of each plot indicate the model's performance at the end-state of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

Section 2.1 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. 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. Thus the semantic vector (0, 0, 1, 1, 0), for example, might become (0.001, 0.0005, 1.002. 0.099, -0.001). 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 nodes 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* is, thanks to the addition of a tiny bit of Gaussian noise, now modeled as very similar to the TREE sense of German *Palme*, but not completely identical. 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'. 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.

5.2 Non-balanced bilingual lexical learning: 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

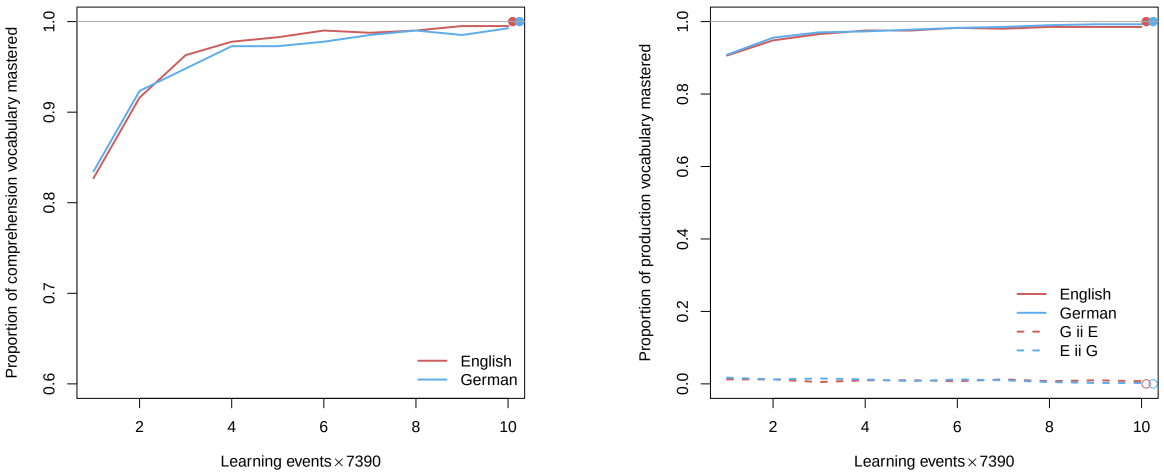


Figure 9: Vocabulary size as a function of exposure for comprehension (left) and production (right), for English-German bilinguals. In this simulation, a small amount of random noise was added to the semantic vector of each individual word. The dots to the right of each plot indicate model performance 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.

which German (L2) was introduced at different stages. For the simulation of early bilinguals, L2 learning started after the second evaluation, i.e. after 14,780 learning events of only English words. For the simulation of late bilinguals, the onset of L2 learning was after the fifth evaluation, i.e. after 36,950 learning events of only English words. In both cases, 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. For these two simulations, during the bilingual learning phase, 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 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.[9] 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, English rarely suffers from intrusion, regardless of the onset time of L2 and the amount of L2 input. For L2, in the absence of homophones, intrusion errors occur primarily at the beginning of learning, after which the amount of intrusion tapers off. The scarcity of intrusions of L1 into L2 is a consequence of the L1 mapping of meaning to form having a substantial head start on L2 learning (see Figure A1 in Appendix B for graphs representing non-balanced bilinguals without homophones).

6 Simulations of trilingual lexical learning

In the preceding sections, 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. 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 German as L2. In a first set of simulations, the form representations are the same as those used for the monolingual and bilingual models presented above, namely vectors of triphones. However, given that Mandarin is a tone language, we subsequently extend the model with representations for tone and intonation, and then explore the consequences for lexical learning of having to master both segmental and suprasegmental structure. In a final simulation, 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.

6.1 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, e.g. 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 comparison, 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 2 triphones in common. Because Dutch is phonologically similar to English and German, whereas Mandarin is

[9] At the first evaluation after L2 onset, all but three German words have been presented to the model at least once. All German words have been encountered at the end of learning.

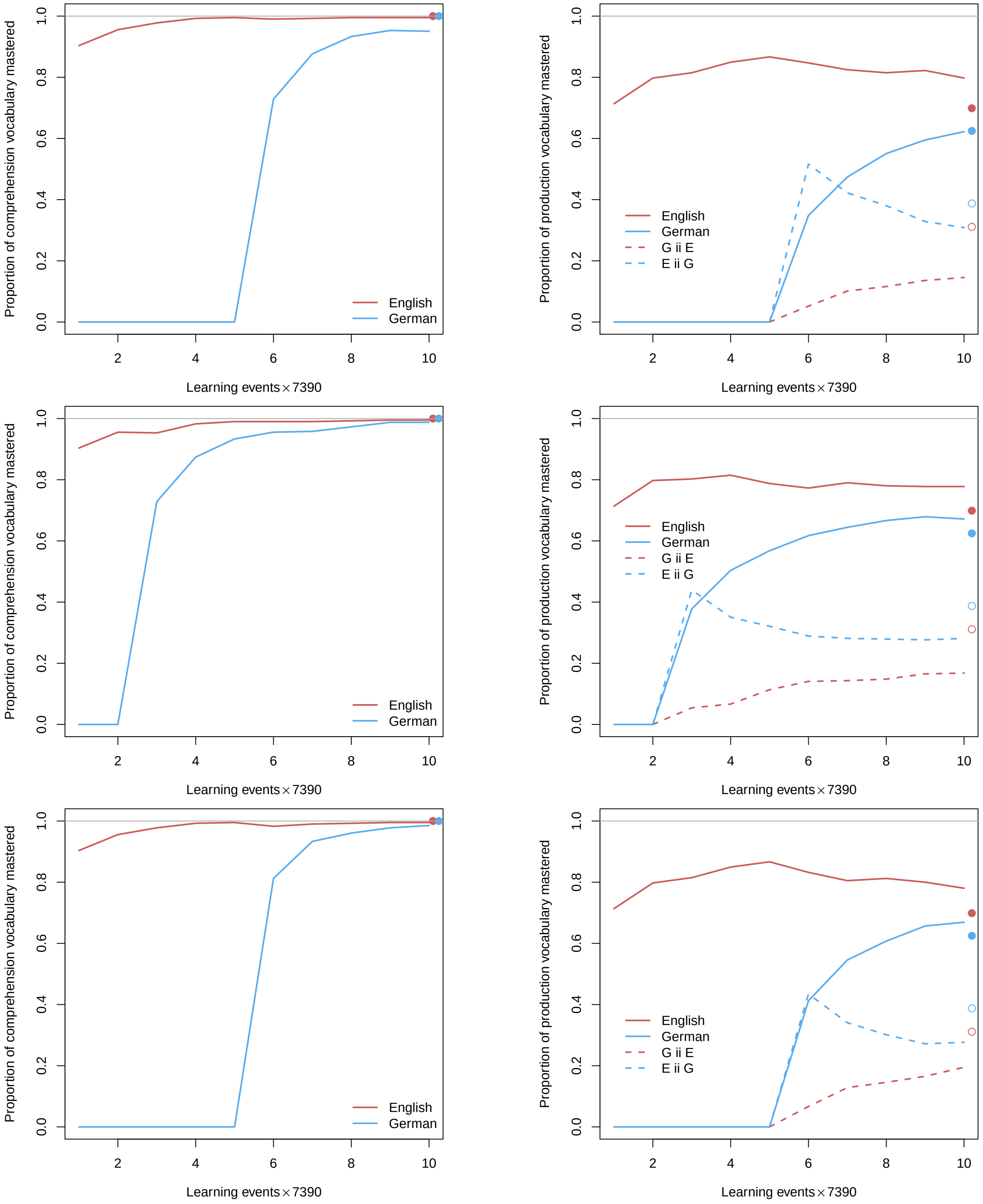


Figure 10: Vocabulary size as a function of exposure for comprehension (left) and production (right), for L1-English L2-German bilinguals. For the upper panels, L2 learning starts after the fifth evaluation, whereas for the middle panels, L2 learning starts earlier, after the second evaluation. The lower 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.

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 Section 5.2. In other words, each simulation started with 5 learning periods of 100% L1 English, followed by 5 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. 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.

Figure 11 presents the development of lexical learning when the third language is Mandarin. Figure A2 (in Appendix C) presents the corresponding developmental curves for L3 Dutch. As the curves for Dutch are very similar to those for Mandarin, they are not shown here. 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 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 12 for Mandarin and A3 (in Appendix C) 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 green and orange dots in the lower panels of Figure 11 and Figure A2, 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 differences in phonotactic restrictions governing Mandarin as opposed to the three Germanic languages. The many triphones that are unique to Mandarin in our dataset, enables 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, i.e. 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 13, which presents by means of boxplots the distributions of the amount of support that words' triphones receive from the semantics. These distributions are presented for both Mandarin (left) and Dutch (right). For each language, the distributions are visualized separately for words without intrusions, words with intrusions from English, and words with intrusions from German. For both Mandarin and Dutch, the semantic support for triphones is weakest for German. For Mandarin, the semantic support for triphones is intermediate for intrusions from English, whereas for Dutch as L3, the semantic support for words with intrusions from English does not differ much from the semantic support for words without any intrusions. Given

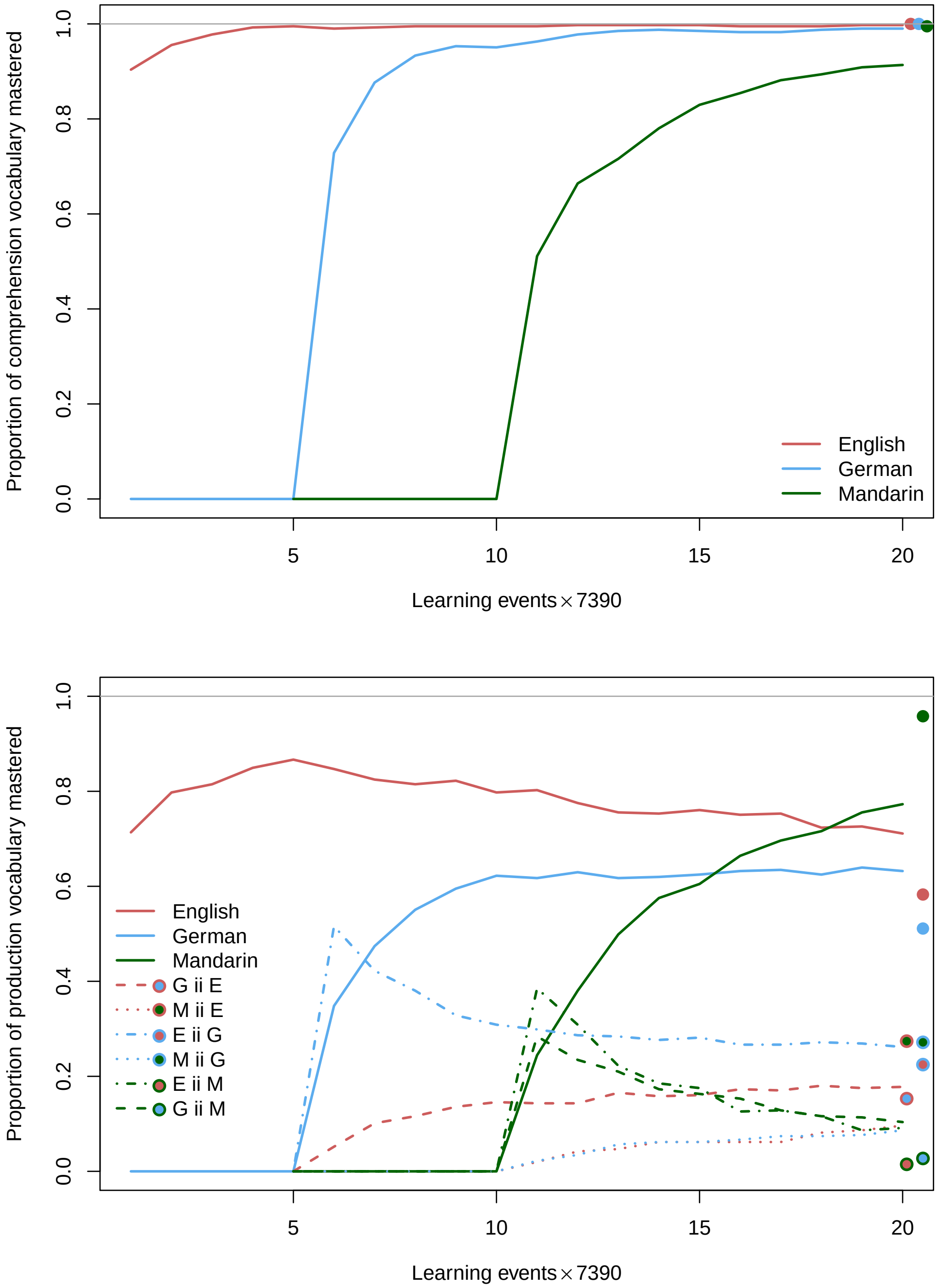


Figure 11: Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Mandarin trilinguals. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the bottom panel 'X ii Y' represent the proportion of intrusions from language X into language Y.

that there are twice as many triphones shared between Dutch and German (263) as compared to the triphones shared between Dutch and English (127), the triphones of Dutch and German are more in competition during learning, resulting in weaker connection weights.

6.2 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. In what follows, we explore to what extent 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 tone. 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 et al., 2005), which can be formalized using ToBI notation as a high pitch accent (H*), followed by a low boundary tone (L%) (Pierrehumbert and Hirschberg, 1990; Beckman and Ayers, 1997; Grice et al., 2005). 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, 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.

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 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

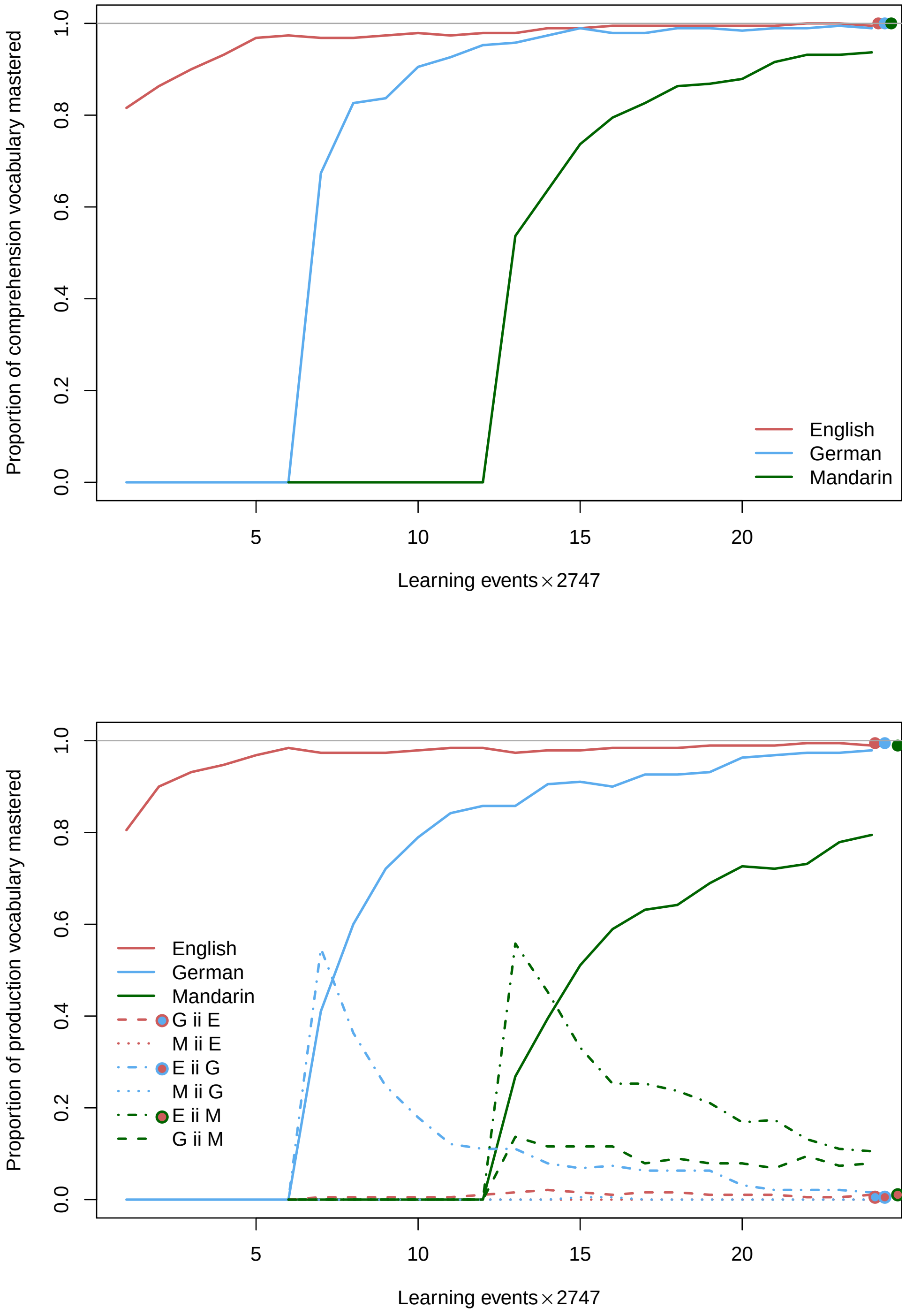


Figure 12: Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Mandarin bilinguals. The simulations were run with the smaller dataset without homophones. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the bottom panel 'X ii Y' represent the proportion of intrusions from language X into language Y.

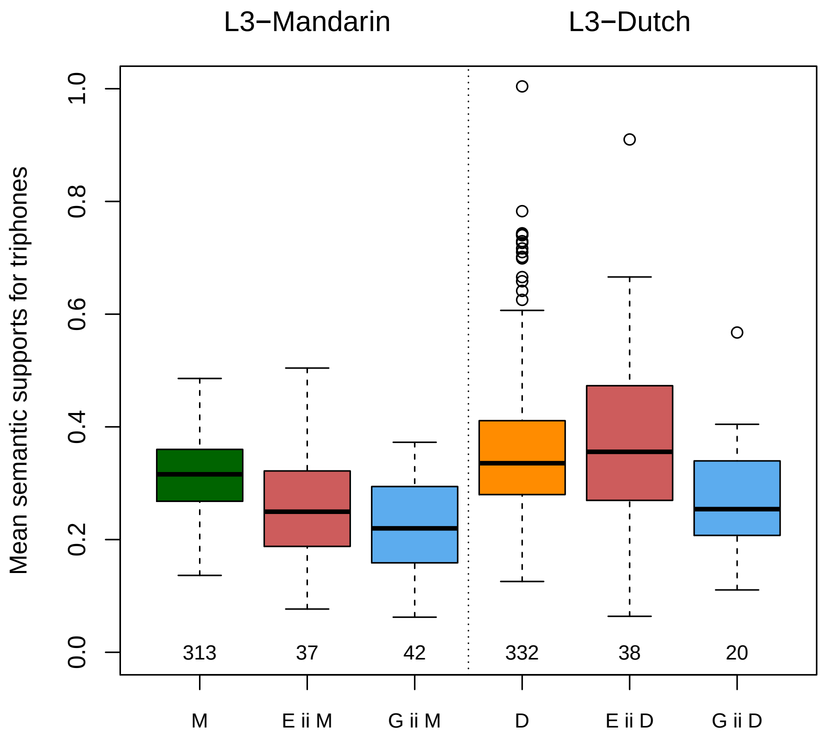


Figure 13: Boxplots for the distributions of the amount of support recieved by words' triphones from the semantics, for Mandarin (left) and Dutch (right) calculated at the end of the simulation period. The first boxplots of the two panels are for words without language intrusion, which are correctly 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.

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	(H)L(H)	mǎ *'horse'*
Tone 4	High-falling	HL	mà *'scorn'*

Table 7: Tonal representations of the falling tonal pattern for English and German. Stressed syllables are underlined.

Representations	English examples	German examples
HL	*bark, organ*	*Tee* 'tea', *Affe* 'monkey'
H–L	*customer*	*anrufen* 'to call'
–HL	*piano, bouquet*	*Kartoffel* 'potato', *Violett* 'violet'
–H–L	*electricity*	*Olivenbaum* 'olive tree'

H–L respectively. With tritones included, the form vectors of these words are now as follows:

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

The number of different tonal patterns for Mandarin was 41, whereas that for English and German was 4.

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 does 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 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/$_{\text{HL}}$, /p,m/$_{\text{H–L}}$, /h&ntflEx@/$_{\text{HL}}$, and /h&ntflEx@/$_{\text{H–L}}$. The predicted form was selected from this set, using synthesis-by-analysis.

Comprehension and production accuracy for the simulation including suprasegmental features are shown in Figure 14. 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 11). 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 production vocabulary 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 reduced (1/3) compared to initial exposure to German (1/2).

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 words' form representations. The tonal features, however, render production more demanding. 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 is simply more difficult as there are more ways in which things can go wrong. At the end of the simulation period, 46% of the Mandarin words obtain wrong predictions for their tonal patterns, 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 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 green dot in the right-hand margin of Figure 14.

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, especially from L3 Mandarin into L1 and L2. L1 English and L2 German suffer on average 17% and 14% less language intrusion in this simulation than in the phone-only simulation. For Mandarin, the number of intrusions is negligible. 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, i.e., 'HL' instead of 'H–L'. There are also cases where only tonal features intrude, e.g., 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 intrusion 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 (for details, see Figure A4 in Appendix C). Given the similar suprasegmental features of the English, German and Dutch, the learning does not differ substantially from the learning of phones alone (Figure A2, Appendix C). 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 (97% *vs* 90%), as indicated by the green and orange dots to the right of the lower plots in Figures 14 and A4 respectively. The superior end-state production accuracy of L3 Mandarin in simulations using our full dataset results from the the fact that Dutch has more homophones in this dataset than Mandarin does. This can be seen by comparing the lower panels of Figures 12 and A3: with homophones excluded from the dataset, the model ultimately achieves full production accuracy in both Mandarin and Dutch.

6.3 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 11 and 14). 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, 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 15.

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 14). 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 regards to production, comparing the lower panels in Figure 15 and Figure 14, 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 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

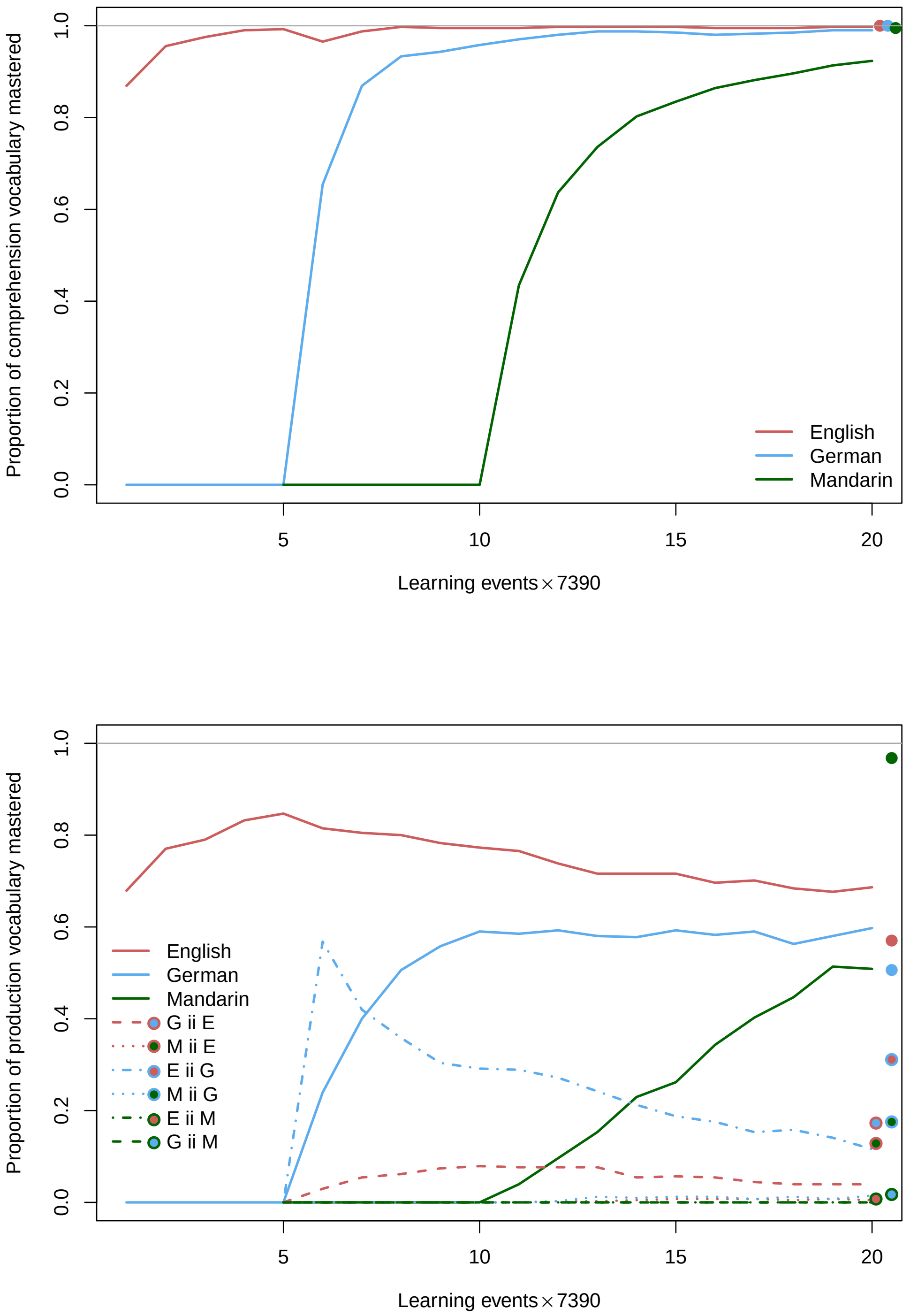


Figure 14: Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Mandarin trilinguals. In this simulation, words' form representations included suprasegmental information. The dots to the right of each plot indicate the model's performance 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.

from English, Mandarin, with very few homophones, 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 nor L3 learning follows straightforwardly any fixed pattern of learning. Instead, much is dependent on which language is learned at what time, and the distributional properties of these languages.

7 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), Ellis and Larsen-Freeman (2009) on the role of discrimination learning in L2, and work by Ramscar et al. (2010, 2013) on discrimination learning in L1 acquisition.

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 are two main findings. First, just as in human learning, in the simulations, production accuracy lagged behind comprehension accuracy. Second, within-language homophones give rise to frailty in the mappings of form to meaning, consequently negatively affecting the learning of both comprehension and production. For comprehension, this frailty gives rise to less good approximations of the targeted semantic vectors, and for production, the reduction in quality in the predicted semantic vectors gives rise to semantic errors.

According to Wilhelm von Humboldt's universal (van Marle and Koefoed, 1980), also known as the bi-uniqueness principle, ideally lexical forms and lexical meanings should be in a one-to-one correspondence. Violations of this principle are hypothesized to be dysfunctional. 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 dysfunctional.

The second set of simulations addressed bilingual lexical learning, with English as first language and German as second language. In these simulations, the 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.

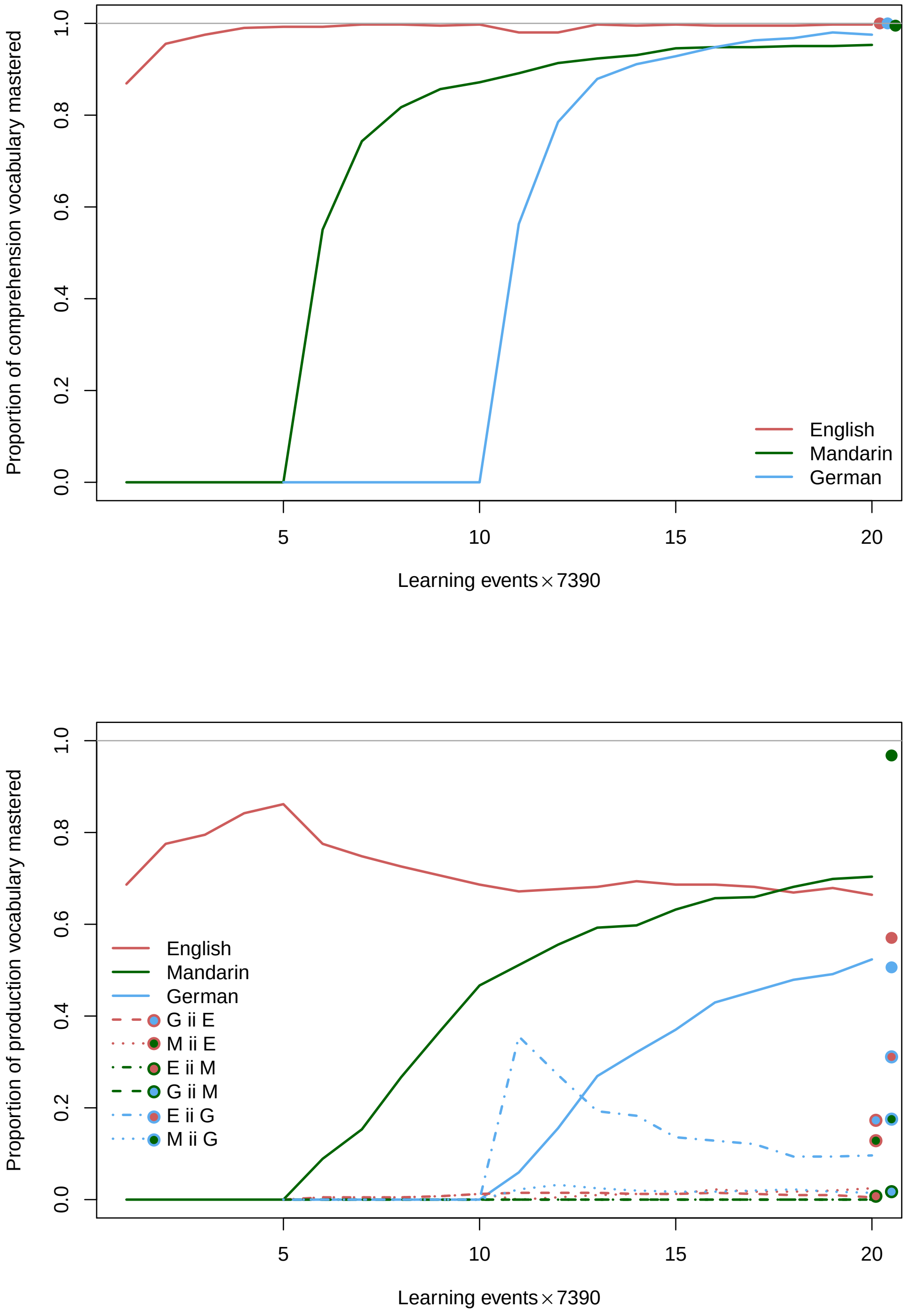


Figure 15: Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-Mandarin, L3-German trilinguals. In this simulation, words' form representations included suprasegmental information. The dots to the right of each plot indicate the model's performance 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.

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 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. By integrating WordNet-based semantic vectors with vectors derived from corpora using methods of distributional semantics, it may be possible to approximate better a developmental process in which speakers gradually tune in into the lexis of the L2.

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). 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), 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 (Section 6.3). When Mandarin was learned as L2, it received hardly any intrusion from either L1-English or L3-German, which is in sharp contrast with the situation when German was learned as L2. Importantly, when homophones were removed from the dataset, the production learning for L1-English and L2-German steadily increased as L3-Mandarin, no longer showing the downward sloping trends resulting from intrusion when homophones were present (cf. lower panel, Figure 12). In other words, in the absence of homophones, L3 learning is very similar to L2 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 the L3 as compared to L2. However, 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 prouction suffers. Since in our model, internal comprehension is part of production (synthesis-by-analysis), the frailty in the mapping from form to meaning has as its 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 A.

In the light of these 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 suggest that with exactly the same learning mechanism, a qualitatively different pattern of acquisition can be observed, depending on language's distributional properties (in our simulations, the presence of many homophones). 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, i.e, 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, and depending on the distributional properties of the pertinent languages, an L2 or L3 can be either dormant, or actively interfering.

A second 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. Surprisingly, it appears to be the oppositions and contrast, at a type level, that matter. When usage is sent to infinity, tokens give way to types.

A third 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 at the end-state of learning intrusion errors persist depends on how words' semantics are represented.

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. By 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 replicated their result, and at the same time extend it to incremental learning.

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 depended 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.

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:

$$\hat{\boldsymbol{C}} = \boldsymbol{S}\boldsymbol{G}^{(L2)} \quad \text{(direct route)}, \tag{8}$$

$$\hat{\boldsymbol{C}} = \boldsymbol{S}\boldsymbol{G}^{(L1)}\boldsymbol{O} \quad \text{(indirect route)}. \tag{9}$$

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.

In the light of these findings, it is not the case that our simulation experiments unambiguously support one of the three high-level models for syntax (TPM, CEM, LPM) that we briefly discussed in the introduction. Contrary to CEM predictions, L3 does suffer from negative transfer in lexical acquisition from L1 (language intrusions). In general, our results are more in line with LPM than the other two models. Our present results do not rule out that any of these models are actually on the right track, as our simulation are based on a very specific theory of learning. There is much more to the cognition of multilingualism than the implicit, subliminal error-driven learning that our model implements. So claims about transfer from only the L1, for instance, could well be correct. But if so, their origin must lie not in discrimination learning but in higher-level cognitive processes.

Important limitations 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. All these shortcomings of this pilot study can be addressed. Our model scales up well to large datasets. Such datasets will also make it possible to examine in more detail whether the model correctly predicts, for instance, whether nouns are more prone to intrusion than verbs (Marian and Kaushanskaya, 2007).

The amount of L2 and L3 input can also be brought down to that typical for second language learning in high-school settings. For instance, for modeling the adverse consequences of ADHD for learning, error can be injected into the learning process. Individual differences in the ease with which additional languages are picked up can be brought into the model by varying the learning rate of the Rescorla-Wagner and Widrow-Hoff learning rules.

We note here that computationally, our study offers two innovations to the theory of the Discriminative Lexicon, as developed in Baayen et al. (2019c); 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 (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, 2002; Ellis and Larsen-Freeman, 2009; Ellis, 2013) has focused on the learning rule of Rescorla and Wagner (Rescorla and Wagner, 1972). Although their learning rule figures prominently in the naive discriminative learning model (Baayen et al., 2011; Milin et al., 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 the Widrow-Hoff learning rule, which is mathematically related to the Rescorla-Wagner rule, to provide 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.

References

Aertsen, A. M. H. J. and Johannesma, P. I. M. (1981). The spectro-temporal receptive field: a functional characteristic of auditory neurons. *Biological cybernetics*, 42(2):133–143.

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.

Arnold, D., Tomaschek, F., Lopez, F., Sering, T., and 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(4):e0174623.

Baayen, R. H., Chuang, Y.-Y., and Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, 13(2):232–270.

Baayen, R. H., Chuang, Y.-Y., and Heitmeier, M. (2019a). *WpmWithLdl: Implementation of Word and Paradigm Morphology with Linear Discriminative Learning.* R package version 2.0.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (2019b). 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.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (2019c). 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*.

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

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

Bardel, C. and Falk, Y. (2012). The l2 status factor and the declarative/procedural distinction. *Third language acquisition in adulthood*, 46:61.

Beckman, M. E. and Ayers, G. (1997). Guidelines for tobi labelling. Available at http://www.cs.columbia.edu/~agus/tobi/labelling_guide_v3.pdf.

Berkes, É. and Flynn, S. (2012). Further evidence in support of the cumulative-enhancement model: Cp structure development. *Third language acquisition in adulthood*, 46:143.

Blair, D. and Harris, R. J. (1981). A test of interlingual interaction in comprehension by bilinguals. *Journal of Psycholinguistic Research*, 10(4):457–467.

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

Bolinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse.* Stanford university press.

Brysbaert, M., Verreyt, N., and Duyck, W. (2010). Models as hypothesis generators and models as roadmaps. *Bilingualism: Language and Cognition*, 13(3):383–384.

Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language*, 32(2):319–343.

Čavar, F. and 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(1):17–28.

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

Chuang, Y.-Y., Lõo, K., Blevins, J. P., and Baayen, R. H. (2019). Estonian case inflection made simple. a case study in word and paradigmmorphology with linear discriminative learning. *PsyArXiv*, pages 1–19.

Clark, E. V. (1993). *The lexicon in acquisition*, volume 65. Cambridge University Press, Cambridge.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, 47(2):109–121.

Davydova, J., Tytus, A. E., and Schleef, E. (2017). Acquisition of sociolinguistic awareness by german learners of english: A study in perceptions of quotative be like. *Linguistics*, 55(4):783–812.

De Groot, A. M. (2011). *Language and cognition in bilinguals and multilinguals: An introduction.* Psychology Press.

DeAngelis, G. C., Ohzawa, I., and Freeman, R. D. (1995). Receptive-field dynamics in the central visual pathways. *Trends in neurosciences*, 18(10):451–458.

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Mller, S., Cieliebak, M., Hoffmann, T., and 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.

Dijkstra, A. and 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.

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

Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., and Rekké, S. (2019). Multilink: a computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, 22(4):657–679.

Du Bellay, J. (1549/2013). *Défense et illustration de la langue française.* Presses Électroniques de France.

Duanmu, S. (1999). Stress and the development of disyllabic words in chinese. *Diachronica*, 16(1):1–35.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2017). Multilingual training of cross-lingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 894–904.

Eggermont, J., Aertsen, A., Hermes, D., and Johannesma, P. (1981). Spectro-temporal characterization of auditory neurons: redundant or necessary? *Hearing research*, 5(1):109–121.

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(2):143–188.

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1):1–24.

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(2):164–194.

Ellis, N. C. (2013). Second language acquisition. *The Routledge Handbook of Second Language Acquisition*, page 193.

Ellis, N. C. and Larsen-Freeman, D. (2009). *Language as a complex adaptive system*, volume 11. John Wiley & Sons.

Falk, Y., Lindqvist, C., and Bardel, C. (2015). The role of l1 explicit metalinguistic knowledge in l3 oral production at the initial state. *Bilingualism: Language and cognition*, 18(2):227–235.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Bradford Books.

Flynn, S., Foley, C., and 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(1):3–16.

Grice, M., Baumann, S., and Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology. *Prosodic typology: The phonology of intonation and phrasing*, pages 55–83.

Gurney, K., Prescott, T., and Redgrave, P. (2001). A computational model of action selection in the basal ganglia. *Biol. Cybern.*, 84:401–423.

Halle, M. (1959). Analysis-by-synthesis. In *Record of AFCRL Workshop in Speech Communication, 1959.*

Harm, M. W. and 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.

Hawkins, R., Lozano, C., et al. (2006). Second language acquisition of phonology, morphology and syntax. In Brown, K., editor, *The encyclopedia of English language and linguistics*, pages 67–74. Elsevier, London.

Hermas, A. (2015). The categorization of the relative complementizer phrase in third-language english: A feature re-assembly account. *International journal of Bilingualism*, 19(5):587–607.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154.

Ingram, D. (1974). The relation between comprehension and production. In Schiefelbusch, R. L. and Lloyd, L. L., editors, *Language perspectives—Acquisition, Retardation, and Intervention*, pages 313–334. University Park Press, Baltimore.

Jarema, G. (2017). Polyglossia as a personal journey. In Libben, M., Goral, M., and Libben, G., editors, *Bilingualism. A framework for understanding the mental lexicon*, pages xiii–xvii. John Benjamins, Amsterdam.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

Kroll, J. F. and Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of memory and language*, 33(2):149–174.

Kroll, J. F., Van Hell, J. G., Tokowicz, N., and Green, D. W. (2010). The revised hierarchical model: A critical review and assessment. *Bilingualism: Language and Cognition*, 13(3):373–381.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Linke, M., Broeker, F., Ramscar, M., and Baayen, R. H. (2017). Are baboons learning "orthographic" representations? probably not. *PLOS-ONE*, 12(8):e0183876.

Maciejewski, G. and Klepousniotou, E. (2016). Relative meaning frequencies for 100 homonyms: British edom norms. *Journal of Open Psychology Data*, 4.

Marian, V. and Kaushanskaya, M. (2007). Cross-linguistic transfer and borrowing in bilinguals. *Applied Psycholinguistics*, 28(2):369–390.

McClelland, J. L. and 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.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. (2017). Discrimination in lexical decision. *PLOS-one*, 12(2):e0171935.

Milin, P., Madabushi, H. T., Croucher, M., and Divjak, D. (2020). Keeping it simple: Implementation and performance of the proto-principle of adaptation and learning in the language sciences. *PsyArXiv*, page to appear.

Mosca, M. (2019). Trilinguals' language switching: A strategic and flexible account. *Quarterly Journal of Experimental Psychology*, 72(4):693–716.

Mosca, M. and de Bot, K. (2017). Bilingual language switching: Production vs. recognition. *Frontiers in psychology*, 8:934.

Mulder, K., Dijkstra, T., Schreuder, R., and 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.

Pavlenko, A. (2009). Conceptual representation in the bilingual lexicon and second language vocabulary learning. *The bilingual mental lexicon: Interdisciplinary approaches*, pages 125–160.

Pierrehumbert, J. and Hirschberg, J. B. (1990). The meaning of intonational contours in the interpretation of discourse. In Morgan, P. R. C. J. L. and Pollack, M. E., editors, *Intentions in communication*, pages 271–311. MIT press, Cambridge.

Ramscar, M., Dye, M., and McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, 89(4):760–793.

Ramscar, M. and Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, 31(6):927–960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6):909–957.

Redgrave, P., Prescott, T., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023.

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *American Psychologist*, 43(3):151–160.

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

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(2):179–190.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

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

Sering, T., Milin, P., and Baayen, R. H. (2018). Language comprehension as a multiple label classification problem. *Statistica Neerlandica*, pages 1–15.

Serre, T., Wolf, L., and Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 994–1000. Ieee.

Shafaei Bajestan, E. and Baayen, R. H. (2018). Wide learning for auditory comprehension. In *Proceedings of Interspeech 2018*, pages 966–970.

Shih, C. (1997). Mandarin third tone sandhi and prosodic structure. In *Studies in Chinese Phonology*, pages 81–123. Mouton de Gruyter.

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

Smith, S. L., Turban, D. H., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

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(2):207–221.

Tytus, A. E. (2019). Active and dormant languages in the multilingual mental lexicon. *International Journal of Multilingualism*, 16(3):357–374.

van Heuven, W. J. B. and Dijkstra, T. (2010). Language comprehension in the bilingual brain: fmri and erp support for psycholinguistic models. *Brain research reviews*, 64(1):104–122.

van Marle, J. and Koefoed, G. A. T. (1980). Over humboldtiaanse taalveranderingen, morfologie en de creativiteit van taal. *Spektator*, 10:111–147.

Westbury, C. (2014). You can't drink a word: Lexical and individual emotionality affect subjective familiarity judgments. *Journal of Psycholinguistic Research*, 43(5):631–649.

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., and Jacobs, A. M. (2014). 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*.

Westergaard, M., Mitrofanova, N., Mykhaylyk, R., and Rodina, Y. (2017). Crosslinguistic influence in the acquisition of a third language: The linguistic proximity model. *International Journal of Bilingualism*, 21(6):666–682.

Whorf, B. L. (1953). *Language, Thought, and Reality: selected writings of Benjamin Lee Whorf*. The MIT Press, Cambridge, Mass. edited and with an introduction by John B. Carroll.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, pages 96–104.

Wieling, M., Margaretha, E., and Nerbonne, J. (2012). Inducing a measure of phonetic similarity from dialect variation. *Journal of Phonetics*, 40(2):307–314.

Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., and Baayen, R. H. (2014). A cognitively grounded measure of pronunciation distance. *PLOS-ONE*, 9(1):e75734.

Appendices

A The effect of within-language homophones in bilinguals

Table A1 presents a small lexicon that contains six English words and six Dutch words. Two of the English words are homophonous (*palm*), and all except for the two homophones have Dutch translation equivalents. In other words, the two homophone meanings in English (i.e., TREE and HAND) do not have corresponding Dutch forms. At the end-state of learning, results show that all English and Dutch words are correctly understood and produced, including the two English homophones. Now consider the situation in which the two homophone meanings in English also have corresponding Dutch forms which are not -homophones (*palmboom* and *handpalm*, as shown in Table A2). Under such circumstances, while the English homophones can still be correctly recognized, their production, however, suffers from the intrusion of their Dutch counterparts. That is, although the intended form to be produced is *palm*, for either the TREE or the HAND sense, the predicted form is however *palmboom* and *handpalm* respectively. This pattern of results clarifies how the frailty that characterizes homophones in the comprehension system only leads to intrusion errors when the homophones of one language have non-homophonous translation equivalents in the other language.

Next, consider what happens when the homophones in one language also have homophonous equivalents in the other language, as shown in Table A3 (English *card* and Dutch *kaart* can both mean the cards with which one sends a greeting, and the cards used in games). Now, at the end-state of learning, no intrusion errors are present, a result that differs from that of the previous example (Table A2). The lack of intrusion errors, despite the presence of homophones, follows from the fact that the homophones now give rise to frailty in both languages. Since for the homophonous meanings no specific form becomes dominant, language intrusions do not occur.

Since homophones render word learning more difficult as a result of double violations of the one-form, one-meaning principle, we can ask whether it will be the case that when there are more

Table A1: A small English-Dutch bilingual lexicon, with one homophonous pair in English.

English	Dutch	Meaning
palm	—	TREE
palm	—	HAND
—	*hemel*	SKY
—	*lidmaat*	MEMBER
bridge	*brug*	BRIDGE
river	*rivier*	RIVER
eye	*oog*	EYE
foot	*voet*	FOOT

Table A2: A small English-Dutch bilingual lexicon. The homophonous pair in English has non-homophonous translation equivalents in Dutch.

English	Dutch	Meaning
palm	***palmboom***	TREE
palm	***handpalm***	HAND
bridge	*brug*	BRIDGE
river	*rivier*	RIVER
eye	*oog*	EYE
foot	*voet*	FOOT

Table A3: A small English-Dutch bilingual lexicon, with homophones in both English and Dutch.

English	Dutch	Meaning
card	***kaart***	GREETING
card	***kaart***	GAME
bridge	*brug*	BRIDGE
river	*rivier*	RIVER
eye	*oog*	EYE
foot	*voet*	FOOT

Table A4: A small English-Dutch bilingual lexicon, with two and four homophones in English and Dutch respectively.

English	Dutch	Meaning
card	***kaart***	GREETING
card	***kaart***	GAME
map	***kaart***	MAP
menu	***kaart***	MENU
bridge	*brug*	BRIDGE
river	*rivier*	RIVER
eye	*oog*	EYE
foot	*voet*	FOOT

homophones in one specific language, this causes this language to become more prone to intrusions. To address this possibility, Table A4 presents a final bilingual lexicon in which English *card* is mapped onto two meanings, whereas Dutch *kaart* is mapped onto four meanings. The greater number of homophones indeed makes the Dutch system more vulnerable, as intrusion errors are now found for all four Dutch homophones. Thus, the degree of frailty in a given language is postively correlated with the number of homophones in that language.

B Word learning of English-German Bilinguals, without homophones

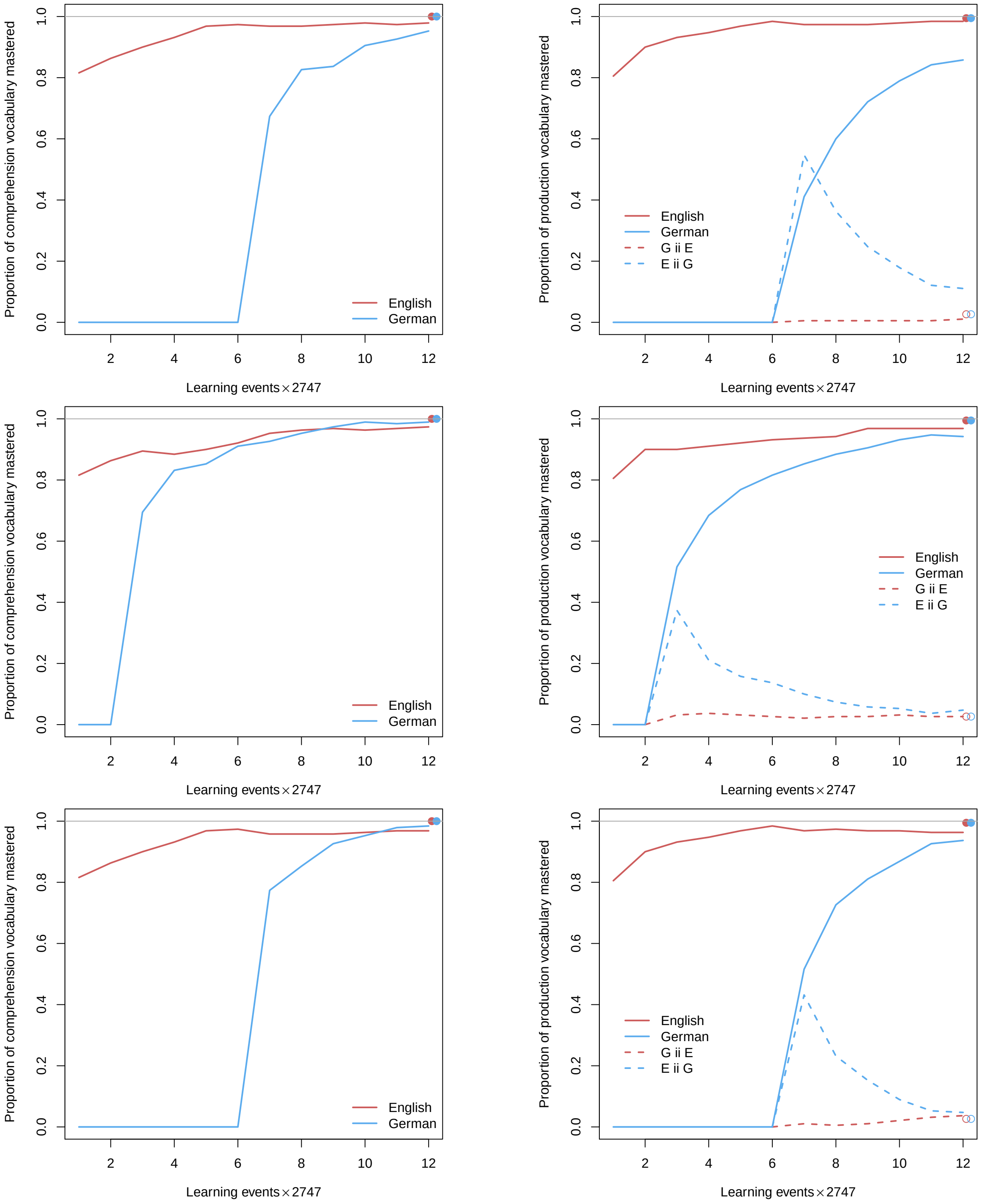


Figure A1: Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English L2-German bilinguals. The simulations were run with the smaller dataset without homophones. For the upper panels, L2 learning starts after the fifth evaluation, whereas for the middle 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.

C Late trilingual learning: Dutch as L3

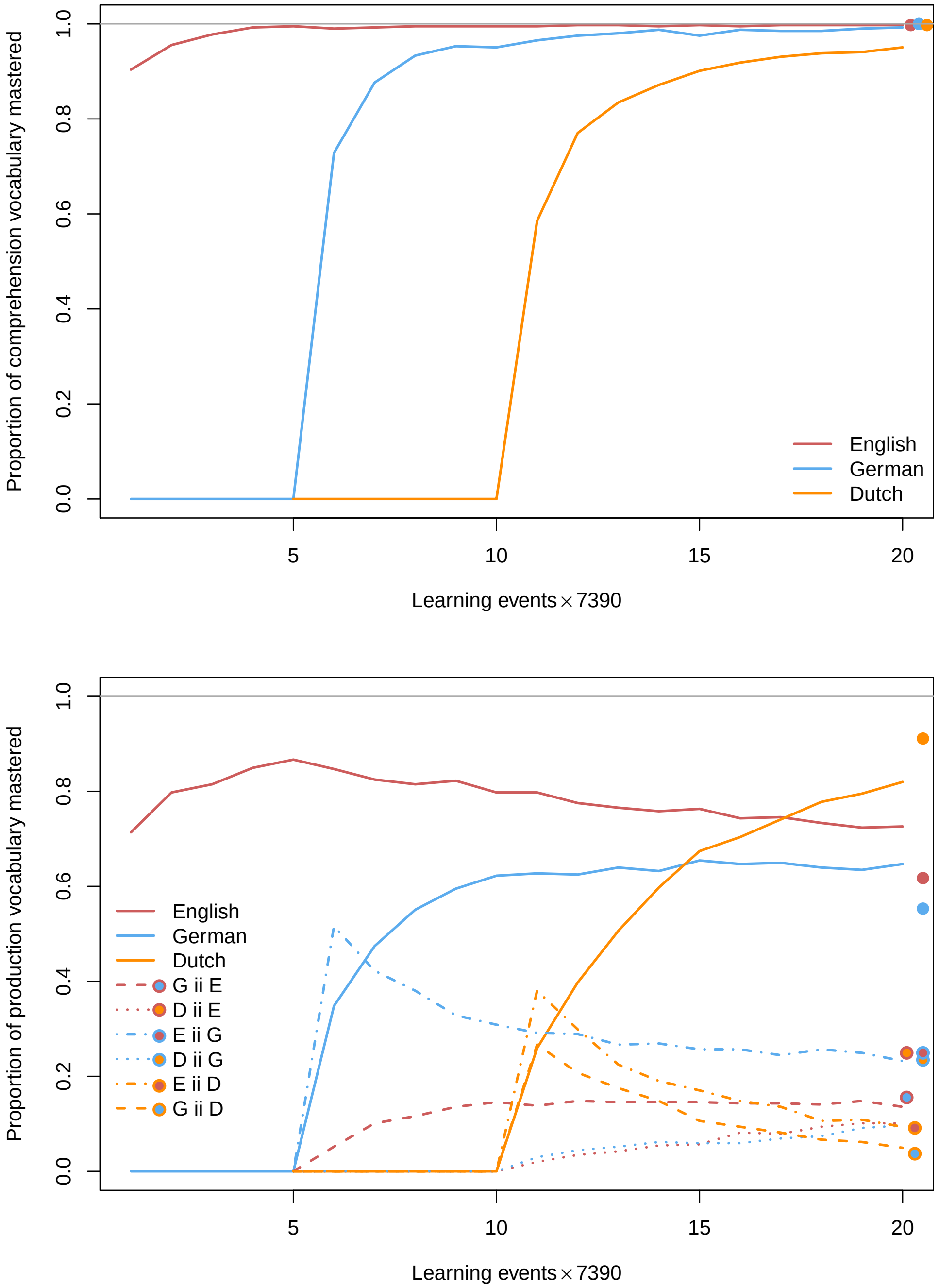


Figure A2: Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Dutch trilinguals. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel 'X ii Y' represent the proportion of intrusions from language X into language Y.

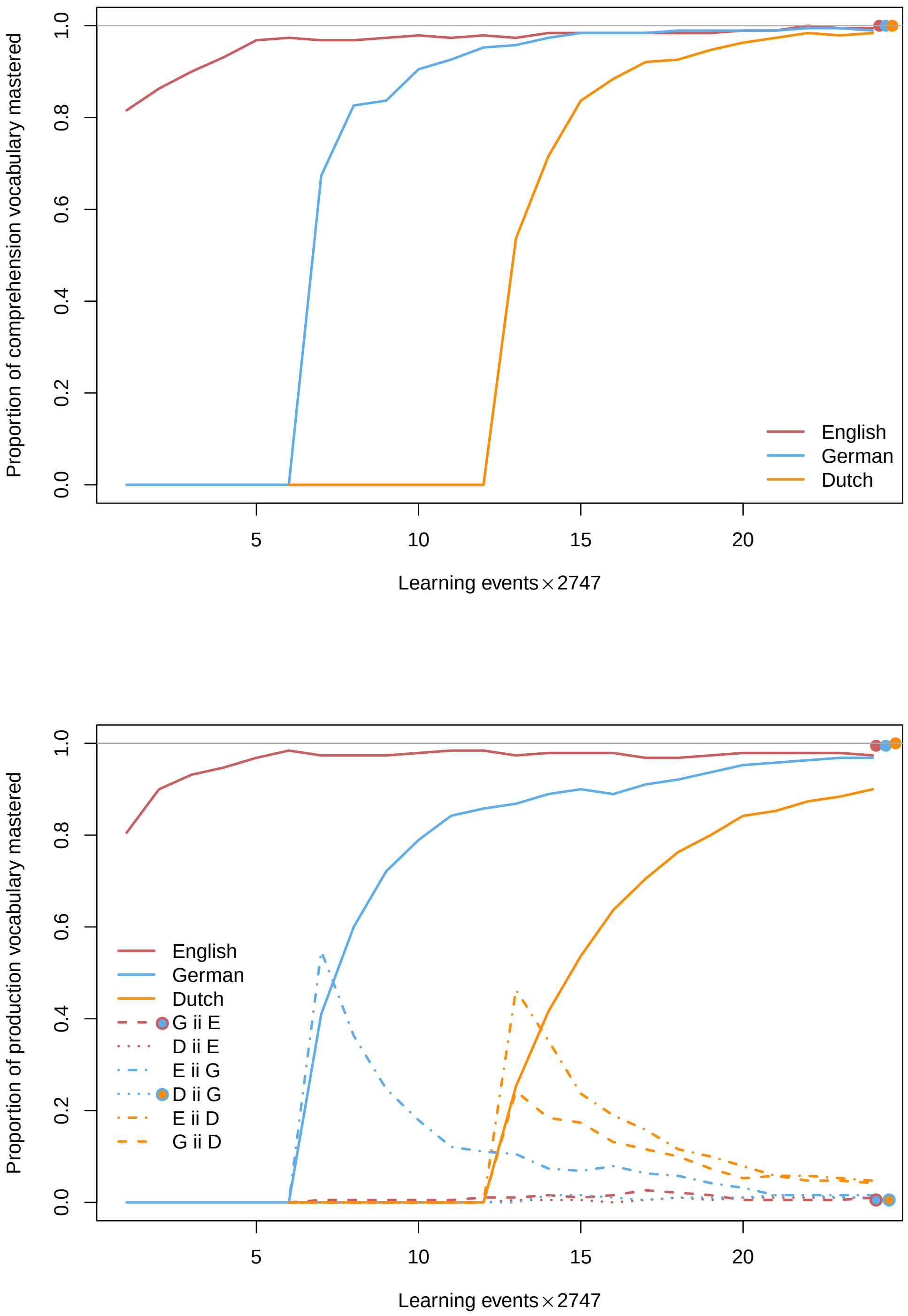
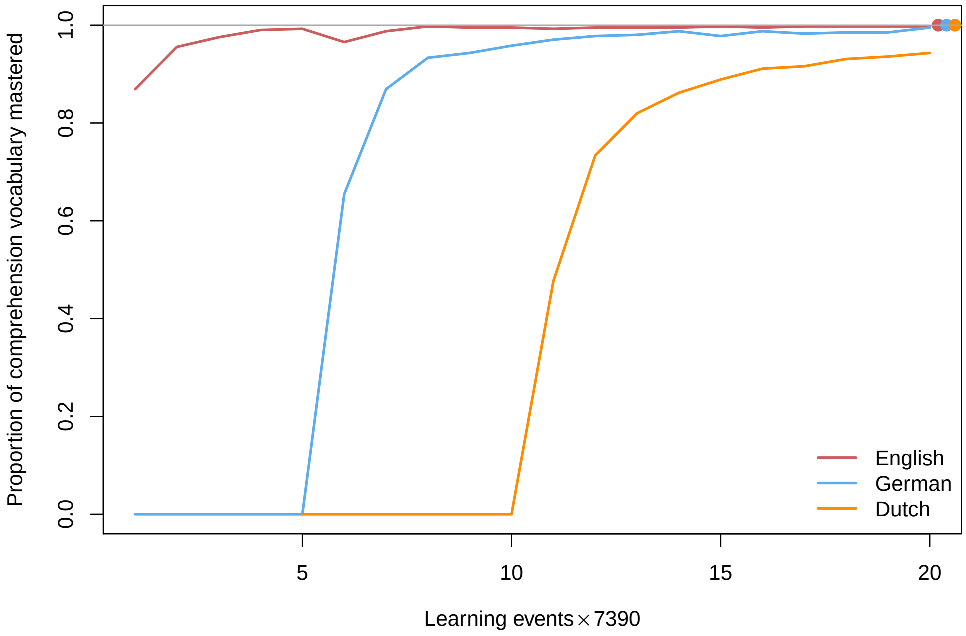


Figure A3: Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Dutch bilinguals. The simulations were run with the smaller dataset without homophones. The dots to the right of each plot indicate the model's performance at the end-state of learning. The dashed lines in the right panel 'X ii Y' represent the proportion of intrusions from language X into language Y.



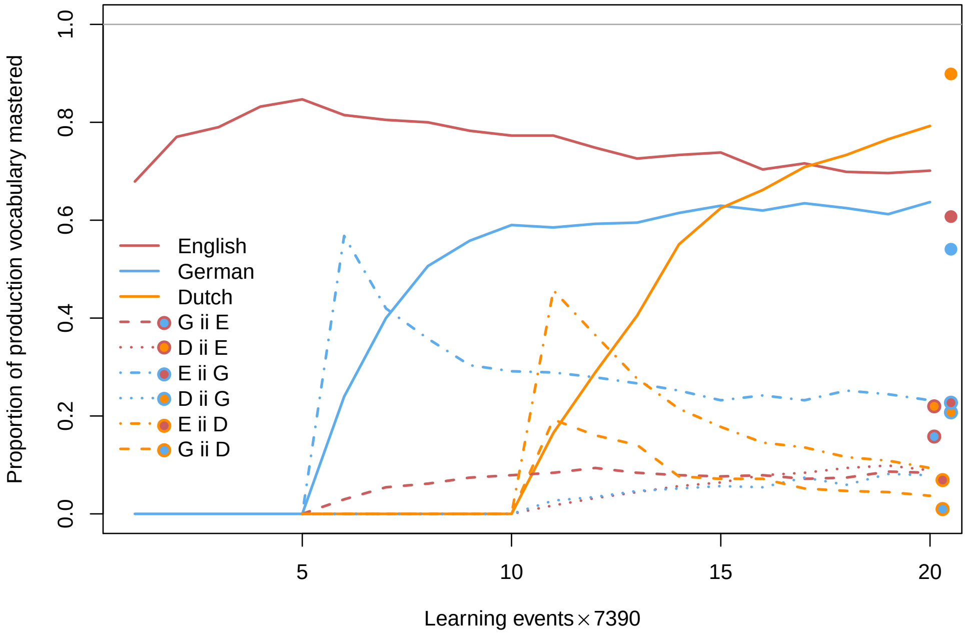


Figure A4: Vocabulary size as a function of exposure for comprehension (top) and production (bottom), for L1-English, L2-German, L3-Dutch trilinguals. In this simulation, words' form representations included suprasegmental information. The dots to the right of each plot indicate the model's performance 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.