A self-organizing model of sequential and simultaneous late language learning

Frank Leoné, Supervisors: prof. dr. A.F.J. Dijkstra, dr. P.A. Kamsteeg
Donders Centre for Cognition, Radboud University Nijmegen

Language learning is typical a sequential process, in which one language is learned after the other. There is reason to believe however that simultaneous language learning, or learning words from multiple languages for one concept at same time, is more efficient. Not only can early learners successfully learn languages simultaneously, associative learning also predicts simultaneous learning to be advantageous in general. Moreover, the integrated nature of the lexicon, with all languages in one storage, seems well fit for simultaneous multilingual learning. To test the likelihood of the hypothesis that simultaneous language learning is indeed beneficial, we developed a model of the lexicon called the Self Organizing Model of MUltingual Processing (SOMMUP) using self-organizing maps. One map successfully learned semantic similarities, the other one orthographic similarities. Importantly, none of the maps developed any language-specificity. The model was able to successfully predict the patterns in reaction times as found in specific and generalized lexical decision tasks depending on word frequency, neighborhood density, and neighborhood frequency. Using the validated model, we tested the effect of sequential, mixed, and simultaneous language learning. Due to imbalances in the tests we could not draw conclusions on the results however, though signs of relevant patterns were found. Combined, these results not only warrant further research into the possibility of simultaneous language learning, but also have interesting consequences for our view of the human lexicon and models thereof.

1. Introduction

Young children have the impressive ability to learn languages at a greater pace and to a greater proficiency than adults. Moreover, they can do so sequentially as well as simultaneously, without obvious detrimental effects on speed and level of acquisition (Snow, 1993). Do only children have this ability to learn languages simultaneously and is it their developing brain that allows for such amazing feats, or would it also be possible, even advantageous, for adults to learn languages simultaneously?

At first sight, language learning in children seems to be qualitatively different from that in adults. For a long time, researchers believed in the existence of a critical period, in which the brain would be optimally equipped for language acquisition (Lenneberg, 1964). More recently, however, the distinction between early and late learning is considered to be less strict, and the concept of a critical period has been questioned (Birdsong, 2005). Instead, a more quantitative approach has been proposed: The ability to learn languages is thought to decline gradually with age, in contrast to a sharp decline after a circumscribed period. A gradual decline does not only imply that at a later age it is still possible to learn languages to a certain proficiency, but it might also entail that simultaneous language learning is possible for late language learners too.

Whether it is indeed possible for late learners to learn multiple language simultaneously has not yet been subject of research. One reason lies in the intuitive expectation that simultaneous language learning is not beneficial at all. At first sight, simultaneous language learning would appear to be detrimental, because the increased cognitive load of simultaneous learning could result in a mixing up of languages. The abundant similarities that exist between languages, especially for languages from the same language family (Ruhlen, 1991), would only increase this effect, because one would no longer be able to tell whether a particular variant of a word belongs to one language or the other. According to this line of reasoning, keeping languages separate in the process of learning is needed to keep them separate in the lexicon, as well as in actual usage.

However, the correctness of this intuitive account can be questioned for several reasons. For instance, some

For this project to succeed, the guidance and support of a number of people was important. First and most important of all, my supervisors, who helped me keep focused from the beginning. At the same time, prof. dr. Dijkstra and dr. Kamsteeg gave me enough free room to try and test new ideas. I am really grateful for the confidence they have showed to have in my and I hope the end result lives up to the expectations. Secondly, I am indebted to S.R. van Eck, MSc., who stood by me almost to the end. She is the one who gave me my confidence to start large projects such as the current one throughout my student career. Thirdly, the feedback of both S.R. van Eck, MSc., and M.B.M. Leoné, MSc., on earlier versions of the manuscript was greatly appreciated. Finally, the last tests of the model could not have been done without the facilities of the Donders Centre for Cognitive Neuroimaging. My gratitude goes out to all those mentioned and not mentioned who helped to realize this project.

interactions between languages are unavoidable and also emerge in sequential learning in the form of transfer (Odlin, 1989) from the native language to foreign languages and vice versa (Pavlenko & Jarvis, 2002). These can actually have both positive and negative effects on the rate of acquisition. The effect is positive for shared parts of languages, such as cognates[1] (Lotto & Groot, 1998), but is negative for aspects that differ, such as phonemes in the foreign language that do not exist in the native language (Gathercole & Thorn, 1998; Groot, 2006). This interaction between languages is in line with the demonstration that the human lexicon consists of one store for all words, irrespective of language, rather than of several stores, one for each language (Dijkstra, 2005). This counterintuitive organisation of the human lexicon also has consequences for simultaneous learning, because if words from all languages end up in one big store even after sequential learning, there is no direct reason left to expect detrimental language mixing effects of simultaneous learning.

In sum, whether late simultaneous language learning is possible, even beneficial, remains an open question, waiting to be answered. Three different outcomes of research into this issue are possible. In the worst case, the greater cognitive load of simultaneous language learning and the smaller segregation between languages could lead to a decreased rate of foreign language acquisition, both for similar and dissimilar language aspects. We call this possibility the 'Interference hypothesis'. Alternatively, one could expect facilitation for similarities between languages, but detrimental effects for dissimilarities, as in the case of transfer. If this is the case, the question remains which effect is the strongest; A cost-benefit analysis would then determine whether or not simultaneous language learning is worth the effort. This hypothesis we call the 'Similarity-dependent facilitation hypothesis'. The third possible effect is that both similarities and dissimilarities are learned more effectively due to the active (conscious) and simultaneous comparison between the words in different languages, allowing them to be stored more effectively in the integrated lexicon. This implies that the learning of both similarities and differences between languages should be influenced positively. This last hypothesis we will refer to as the 'Facilitation due to comparison hypothesis'.

The goal of the present study was to assess these hypotheses (see table 1) and their associated predictions on how late simultaneous learning influences language learning. The obvious way to test them would be to let human participants learn lists of words from existing or non-existing languages simultaneously and sequentially, and to examine the effect on error rate, error types, and speed of acquisition. However, we instead adopted a different approach, namely to construct a model of the human lexicon with which these hypotheses can be tested both qualitatively and quantitatively. Following this approach, all aspects are under the control of the experimenter, in contrast to studies with human subjects. Human subjects can, for example, know more of a language than they consciously report, use unpredictable learning strategies, or just not pay attention. The drawback of building a model instead of doing human experiments is that the validity of a model is hard to verify; one can only try to make the model as plausible as possible, paving the way for subsequent studies in human subjects. So the second goal of the study was to develop a model which was structurally plausible and would also allow us to study the way and the speed in which a multilingual lexicon develops.

Table 1
Three possible hypotheses concerning the effect of simultaneous language learning, with the minus sign ('-') signaling a negative effect and the plus sign ('+') a positive effect on learning.

Hypothesis	Similarities	Dissimilarities
Interference	-	-
Similarity-dependent facilitation	+	-
Facilitation due to comparison	+	+

In sum, first a model of the fully learned multilingual lexicon we called SOMMUP (Self-Organizing Model of MUltilingual Processing) was built and validated. After confirming that the learned performance of the model was in accordance with data from experiments, we attempted to determine to what extent the model's language acquisition method, simultaneously or sequentially, influenced error rates and speed of acquisition. Finally, we formulated proposals for behavioral experiments to test model predictions, as well as a proposal for a user model on the basis of the cognitive model. In total, this amounts to a quantitative test of the hypotheses in a new model of the bilingual lexicon and the acquisition thereof, in order to shed light on the advantages or disadvantages of a simultaneous instead of sequential language acquisition approach.

2. Sequential versus simultaneous learning

As a preliminary to model construction, we will first consider differences between the different forms of learning and counterarguments against the intuitive account of a detrimental effect of simultaneous learning. Sequential and simultaneous learning modes are not as distinct as they might seem, but form the extremes of a continuum. Sequential language learning on the one end involves learning one language after another. First, one becomes proficient in a language, and then, possibly after some time, one starts to learn the next. This is the mode of learning often seen in specialized language courses. On the other end of the continuum is totally simultaneous learning, in which words for a concept are presented at the same time in multiple

[1] For explanations of the vocabulary, the reader is referred to appendix A, in which the most important psycholinguistic concepts are listed for reference.

languages. In-between forms of learning also exist. For example, in high school multiple languages are learned within the same time period, but with separate sessions for each language. This latter type of learning we will refer to as mixed learning.

The focus of the current study is to determine to what extent these different forms of learning affect speed and accuracy of learning. There are at least two important factors that may codetermine the learning effect:

1. The structure of the human lexicon
2. The effect of simultaneous learning

When considered in combination, there appear to be good reasons to expect different effects than intuitively expected by most. An elaboration of these reasons follows in the subsequent sections.

2.1 The structure of the human lexicon

It seems trivial that simultaneous language learning can only take place successfully if the lexicon is able to differentiate the language streams of two or more languages that arrive more or less in parallel. In other words, it should not depend on sequential input to keep the languages separate. Intuitively, an advantage of sequential language learning is the clear segregation of languages, facilitating separate storage in the human language system. This segregation could help to keep languages apart in both perception and production. However, it turns out that the human language system actually has an integrated organization. Simultaneous lexical access in multiple languages has been demonstrated to be part and parcel of human language processing, implying that simultaneous learning may be less problematic or detrimental to the learning process than expected. Considerable evidence converges on this view of an integrated, simultaneously accessed lexicon (see Dijkstra, 2005). In the following, we consider three lines of evidence in support of the integrated nature of the lexicon. Studies like the reviewed ones will be important later for testing the cognitive model we developed.

The first line of evidence in support of an integrated lexicon involves interlingual homographs. The rationale underlying this research is that, if the lexicon is integrated across languages, interlingual homographs should yield different response times than non-homographs, because the active readings from both languages can affect processing. This effect should only be present for bilinguals, as they know the multiple readings of the word, and no such effect should be present for monolinguals. Many studies confirm this view. For instance, Lemhöfer and Dijkstra (2004) tested Dutch-English bilinguals in both an English and a generalized lexical decision task. In an English (L2)[2] lexical decision task, they found that homographs were recognized faster than English control words. In a generalized lexical decision task, homographs were again found to be recognized faster than L2 control words, but about equally fast as L1 control words. Lemhöfer and Dijkstra state that this difference between tasks is probably due to a difference in the homographs' relative frequency in the two languages: L1 words are subjectively much more frequent than L2 words. This difference in subjective frequency leads to faster recognition of L1 (Dutch) words compared to L2 (English) words. In the English lexical decision task, the slow recognition for the L2 reading of homographs is facilitated by the faster L1 recognition, resulting in in-between reaction times. In the generalized lexical decision task on the other hand, the L1 reading of a homograph can be used exclusively to recognize the homograph, making recognition of homographs as fast as the recognition of non-homograph L1 words. The contribution of the slower L2 reading to the reaction time is probably negligible in this case. Other researchers confirmed that no homograph effect exists for monolinguals (Studnitz & Green, 2002).

The second line of evidence focuses on the cross-linguistic effect of interlingual neighbors. If the lexicon is integrated, an effect of the number of *inter*lingual neighbors on word recognition is expected, just as there is an effect of *intra*lingual neighbors (Andrews, 1989; Grainger, 1990). This is exactly what was found by Grainger and Dijkstra (1992), who reported that the number of neighbors in L1 influences recognition of words in L2 in a lexical decision task. The more neighbors a word had in L1 compared to L2, the slower the responses of the participants were. L2 words with more neighbors in L2 than in L1 were recognized faster, possibly because the same-language neighbors help to recognize the word as a member of a particular language. In a follow-up study, Van Heuven, Dijkstra, and Grainger (1998) replicated the earlier results in both progressive demasking and lexical decision experiments: The greater the number of neighbors in L1, the slower the reaction times on L2 recognition.

The third and last line of evidence concerns the effect of context and prior knowledge of the expected language on word recognition. If a specific language context or prior knowledge could help the language system to exclude words from non-target languages, access would be language-specific and words in different languages would still be separable to a certain degree. To test this effect, Dijkstra et al. (2000) did three experiments using mixed lists of Dutch-English homographs that were either of high-frequency in one language and low-frequency in the other, or of low-frequency in both. In the first task, participants had to judge which language a word belonged to (a language decision task), while in the other two tasks they only had to respond to items either in Dutch or English (a go/no-go go task). Results were comparable to the results discussed earlier for homographs, with a striking additional effect: Participants often missed the low-frequency meaning of a word if a high-frequency one also existed in the other language, even if they did not need to respond to the language of the high-frequency word. For example, subjects failed to correctly classify the English-Dutch homograph ANGEL

[2] For the clarification of conventions such as the L1-L2 distinction and the representation of orthography and semantics, see appendix B.

as a Dutch word, as the English reading is more frequent than the Dutch one. This finding shows that information about the target language cannot be used to exclude words from a non-target language. Other potential evidence for the language membership of a target word, like the language of the previous word in a list (Studnitz & Green, 1997; Thomas & Allport, 2000) or prime, unconscious knowledge of expected language (Bruijn, Dijkstra, Chwilla, & Schriefers, 2001), can also hardly be used to facilitate word recognition. In total, the available evidence clearly points to a language non-selective access procedure and an integrated lexicon.

To summarize, research has found many interlingual interactions in language comprehension. Taken together, the currently dominant view is that the lexicon is integrated and is accessed in a language non-selective way. Instead, bottom-up competition between semantically similar concepts and orthographic similar words, across languages, guides the process of lexical access. The implication is that, as the language system is using a mixed representation of words from different languages and language context effects hardly influence lexical selection, there is no direct reason to expect negative effects of simultaneous language learning on the representations on the lexical level; a facilitatory effect is at least as likely.

2.2 The effect of simultaneous learning

Even though bilinguals possess an integrated lexicon, they are still able to distinguish the languages of the words within the lexicon, both when judging the language of a word, as when producing speech. This property of words, which could well be extralexical (Dijkstra & Heuven, 2002), needs to be learned during language acquisition and hence could be distorted by simultaneous learning due to the mixing of languages. Intermixing of words from different languages during acquisition does indeed occur (Odlin, 1989), though with both positive and negative effects that depend on the specific similarities and dissimilarities between languages. There are, however, no strong reasons to expect increased language confusion due to simultaneous learning; in fact, less confusion appears more likely.

When one learns a new language, abundant interactions occur between the native and foreign language, due to transfer from one language to the other and back. Especially lexical transfer, i.e., the transfer of words from one language to another, takes place quite frequently (MacWhinney, 2005). A high degree of transfer implies that, initially, L2 learners use their L1 lexical knowledge in L2 understanding, making L2 totally dependent on L1 (MacWhinney, 2005). With increased L2 proficiency, this dependence decreases and L2 develops a language system of its own, especially when L2 language structure is significantly different from L1. High proficiency in L2 can even lead to opposite transfer, from the foreign to the native language (Pavlenko & Jarvis, 2002). This makes sense, because in general the direction of transfer is determined by the relative strength of the languages, modulated by the applicability of the rules, categories, and words from one language to another (Pienemann, Biase, Kawaguchi, & Håkansson, 2005; MacWhinney, 2005). In addition, as mentioned, transfer between languages can have both positive and negative effects, since similarities between languages are learned faster due to transfer, while differences are often found to be more difficult to acquire (De Groot & Van Hell, 2005).

Although transfer allows language learners to make use, to some extent, of cross-language similarities, this transfer is mostly an automatic process. It may result in overgeneralization, but, since only salient similarities transfer, also in missing out on similarities that remain hidden due to slight differences in, for example, word form. For instance, the similarity between NOTTE, NOCHE and NUIT, is understandable from a historical perspective, but is not striking enough to automatically facilitate learning when learned separate from each other. In addition, in the case of multiple non-native languages, transfer mainly occurs from the stronger L1 *to* the weaker non-native languages and to a lesser degree *between* the non-native languages.

Thus, a differential effect of simultaneous compared to sequential learning is not a case of intermixing versus non-intermixing of languages. Neither is it a case of positive versus negative effects of such intermixing, because intermixing with both kinds of effects is also found in sequential learning. Rather, the remaining question is whether this intermixing will become worse when a languages are learned simultaneously or that simultaneous learning will actually lead to less intermixing and improved language learning.

To answer this question, we now turn to the study of associative learning, which is thought to be the basis of most, if not all, of the learning in both animals and humans (Lieberman, 2000; Skinner, 1953). Associative learning is based on the development of associations between stimuli, primarily induced by simultaneous presentation. In vocabulary learning, the foreign word is normally presented together with the native translation or a (graphical representation of) the concept in order to form such an association. This method is called paired associative learning (De Groot & Van Hell, 2005). If one would apply this method in a simultaneous way, one stimulus would be presented together with its translational equivalents in multiple languages and the learners would need to learn the similarities and differences between them: They have to learn to discriminate the different words for the same concept, making their task essentially a discrimination conditioning task.

In contrast to research on language learning, for discrimination conditioning comparisons have been made between simultaneous and sequential learning. In a variety of tests, e.g., on object naming (Cuvo et al., 1980) and concept learning (Tennyson, Tennyson, & Rothen, 1980), simultaneous discrimination conditioning proved more effective than successive discrimination conditioning with respect to learning speed, number of errors, and retention. The explanation often given is that simultaneous presentation allows for easier comparison and discrimination, allowing for better separation and storage of the stimuli. On the other hand, successive presentation, certainly over a long period of time, results more in generalization than discrimination. Apparently, instead of making distinctions between slightly different words for the same concept, the representation of the native word is generalized as much as possible to try to incorporate the new words, so no distinction between native and foreign words is made until this is absolutely necessary. This slows down the process of learning to distinguish words from different languages in sequential learning, in contrast to the facilitating effect found in simultaneous learning.

Even if one is skeptical about the extent to which words can be reduced to simple stimuli, there is evidence that simultaneous presentation also facilitates rule formation and reasoning skills (Lee, 1982). Rule formation in this case involves the induction of rules upon the confrontation of the stimuli only, both implicitly and explicitly. There are plenty of rules in the comparison between words in different languages that could help the discovery of similarities and differences between translational equivalents.

This is nicely shown by a number of European projects that focused on determining the rules of conversion between languages on the basis of the similarities and differences between languages, and tried to put the results of this comparison to use in teaching. Examples are the Eurom4 (Castagne, 2001), Galanet (Degache, 2003), IGLO (Mondahl, 2002), and EuroCom projects (McCann, Klein, & Stegmann, 2003). The first and second concentrated on the similarities between the Romance languages (Italian, Spanish, Portuguese, French), the second on Germanic languages (Danish, Norwegian, Swedish, Icelandic, English, Dutch, German), and the third on all languages in the European Union. The EuroCom project, the largest project and the only one still active, distinguishes seven sieves, or conversion rules, which are mostly based on lexical similarities and are depicted in table 2. Knowing these conversion rules could, according to the founders of the projects, greatly facilitate language learning. These projects confirm that, at least for European languages, translational equivalents are often so orthographically similar that they can be converted into each other using rules. Thus, they are similar enough to expect a facilitating effect on language learning.

Instead of explicitly teaching these rules, the current study assumes that language learners can derive these rules themselves to some extent when confronted with simultaneous language learning. In addition, teaching the rules to the language learners should lead to even further facilitation. In contrast, sequential learning is expected to separate languages too much, hindering an active and elaborate comparison for useful similarities and differences.

We conclude that the expectation that simultaneous language learning will result in increased confusion between languages is not founded on empirical evidence. Admittedly, in language learning intermixing of languages occurs, but it also does in sequential learning, for better and for worse. Moreover, there is no reason to expect that the confusion effect increases with simultaneous learning. To the contrary, experimental studies on simultaneous versus sequential conditioning have shown that simultaneous presentation of stimuli facilitates both stimulus discrimination and rule formation. These are expected to facilitate the acquisition of words in foreign languages, which would be in line with the 'Facilitation due to comparison hypothesis'.

2.3 Summary: the likelihood of simultaneous learning

To summarize, the expectation that simultaneous language learning will have a negative influence on language acquisition seems based on two premises, which on closer inspection both do not hold. The first states that the lexicon is not made to process languages simultaneously and needs sequential learning to keep languages apart. However, the lexicon is *not* organized on the basis of language membership, but on the basis of orthographic similarity. Moreover, lexical access is language aspecific: Or languages are accessed simultaneously all the time. As such, there should be no difference for the lexical processing between simultaneously and sequentially presented languages. The second premiss concerns the acquisition process itself, predicting more interference when stimuli are presented together. Evidence from associative learning shows the opposite though: Simultaneous presentation is beneficial for the learning of discriminations, which is essentially what needs to be learned in the acquisition of a new language. Without a clear basis for the common-sense notion, it is again an open question whether simultaneous language learning will work or not in practice. We aim to provide the first answers to this question in this thesis.

3. SOMMUP: A new model of multilingual vocabulary learning

We took a modeling approach in order to answer the question what the effect of simultaneous language learning is. This means we required a valid model of the multilingual lexicon. Because there is no existing model that completely incorporates the properties of the lexicon as described in section 2.1 and is actually a learning model that allows to test the hypotheses, a new model is proposed. The new model, called SOMMUP, was built, first concentrating on general plausibility and then zooming in on the effect of learning schemes. In this section, the design of the new model, structure, data, training and performed tests are described.

3.1 Design of the model

A number of choices had to be made in order to construct a plausible and usable model of the lexicon and lexical learning. These choices were largely based on properties of the human lexicon as given in the previous chapter. More details on the choices and their implications are provided in the subsequent sections.

3.1.1 Restriction of the domain. Language learning is a large domain, because many aspects of a language have to be learned (e.g., grammar, orthography, phonology) and many

Table 2
The seven sieves distinguished by EuroCom for facilitated learning of most of the European languages, focusing on vocabulary acquisition.

Nr.	Sieve focus	Description
1	International vocabulary	Focuses on the 5000 words which are shared across languages, largely based on Latin or Romance.
2	Pan-Romance vocabulary	About 500 words that are common to the Romance language family.
3	Sound correspondences	Educates the sound correspondence formulas, or letter combinations which diverged during the development of languages, but actually share a common root and meaning.
4	Spelling and pronunciation	Establishes the conversion rules from spelling to sound, showing which regularly occurring letter combinations in different languages correspond to common sounds.
5	Pan-Romance syntactic structures	Educates the nine basic sentence types found in Romance languages.
6	Morphosyntactic elements	Provides the basic formula for discovering the common grammatical elements.
7	Pre- and suffixes	Describes the common and specific pre- and suffixes, allowing to separate these parts from the root words for easier identification.

words and rules exist. As a consequence, the first choice in the design of any model of multilingual language learning is in terms of content: Which aspects should be incorporated and which should be excluded? In our case, we restrict our model to vocabulary learning, leaving all grammar rules out of the model. This choice significantly reduces the required complexity of the model, but still keeps its applicability to real world situations, because the vocabulary is thought to be the most important part of a foreign language to be mastered (De Groot & Van Hell, 2005).

A second restrictive choice concerns whether orthographic and/or phonological aspects of vocabulary should be included in addition to semantics and language membership. Orthography has the advantage that it is (mostly) equal across alphabetic languages, whereas phonology shows more variations in sound repertoire and is harder to encode. Moreover, more databases of orthography are available than of phonology. Because a large dataset containing words in a significant number of languages is required for a model of multilingual learning, this makes orthography the preferred aspect of language to include. The choice for orthography implies that some effects, such as phonological neighborhood effects, cannot be accounted for by the model when they are not accompanied by orthographic neighborhood effects (e.g., the English word LANE and Dutch word LEEN).

In summary, the model was restricted to vocabulary learning using semantics, language, and orthography, which constitute three essential ingredients for successful word translation.

3.1.2 Localist vs. distributed model. Models can be of a localist or a distributed type. A localist model uses single nodes to represent single symbolic entities, while distributed networks use the pattern of activation in a number of nodes to represent such entities. The choice for a localist or distributed model depends largely on the purpose of the project. We wished to build a learning model that ideally should scale well when concepts and words are added in the future.

For this purpose, a localist model does not seem to be the best choice. In this model type, one node would be assigned to each concept or word form, as, for example, in the Bilingual Interactive Activation model (BIA)(Dijkstra & Heuven, 2002). The implication is a linear increase in the number of nodes with the number of concepts and words, achieving no dimension reduction at all of a given input database. Even more importantly, the weights within these models are often set by hand and no learning or development occurs.

The second type, that of distributed models, is inspired by the biological neural coding of information: It is the combined pattern of activations in a group of nodes that represents a concept or word, which is an efficient way of reducing dimensionality. Moreover, distributed models in general are learning models for which a wide range of learning algorithms exists. Therefor, a distributed model appears the best choice for our model, implying that distributed representations for semantics, language membership, and orthography are needed.

3.1.3 Choice of algorithm. As the next step in setting up the model, we needed to choose a learning algorithm from the existing series of learning algorithms for distributed networks. The algorithm should be able to incorporate the most important aspects of the lexicon and the word learning process. In this regard, the lexical competition for both words and concepts, based on similarities and dissimilarities within and between languages, is of great importance. In addition, the algorithm should be able to learn to recode between combinations of semantics, language membership, and orthography in several directions. This latter restriction makes many algorithms unusable, because most are only suited for learning in one direction, and only allow learning in other directions by explicitly training the model also for these directions. Two algorithms that do not have these restrictions are Radial Basis Function (RBF) networks and Self Organizing Maps (SOMs).

RBF networks are built of neurons incorporating different kinds of non-linear function, the so-called basis functions (Bishop, 2006). The properties of these functions are often trained first, after which a linear combination of the basis functions fitting the output is found in a second training step. The basis functions can be chosen to be bidirectional, if functions are used with such properties (e.g., Gaussians as in Deneve, Latham, & Pouget, 2001). However, RBF networks used in such a bidirectional way are often not trained, but set by hand and are not suited for representing neighborhood relations.

In contrast, SOMs have been used extensively to represent neighborhood relations (Kohonen, 2001). SOMs were developed to distribute multidimensional data on a lower dimensional map, often as low as two dimensions. In the context of language learning and multilingualism (Li, 1999, 2000, 2001; Li & Farkas, 2002; Li, Farkas, Zhao, & MacWhinney, 2004; Li, Zhao, & MacWhinney, 2007), this approach has proven to be fruitful, and it provides an effective and intuitive way of explaining neighborhood and other effects. Learning by SOMs is also regarded as a biologically plausible way of learning, implementable even by mere neuronal Hebbian learning.

Importantly, the SOM-algorithm is an unsupervised algorithm, i.e., there is no feedback-signal to drive learning. This might seem to be a problem, because the model needs to learn translations, for which feedback is standardly used. Li and colleagues also built a SOM model of bilingual language learning and found a solution to this problem (Li & Farkas, 2002). They trained the network by linking two SOMs, representing phonology and semantics, with Hebbian learning. Training of the associations between the semantics and phonology SOMs occurred by presenting data to both, which can be thought of as representing an input and a target, and correlating the activations in the maps using Hebbian learning. After learning, the weights between the two SOMs represented the correlations between the unique word and unique semantic representations. In this way, activating a word in the phonology SOM automatically activated the appropriate language-specific concept in the other SOM and the other way around. However, this method is not applicable to language unspecific semantic representations with a separate language representation, because there is no longer an unique one-to-one relation between words and concepts: The mapping problem of phonology to semantics is no longer linearly separable and cannot be resolved by Hebbian learning. Interestingly, mere unsupervised learning in a SOM can instead be used to learn the input-to-output mappings, a process called autoassociative mapping (Kohonen, 2001). By means of this technique, the activated units in the semantics and phonology or orthography maps, combined with the language information, can be mapped together on yet another SOM to learn the associations, which suits the current purposes well (see figure 1A for an explanation).

A possible disadvantage of the use of SOMs becomes apparent from the work of Li and colleagues: These models are essentially localist in nature, because following learning, each concept or word is linked to one representational unit and far more units are needed than there are words or concepts. This goes against the principle of dimensionality reduction. Using interpolation methods (Göppert & Rosenstiel, 1993, 1995, 1997; Aupetit, Couturier, & Massotte, 2000; Campos & Carpenter, 2000; Flentge, 2006) that allow to determine points in-between nodes, this shortcoming can be corrected. However, preliminary tests indicated that the autoassociative capabilities of SOMs depend heavily on one unit per pattern in case there is no direct relation between the neighborhoods in the to be associated subspaces. Figure 1B graphically depicts this problem. This meant that we had to use a localist representation in the hidden layer. In theory, more generalization might be reached by turning each neuron into a convertor for a small part of the subspaces, a convertor that is 'mappable' from one subspace to the other, a combination of NG and RBF networks (figure 1C). Time limitations on the project prohibited the implementation of this solution.

In sum, in the new model the SOM algorithm was incorporated, because it can be used bidirectionally and is sensitive to neighborhood relations and lexical competition. Another SOM was incorporated for the mapping from orthography to semantics. Preliminary tests indicate that this mapping could only be achieved by means of a localist representation, which means no dimension reduction was reached, even though theoretically interpolation methods should be able to resolve this problem.

3.1.4 Representation of semantics. A major issue regarding the representation of semantics is whether concept representations are shared between languages or not. Apart from this point, the semantic representation must be sufficiently high in resolution to allow for a detailed discrimination of concepts, and it must be upscalable, because the lexicon needs to incorporate a large number of concepts.

Li and colleagues (Li & Farkas, 2002; Li et al., 2004) chose to encode the semantic properties of a word by means of the accompanying words in native texts for both their DevLex and SOMBIP model. For example, the fact that RIVER is frequently accompanied by WATER tells us something about its meaning. In addition, it tells us something about the semantic similarity of RIVER and BOAT, because BOAT will also frequently be found in combination with WATER. However, this approach automatically results in language-specific representations, because the accompanying words in a native text are in a specific language, and therefore different for different languages. This is in strong contrast with the language-independent representations that are thought to be present in the brain and used in a number of other models (e.g., Dijkstra & Heuven, 2002).

For obtaining proper language aspecific semantic representations, at least three approaches exist. The

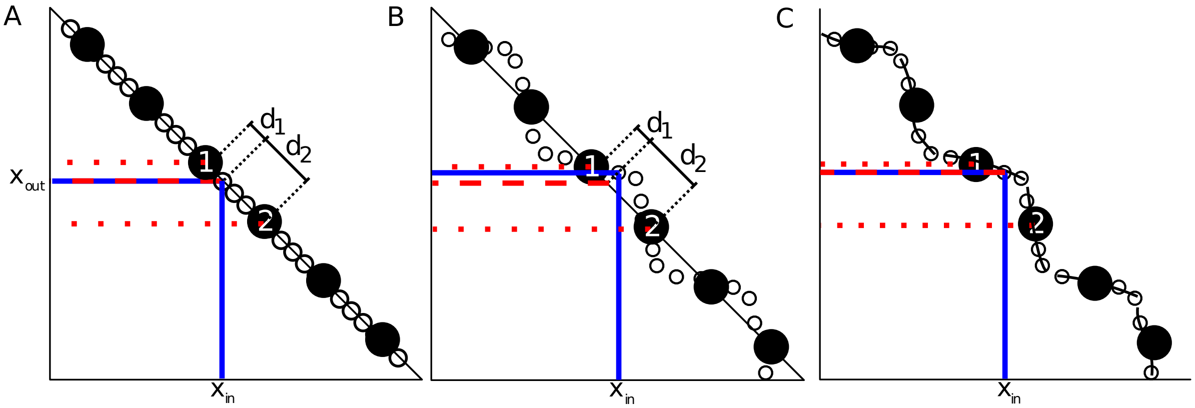


Figure 1. **A**. A simplified schematic view of autoassociative mapping of two-dimensional data on a one-dimensional SOM. The axes represent the input and output part of the data. The SOM is represented by the large filled circles, the model vectors, and the black line, which describes the surface defined by the model vectors. The blue line indicates the data vector *x*, of which applying only the input part (x_{in}) should lead to the appropriate output part (x_{out}). The closest model vector is denoted by 1, the second by 2. Only taking the output part of model vector 1 leads to an overestimation of x_{out} (upper red line), while using model vector 2 leads to an underestimation (lower red line). Taking the weighted average with respect to distances *d*1 and *d*2 of model vectors 1 and 2 leads to a correct approximation (middle dashed red line). **B**. If the data is not as regular, meaning that the input coordinates cannot be converted to the output coordinates as straightforward as in the case described by A, applying a weighted average does not lead to a correct approximation: the red dashed line is not on top of the output part of the blue line. **C**. A possible solution to the non-mappable input and output data of panel B is to train each unit to describe a function that converts the input to output data and vice versa. Instead of learning functions for each unit, functions could be learned for the space *between* each pair of units.

most obvious approach is to encode semantic properties of words by means of conceptual features, for example, representing an object's size and its color. However, it is hard to determine how many and what features are needed to obtain a fine-grained distinction between a large number of concepts, and many, more abstract, concepts are hard to reduce to features (e.g., *game*). A second procedure is to apply Li and Farkas' (Li & Farkas, 2002; Li et al., 2004) method to texts from one language only. This results in a language aspecific representation of semantic meaning. Instead of using different texts for each language, only the texts for, for example, English can then be used to represent all semantic properties. A third approach would be to take the distance between concepts in networks describing semantic relations, so-called semantic networks, as a measure of similarity. This method is used, together with text based measures, in the DevLex model (Li et al., 2004). Using texts or semantic networks both can offer a high resolution representation, with the semantic network being the most extendible, as long as any added words are also included in the semantic network. A disadvantage here is the relative unavailability of data. Only a few databases of sufficiently large semantic networks exist and often only for English concepts. The same is true for texts with sufficient semantic information to distinguish a set of words, balanced with respect to the amount of information available for each word, are not easily found. Moreover, semantic networks or texts in one language offer word meanings for that language only, implying that subtle differences in meaning between translational equivalents are not captured.

Of the three methods just discussed, the use of a sufficiently large semantic network offers the most flexibility and highest resolution. We opted for inclusion of the semantic network WordNet (Fellbaum, 1998), which represents the word meaning for about 150.000 English words. Because words from all languages are mapped onto the English meaning, this approach fails to take into account the differences in exact meaning between languages. Unfortunately, the Global WordNet project has not progressed sufficiently to allow WordNet based representations for all languages we are interested in (Vossen, 1998; Fellbaum & Vossen, 2007) and not all projects that are part of the Global WordNet project are freely available, otherwise these language-specific WordNets could have been used. Nevertheless, the lack of language-specific semantic representations is not expected to affect the results of the model in any way related to the characteristics of the human lexicon or the hypotheses regarding the effect of simultaneous learning.

3.1.5 Representation of language. To model word translation using language-specific orthography and language aspecific conceptual information, language membership information is needed. Otherwise, it would not be possible to proceed from language aspecific semantics to language-specific orthography. However, there is no consensus on whether language membership should be explicitly included in a model (French & Jacquet, 2004). On the one extreme, explicit language nodes are used that represent language activation and may bias the word selection process depending on context. This approach is implemented, for instance, in the first version of the

Bilingual Interactive Activation model (BIA) (Dijkstra & Van Heuven, 1998). However, as mentioned in section 2.1, empirical studies indicate that language context does not strongly affect lexical selection. An alternative method is to keep language membership as a completely implicit representation. This still allows word translation if both the orthographic and semantic representations contain enough information to keep languages apart. In the SOMBIP model, for example, the semantic and orthographic representations are language-specific, which implicates no language representation is needed (Li & Farkas, 2002). However, when a shared conceptual representation between languages is assumed, this is not feasible.

An intermediate approach is to represent language information, as required for the translation of words without context, but give it a low weight compared to orthography and semantics, resulting in a small effect on the translation but not enough to totally exclude words from other languages. A possible distributed representation is a bit-wise code with the length of the number of languages, i.e., a string of zeros for non-target languages and a one for the target language. Importantly, each representation of language membership should be unrelated to all others, because the languages are initially assumed to be unrelated. Any underlying language similarities and relations should be determined by the model itself and should not be predefined in the language representation. In other words, the distributed representations of the languages should be orthogonal.

3.1.6 Representation of orthography. With respect to orthography, it is important that letter identity, letter order, and possibly letter similarity are captured. The most biologically plausible and still rather efficient method for this purpose, compared to alternatives like position encoding, currently is using open bigram counts (Dehaene, Cohen, Sigman, & Vinckier, 2005). N-grams represent all sequential letter combinations of length n in a word, in the case of bigrams 2 (e.g., a bigram representation of TREE is _t, tr, re, ee, e_). Open bigrams are a generalization of bigrams and include all combinations of two subsequent letters in a word, with or without in-between letters. The more letters there are between the two letters of the bigram, the lower the value assigned to the bigram (e.g., an open bigram representation of TREE is tr, t_e, re, r_e and ee, where t_e and r_e have a lower count value, e.g., 0.6, while the rest has count 1). It is also possible to capture letter similarity using open bigrams, for example, by generalizing the activation on a bigram to bigrams with similar letters (e.g., activation from the bigram p_p generalizes to p_b). Because the method is seen as most similar to the one used in human cognition, it is also most likely to be the method resulting in human-like behavior.

However, preliminary tests showed that open bigrams did not result in correct orthographic maps for the dataset used and that many bigrams were needed to capture all the differences between words. Instead, we therefore chose to use orthographic edit distances between words (Damerau, 1964; Levenshtein, 1966). Thus, each word was represented by its orthographic distance to all other words. This approach allowed for fine grained distinctions, while the number of features could be reduced by using the distances to just a subset of words, because there is much redundant information in the edit distances to all other words. Letter position and letter identity are not directly captured using edit distances, but edit distance does allow a determination of the orthographic similarity between words. Letter similarity could also be captured by setting lower switching costs for more similar letters, but this proved not to be necessary for the purposes of this thesis.

Table 3
The choices made for the most important aspects of the model.

Aspect	**Choice**
Domain	Vocabulary learning, mapping orthography to semantics, modulated by language.
Model type	Distributed
Algorithm	Self Organizing Map
Representations	
Semantic	Language aspecific edit distances in WordNet
Language	Bit-wise numerical representation with a low weight compared to orthography and semantics
Orthography	Edit distances

3.1.7 Summary. In sum, the selected properties of the model were as indicated in table 3. This set of choices combines to a model with topographical representations in all layers, due to the SOMs, and with language aspecific semantic and orthographic representations, resembling the dominant view of human language processing (Dijkstra, 2005). Moreover, the model makes only a few assumptions with respect to the representations of languages, word semantics and word forms. The main assumption is that all three are represented in a distributed way. More specifically, we used a orthogonal representations for language, without any assumptions on language relatedness, and edit distances for word forms and semantics. Essentially, this mainly assumes that human cognition can assess similarity for both words and concepts, and not directly what features it uses.

3.2 Implementation

As was discussed in the previous section on the design, the model was implemented using the SOM algorithm. In the next section, this algorithm is briefly described. Furthermore, the structure of the model is discussed. All implementations were done using the SOM Toolbox for Matlab (Vesanto, Himberg, Alhoniemi, & Parhankangas, 2000).

3.2.1 Algorithm. The model was implemented using the Self-Organizing Map (SOM) algorithm, applied for autoassociative mapping. For more details than the short overview given here, the interested reader is referred to Kohonen (2001).

The SOM algorithm describes a way to represent multidimensional data on a lower dimensional, often two-dimensional, map. This is done by defining a grid of reference points and a metric that defines the distance from the reference points to the data points. Next, the reference points are updated iteratively or batch-wise to reduce the total distance between reference and data points. Reference points that are close together, learn together to develop and maintain the topological representation.

Formally, this can be described in the following way. The algorithm starts with a dataset containing vectors $x_k = [\eta_1, \eta_2, \ldots, \eta_n] \in \Re$, with n the dimensionality of the data. To model this data, a set of model vectors $m_i = \mu_{i1}, \mu_{i2}, \ldots, \mu_{in} \in \Re$ is defined at random, with n again being the dimensionality and i the number of the model vector. These model vectors, or units, all have indices defined by the topology of the map. For most purposes, a rectangular map is used, with the ratio between dimensions determined by the ratio of the two most dominant eigenvectors in the data. In such a rectangular map, the indices can be described by $r \in \Re^2$. For example, the first node on the second row would have index $r = (2, 1)$. Combined, a SOM is thus defined by a set of model vectors μ, with indexes r organized in a rectangular map, on which the data vectors x are projected.

Next, after random initialization of the reference vectors, the map can be trained in a sequential or a batched way. For sequential learning, the map is trained using the update rule:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[d(x, m_i)] \tag{1}$$

The last part of the formula describes the distance d between input vector x and model vector m_i. Usually, the standard Euclidean distance measure $d(x_i, m_j) = \|x_i - m_j\| = \sqrt(\sum (x_i{}^2 - m_j{}^2)$ is used. The degree to not only the winning model vectors, the Best Matching Unit (BMU) with index c, are updated, but also of neighboring units, denoted with index i, is defined by the neighborhood function h_{ci}. This neighborhood function is essential to develop or maintain the map topology, as it makes neighboring units learn in a comparable direction and thus represent comparable values. The neighborhood function defines what this influence looks like, for which often a Gaussian shape is used,

$$h_{ci}(t) = \alpha(t) * \exp\left(- \frac{\|r_c - r_i\|^2}{2\sigma(t)^2} \right) \tag{2}$$

where $\alpha(t)$ is a scalar-valued learning rate factor and the parameter $\sigma(t)$ defines the width of the neighborhood function. Both decrease over time t to allow for finer distinctions. In addition to the neighborhood function, neighborhood is also defined by the shape of the connections units make: These can be either square, only connections to the horizontal and vertical neighbors, or hexagonal, with connections to the horizontal, vertical and diagonal neighbors. The latter is less biased towards horizontal and vertical orientations in the map development and is often the shape of choice.

Batch-learning follows the same line of reasoning, except that all vectors in the input data are presented at once. This means that in batch-learning, after initialization, the model vectors are set to the weighted average of the input vectors in their neighborhood:

$$m_i(e+1) = \frac{\sum_{i=1}^{N} h_i(d_i)}{\sum_{i=1}^{N} h_i} \tag{3}$$

where d_i defines the distances of all N input vectors to node i, which is determined by comparing the data vectors x to the current reference vector positions $m_i(e)$, modulated by neighborhood $h_i(e)$. For clarity, t is replaced by e in the batched version of the formula, because instead of iterating over individual patterns, batched learning iterates over epochs.

The advantage of the batch version of the SOM algorithm is that it converges faster to an optimal solution. For the purpose of manipulating the learning scheme in multilingual learning, both types of learning could proof important though, as sequential SOM training resembles sequential concept-word presentations, while batch training is better comparable to simultaneous presentation of multiple associations. Preliminary tests pointed out, however, that batched learning did not work if not all patterns from the set are presented. Because, with only partial data, the model vectors change to the mean of only the presented part of the neighborhood, leading to the loss of representation of already learned patterns that are not included in the partial data. Details on how simultaneous learning was implemented instead follow in section 6.1.

The quality of a SOM is often determined using two measurements (Kohonen, 2001), the first based on the remaining error in the map and the second on the preservation of topology. The remaining error is called the average quantization error, calculated as the squared sum over the difference between the data vectors and the corresponding BMUs, or $\|x - m_c\|$. An often used measurement of topology quality is based on the fact that when the representation is topological, the reference vectors closest to a data vector should be neighbors of each other. This can be formalized by calculating the proportion of data patterns for which the two closest reference vectors are not adjacent on the map: the lower this proportion, the better the topology. Note that for large maps, the topology value is slow to decrease because the large number of nodes increases the chance that two nodes are not located next to each other.

As the SOM algorithm is in essence unsupervised, an alternative way is needed to make the network learn word-language-concept associations. This can be done by making use of the pattern completion abilities of SOMs, also called autoassociative mapping. For clarification, let us divide a data vector x into an input and output part, called x^{in} and x^{out}, which also results in an input and output part for the model vectors, respectively m^{in} and m^{out}. If a network is trained on the combined vectors x, representing both input and output, the model vectors learn to represent both the input and output side of the data. If the map is subsequently tested on only the input part x_{in}, which is compared to the m_{in} part, the same BMUs should be found as when the entire vector would be presented, as long as there is sufficient redundancy in the data. In the current application, there is such redundancy: if two of the factors orthography, language and semantics are known, the third is also uniquely defined. This means that an approximation of x_{out} (one of the three factors) can be found by looking at the output part of the winning node, m_c^{out}. It is an approximation because the winning node m_c is normally situated near vector x_{out}, not exactly on it.

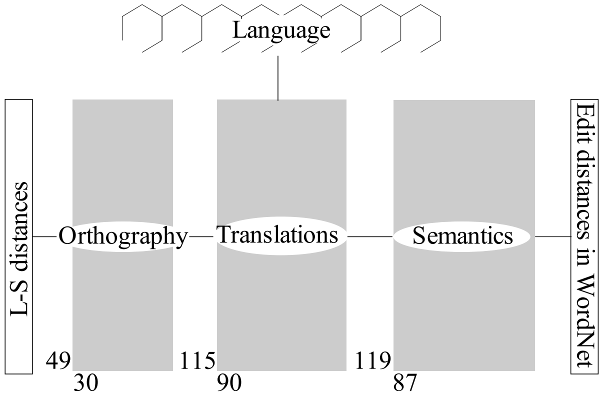


Figure 2. A schematic overview of the structure of the model. The model consists of four layers (orthography, translation, semantics, and language membership), of which the first three are SOMs. The numbers on the sides of each layer represent the number of model vectors, or nodes, along the length and width of the layers. The inputs to the orthographic and semantics layer are shown in the rectangular boxes on the sides. The layers are connected by lines, depicting that the output of one layer is used as *features* for the next map. For example, the Levenshtein-Schepens (L-S) distances are the features for the orthography map and the coordinates of the BMUs in the orthography and semantics SOMs, combined with the language information, are the features for the translation layer.

3.2.2 Structure. In the brain, semantics and orthography are stored in separate areas, with strong interconnections in-between (Münte, Heinze, & Mangun, 1993; Crosson et al., 1999; Tagamets, Novick, Chalmers, & Friedman, 2000). An analogous division in structure was used in the model, with a separate SOM for orthography and for semantics, plus a translation SOM in-between, which in turn was mediated by a language layer representing contextual language information. The complete model is shown in figure 2.

The orthography and semantics SOMs were both two-dimensional SOMs. The ratio between the two sides of the rectangular maps was chosen to roughly correspond to the relative length of the two dominant eigenvectors of the two datasets, which should facilitate topology development (Kohonen, 2001). For consistency and to improve resolution, we also used three times the number of words and concepts in the orthography and semantics map, as in the hidden map, resulting in 10353 (119 times 87) and 1470 (49 times 30) units in these two maps respectively.

The third factor, language, was represented as a one-dimensional layer with the number of units equal to the number of languages, with no topographical properties. This is not to say there is such a language representation in the brain; instead the incorporated language signal should be viewed as a contextual signal guiding the translation process and as such has no direct corresponding neural correlate.

These three layers, orthography, semantics, and language, were combined in another SOM, the hidden layer we call translation layer. The data vectors for the translation layer consisted of three parts (orthography, language and semantics) and instead of autoassociative mapping with only an input and output side, a three way mapping was used. In other words, after learning, any two parts of the hidden data vector should point to an unique BMU, so orthography and language should define the appropriate semantics, semantics and language the orthography, and semantics and orthography the language. The data used for the autoassociative mapping from the orthography and semantics SOM were the locations of the BMUs. So both SOMs received a pattern, which activated a certain BMU, of which the position was sent to the translation SOM and combined with the language representation to train the hidden layer. For this translation layer, the number of units was chosen to be three times the number of patterns, resulting in a 115 times 90 units map.

In total, a large number of units was used to represent all words and concepts. To be clear, we do not expect the brain to use such an inefficient representation. Instead, the nodes in the model should be viewed as 'resources': The more nodes are close to a word, concept, or relation, the better it is represented and hence known. In section 5.1 we use this rationale to define activation and subsequently reaction time measures which should make clear how this works out in practice.

Table 4
The most important properties of the data set.

Property	Value
Number of concepts	490
Number of languages	8
Number of words	3920
Average word length (SD)	5.77 (1.96)
Average frequency (SD)	96 (203)

3.3 Data

We used a dataset generously provided by Theophilos Vamvakos (Vamvakos, 2006) to both train and test the model. The original dataset contains translations for nouns in 13 languages, of which we selected all languages with the Latin alphabet, resulting in a selection of eight languages: English, Dutch, German, French, Italian, Portuguese, Spanish, and Catalan. In addition, nouns for which not all translations

Table 5
The number of homographs, cognates and false friends in the dataset for each combination of two languages. The values are depicted as homographs (cognates/false friends). Values on the diagonal represent homonyms within a language.

Languages	English	French	Italian	Spanish	Portuguese	German	Dutch	Catalan
English	0 (0/0)	32 (29/3)	1 (1/0)	6 (6/0)	4 (2/2)	20 (20/0)	22 (21/1)	7 (7/0)
French	32 (29/3)	8 (0/8)	4 (4/0)	12 (11/1)	10 (9/1)	11 (11/0)	10 (9/1)	41 (38/3)
Italian	1 (1/0)	4 (4/0)	4 (0/4)	57 (57/0)	61 (61/0)	2 (2/0)	2 (2/0)	35 (33/2)
Spanish	6 (6/0)	12 (11/1)	57 (57/0)	2 (0/2)	126 (126/0)	0 (0/0)	1 (0/1)	76 (75/1)
Portuguese	4 (2/2)	10 (9/1)	61 (61/0)	126 (126/0)	2 (0/2)	2 (2/0)	2 (1/1)	66 (63/3)
German	20 (20/0)	11 (11/0)	2 (2/0)	0 (0/0)	2 (2/0)	2 (0/2)	59 (57/2)	4 (4/0)
Dutch	22 (21/1)	10 (9/1)	2 (2/0)	2 (0/2)	2 (1/1)	59 (57/2)	4 (0/4)	5 (2/3)
Catalan	7 (7/0)	41 (38/3)	35 (33/2)	76 (75/1)	66 (63/3)	4 (4/0)	5 (2/3)	12 (0/12)
Total	92 (86/6)	128 (111/17)	166 (160/6)	280 (275/5)	273 (264/9)	100 (96/4)	105 (92/13)	246 (222/24)

were available were removed and characters with an accent were converted to the non-accentuated characters. Articles were also removed, because we were interested in recognizing words from different languages in the absence of such strong cues. In total, this left 490 concepts in 8 languages, totaling 3920 words.

For the semantic representation of each concept in the dataset, the distance to all other concepts in the dataset was derived from WordNet (Fellbaum, 1998) using the distance rule as proposed by Lin (1998) and implemented by Greenwood (2007). WordNet contains the semantic relations such as hypernyms, hyponyms, holonyms, and meronyms and the lexical categories of about 150.000 English words. The distance measure as developed by Lin is calculated by dividing the number of common semantic properties of two concepts by the total number of properties of the two concepts. This means the value always ranges from 0 to 1. Applying this distance measure results to the current dataset resulted in 490 values between 0 to 1 (mean: 0.07, SD: 0.12) as a distributed representation of each concept.

For the orthographic representation, each word in the dataset was converted into a sequence of edit distances to all words. The edit distance was calculated using the Levenshtein formula (Levenshtein, 1966), which calculates the minimal number of operations required to change one word in the other. Operations taken into account are additions, deletions, and substitutions. Alternatively, the Damerau-Levenshtein distance could be used (Damerau, 1964), also including transpositions. However, transpositions are not thought to induce neighborhood effects, though recent evidence suggests otherwise (Acha & Perea, 2008). We extended the distance formula by applying normalisation, as proposed by Schepens (2008). To keep the number of dimensions in bounds, only the distances to a selection of 490 words out of all words were used, resulting in edit distances ranging from 0 to 26 (mean: 5.90, SD: 1.82) for all 3920 words.

Lastly, for the language representation, a bit-wise representation was used, with the first bit representing English, the second French, etc. resulting in a string of 8 bits, the number of languages, with one being true. This representation is orthogonal, as required by the design (see section 3.1).

These data were combined with the word frequency information found in the CELEX database for English words (Baayen, Piepenbrock, & Gulikers, 1995) to also account for frequency effects. The frequencies, ranging from 1 per 100,000 to up to 1971 per 100,000 (mean: 96, SD: 203), were reduced to ten bins with an equal number of patterns. The bins were numbered 1 to 10, with the bin number representing the number of times the total pattern (orthography, language and semantics) was presented to the network in the training phase. We also tried to use the frequency from CELEX directly as the frequency of presentation. This worked to some extent, but due to the large variation in frequencies the training time increased considerably, because the low frequent patterns were presented too little to be learned. With more training time available, this would be a good option though. For now, possible effects of frequency should be testable using this simplified measure of frequency. We will refer to this binned frequency as the frequency of a word for the rest of this thesis, although it only really roughly corresponds with the actual word frequencies.

Correct topography development and convergence in the SOM algorithm is helped by normalization of the data, in order to make sure all components in the data have the same influence (Kohonen, 2001). We normalized each feature for orthography and semantics to have equal variance. The data for language we left unchanged, because we wanted it to have a lower impact than the other data vectors.

Preliminary tests using this data pointed out a problem though. Using the representations for both orthography and semantics proved computationally complex due to the high dimensionality of both the data and model vectors. Luckily, both the orthography and semantics representations contained a significant degree of redundant information, due to the interrelatedness of edit distances: The edit distance from A to B and A to C is also informative about the edit distance from B to C. The redundancy allowed us to use a subset of 100 out of the 490 edit distances as features, while this did not influence the map development significantly.

More detailed properties of the dataset, such as word lengths and number of homographs, are shown in table 4 and 5.

3.4 Model training

Prior to training, the model was initialized randomly within the ranges defined by the data, because there is no a priori reason to expect an ordered start of the human lexicon. Next, the network was trained using sequential learning (see section 3.2) and the default parameter values for such a network (Kohonen, 2001). This meant the starting value for the neighborhood width was half the width of that particular SOM and the learning rate started at 0.5, both decreasing linearly over the total number of trials, which was 100 for the complete model. The learning rate decreased to 0, the neighborhood radius in the orthography and semantics SOM to 1 and in the translation layer to 0. Decreasing the SOM neighborhood radius to 0 in the translation layer was done to ensure development of localist representations. Afterwards, a finetuning session was done, starting with a learningrate of 0.05 and a neighborhood of 1. Over another 100 trials, the learning rate again reduced to 0, while the neighborhood was constant for the orthography and semantics SOM and decreased to zero for the translation SOM.

Table 6
An overview of all the tests performed on the model.

Test	Subtest
Qualitative properties	
	Map structure
	Language-specificity
	Homograph representation
Quantitative properties	
	Monolingual frequency and neighborhood effects
	Homograph effects
	Neighborhood effects
	Language information effect
Effect of learning scheme	
	Sequential learning
	Mixed learning
	Simultaneous learning

For the specific tests as mentioned in the next section and described in detail in the following chapters, two additional versions of the main model were trained on subsets of the data, as shown in table 7. This was done to ease comparison with experimental data. The sizes of the maps were scaled appropriately for the decreased number of patterns, which decreased training time without influencing results. All other properties of the model remained the same.

3.5 Model tests

The model was tested in three ways, as listed in table 6 and described in the subsequent chapters:

Table 7
The three versions of the model. The column called 'Model' shows the name of the model as it is referred to in the text. 'Data' shows which languages were included and 'Epochs' how many trials the model was trained during the rough and finetune phase. 'Proficiency' shows whether there was an imbalance in the proficiency for the different languages.

Model	Data	Epochs	Proficiency[a]
Monolingual	English	$2*1000$	Balanced
Bilingual	English, Dutch	$2*1000$	Imbalanced $(1/5)$
Multilingual	All eight	$2*100$	Balanced

[a] Proficiency was modified by presenting one language more often than the other. The numbers show the multiplier for the frequencies of the words for each language, if applicable.

- Qualitative tests, focused on the structural validity of the model.
- Quantitative tests, comparing the performance of the model to reaction time data from behavioral experiments.
- Learning tests, testing the effect of different learning schemes on the speed of language acquisition.

Details on the tests used and results found are given in each chapter separately.

All analyses mentioned in these chapters were done using either one of two methods. When the test involved the difference between groups, a two-sided unpaired t-test was used. When it involved quantitative variables, multiple linear regression was applied. In this case, the β-values are reported as slope values and the p-values for the β-values are given, as well as the F, p and R^2 values of the total effect. In either case, the significance border was taken to be .05, while less than .1 was considered marginally significant. All analyses and plots were done in Matlab (Mathworks, 2008).

4. Qualitative test of model validity: internal structure

To reiterate, we built a model of the human multilingual lexicon to predict whether simultaneous language learning is beneficial, compared to sequential language learning. The model was implemented using SOMs and it was trained to learn to convert concepts to words and vice versa in eight languages.

Next, two types of tests of model validity were performed. First the qualitative validity of the model was tested, as described in this chapter. Three aspects were considered qualitative properties of the model. The first was the translation performance, or how well the model translated and which alternatives it considered. The second qualitative property was the representations that developed in the maps contained in the model. More specifically, the degree to which the model was either language-specific versus aspecific was determined. Thirdly, we analysed the representations for cognates, false friends and non-homographs to see whether shared or separate representations were used.

4.1 Translation performance

After training, the first of the three qualitative tests focused on the translation performance of the model, letting the model translate sets of words from one language into another language and checking for the activation of intra- and interlingual neighbors. Translations were modeled by two passes through the network: first from orthography and language membership to semantics (e.g., STONE and English to `stone`) and then from semantics and the target language to orthography (e.g., `stone` and Dutch to STEEN).

After training, the multilingual version of the model showed a quantization error of 0.54, 3.48 and 4.24 on the semantics, translation and orthography SOM respectively. The topology error was 1, meaning that the topology was not yet fully learned, though the large number of units makes it hard to decrease this value much. With these scores, the model succeeded in learning the associations from words to concepts and vice versa for all eight languages to some proficiency, with 784 of the 3920 words converted to the right concept and 491 conversions from concept to the right word. This also allowed the model to translate from one language to another by applying the two conversions successively, resulting in 2736 correct translations or about 45 of the 490 correct translations per language pair in one direction.

Although a success rate of a only 11% is small, note that this was mainly due to the short training time, only 100 epochs, while several thousand epochs is deemed normal for randomly initialized SOMs (Kohonen, 2001). This implies the model should be considered a not fully proficient learner yet. A second source of increased error rate is the unbalanced frequency distribution, even with the altered frequencies (see section 3.3). The average frequency of the successful translations was 9 (SD: 1.6) , which differed significantly from the overall average of 5.5 (SD: 2.9) ($p < .001$). This means the lower frequency ones were not presented frequently enough to become learned to the full extent and the patterns presented more frequently pulled too many of the model vectors towards them, dominating map development. This fits the picture of a low proficient language learner, mainly knowing high frequent words. A third source of increased difficulty for the model, causing longer training times, was the small influence of the language membership signal, leading to language errors. Lack of time kept us from training the model further, though older training sessions indicated that further learning improved translation performance, as well as map and reaction time results, but did not qualitatively change any of the results reported here.

The correct and incorrect translations allowed us to look at the alternatives the model considered and the kind of errors it made. Both alternatives and errors should be orthographically and/or semantically related. Moreover, the orthographic alternatives/errors should be language aspecific, just as found experimentally (see section 2.1). This was indeed what was found for most of the cases, as shown for a number of examples in table 8. More elaboration on the neighbors considered and the effect of neighbors follows in section 5.3. So even though the model made a large number of errors, the errors made sense to a large extent. This probably also explains why the results on other tests did not differ qualitatively when the model was trained further in older tests: The model developed a global structure on the basis of similarity, of which mostly the details changed with extended learning. These first results also showed that the representation that the model developed was probably integrated across languages, since alternatives of all languages were taken into account.

4.2 Map properties and language-specificity

The second series of tests focused on the structural properties of the model by visualizing the maps that arose. Note that the structure within the maps was not set by hand, but developed throughout the learning process. The tests specifically focused on the effect of language in the model: To what extent did language-specific or aspecific representations develop in the three SOMs?

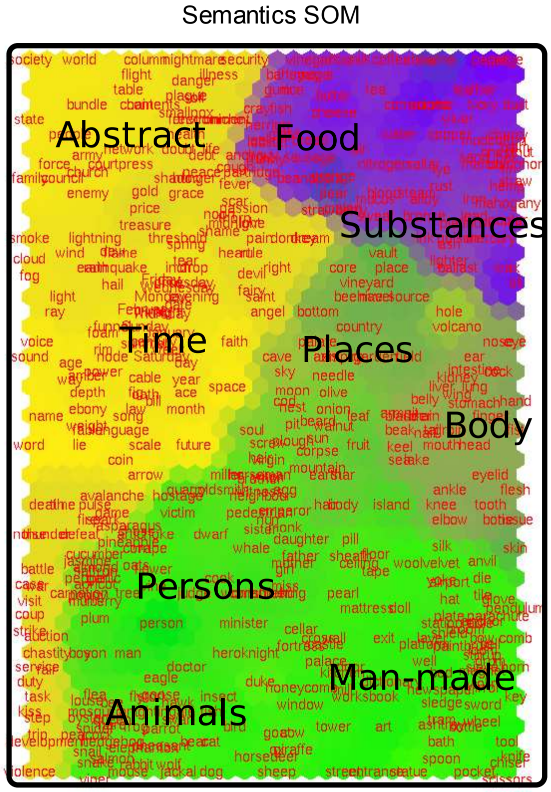


Figure 3. The semantics SOM, with all concepts in red and names for clusters in white. Concept names were automatically applied to the BMUs of the patterns, while clusters were interpreted by hand. Each hexagon represents one unit and the colors indicate similarity: a large shift in color means a large shift in similarity.

The semantic map received language aspecific input on the similarity of the concepts, meaning no language-specificity was to be expected there. This is indeed

Table 8

Example translations as performed by the model, showing the conversion from a word and a language to a concept and from a concept and another language to a word, including the activated alternatives. Both correct and incorrect translations are shown. Note the interlingual nature of the considered orthographic alternatives and the orthographic and semantic nature of the errors.

Word	Language	Concept	Alternatives	Language	Word	Alternatives
Correct						
FRIEND	English	friend	doctor, king	French	AMI	FAIM, PAI
KOPF	German	head	mouth, tooth	Italian	TESTA	PESTA, ESTAT
AREIA	Portugueese	sand	mud, dust	Dutch	ZAND	WAND, HANF
DICCIONARIO	Spanish	dictionary	wall, castle	Catalan	DICCONARI	DICTIONARY DIZIONARIO
Errors						
STOMACH	English	mouth	brain, kidney	Dutch	MOND	MONK, MOON
WOMAN	English	friend	dwarf, father	French	CONSUMENT	CONSUMER, CONSUMATORE
WERKEN	Dutch	service	auction, kiss	Spanish	SERVICIO	SERVIZIO, SERVICO
WINE	English	tea	cheese, coffee	French	THEE	THE, CHEESE
DIMANCHE	French	day	night, week	German	MONTAG	ZONDAG, MONAT

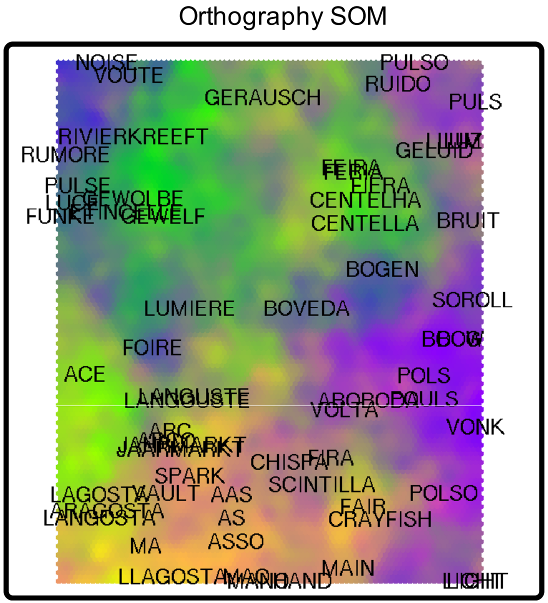


Figure 4. A graphical depiction of the orthography SOM, using the same representation as used in figure 3. Only a small subset of words of a few languages is shown for clarity, in black. Visible is the organization by orthographic similarity (e.g., AAS, AS and ASSO, as well as PULSO, PULS, RUIDO are close together) and the lack of organization by language (e.g., the words for BOW and BOOG, as well as LIGHT and LICHT are on top of each other).

what we found: The map was purely organized on the basis of semantic similarity. A number of categories could be identified in the structure, namely humans, man-made objects, substances, body parts, food, places and abstract concepts, as shown in figure 3. Some concepts were falsely localized (for example, some foods between the places), showing the network did not settle fully yet. Overall, the structure is internally quite consistent, which can also be judged by the homogeneity of color within regions, and was confirmed in reruns.

In contrast to the semantic map, the orthographic map did receive different inputs for each language. Still there was no explicit language membership information, only implicit information in the language-specific orthography. This did not turn out to be a significant factor though: The map that developed for orthography was organized on orthographic similarity only, irrespective of language membership. However, the map turned out to be less well organized than the semantic SOM, shown by regions with less homogeneous colors. Neighboring patterns were not always the most similar ones, which older tests point out does improve with additional training. This decreased organisation did turn out to cause some problems with the quantitative tests though, as will be considered in chapter 5.

To analyse the representation in greater detail, we looked at where words for the different languages were located on the orthography map. Figure 5A shows the result of this analysis, where for each language the activation spots on the orthographic map are shown. Clearly, it is an integrated representation, as there are no definable regions for any of the languages, but words for all languages are scattered all over the map. The number of common hits between languages, i.e., equal nodes activated by words for the same concepts in different languages, confirms this picture (see table 9). If the representations were totally separate, no shared representation should exist, while this table shows a large number of shared representations. When one compares these values with the homograph information in table 5 the same pattern can be seen, though the number of common hits overestimates the number of homographs. The latter seems to be due to two reasons. First, especially in this low proficient version of the model, the high frequent words function as attractors, meaning that high frequent words are activated instead of words in their neighborhood. This increases the number of common hits, because multiple

Table 9

The number of times the word for a particular concept was represented by the same unit in the orthography SOM and the mean distance between words in different languages for the same concept. A higher number of equal common units or a lower distance represents a higher language similarity. Note the similarity with table 5.

Languages	**Catalan**	**Dutch**	**English**	**French**	**German**	**Italian**	**Portuguese**	**Spanish**
Catalan		20 (11.4)	25 (2.8)	54 (5.4)	15 (11.9)	69 (0.5)	88 (0.8)	108 (1.3)
Dutch	20 (11.4)		40 (13.3)	24 (17.7)	37 (4.7)	14 (12)	8 (11.8)	10 (12.3)
English	25 (2.8)	40 (13.3)		49 (1.2)	25 (9.2)	17 (3.2)	17 (3.5)	21 (4)
French	54 (5.4)	24 (17.7)	49 (1.2)		21 (10.3)	16 (5.7)	24 (6)	26 (6.5)
German	15 (11.9)	37 (4.7)	25 (9.2)	21 (10.3)		10 (12.2)	7 (12)	7 (12.5)
Italian	69 (0.5)	14 (12)	17 (3.2)	16 (5.7)	10 (12.2)		86 (0.6)	83 (1)
Portuguese	88 (0.8)	8 (11.8)	17 (3.5)	24 (6)	7 (12)	86 (0.6)		160 (0.7)
Spanish	108 (1.3)	10 (12.3)	21 (4)	26 (6.5)	7 (12.5)	83 (1)	160 (0.7)	
Total	379 (34.1)	153 (83.2)	194 (37.1)	214 (52.8)	122 (72.9)	295 (35.2)	390 (25.3)	415 (38.3)

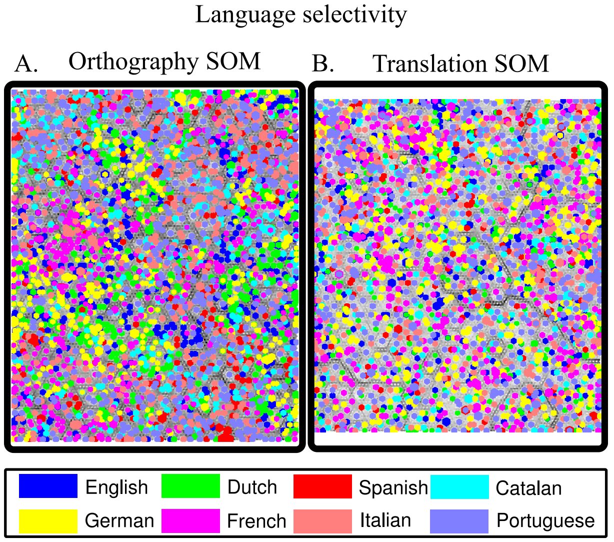


Figure 5. Representation of language-specificity in the orthography (**A**) and translation (**B**) SOM. Each dot indicates a node activated by a certain language (see legend), its size indicating the number of times it was activated for that specific language. This clearly shows the lack of organization by language: The colors are intermingled over the entire map. The colors are overlayed, so the last added languages are overrepresented in the pictures compared to the first presented ones. The order in the legend shows the order of presentation.

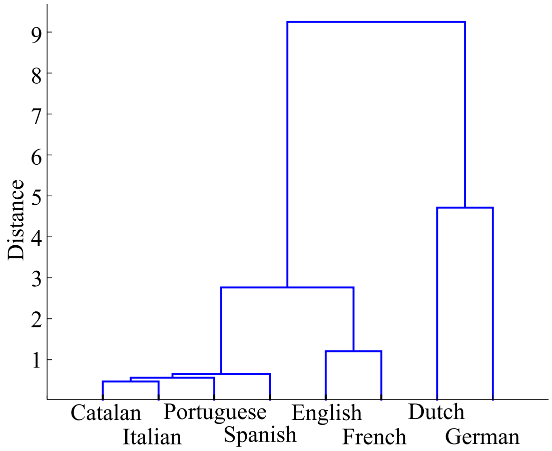


Figure 6. A dendogram of language similarity in the orthography SOM, based on the distances for words for the same concept as shown in table 9. The vertical axis reflects the distance and on the horizontal axis the languages are shown. The shorter the path along the blue line from one language to the other, the more similar they are.

words end up at the BMU of the high frequent patterns. Second, near-homographs, which are not taken into account in table 5, can activate the same BMU if the similarity is large enough.

To get a higher resolution impression of the similarity of the languages, the distances between words for the same concept across language are also shown in table 6. Languages which are similar should have low distances for translational equivalents. The distances showed that the orthography SOM correctly reflects language similarity, as illustrated in figure 6, which confirms the results by Schepens (2008). Combined, the common hits and distances show that the orthographic SOM is organized on orthographic similarity, not on language information, though language similarity is correctly reflected in its organization.

The translation layer was the only layer in the model that did receive language membership information, meaning that language-specific representations could arise there. Still, as the semantic and orthographic representations turned out to be language aspecific and should be dominant compared to the language membership information (see section 3.1.5), an organization on the basis of semantics and/or orthography seemed more likely. The latter indeed proved to be the case: The translation layer was primarily organized on semantic and orthographic similarity, with only a minor effect of language. To scrutinize the latter effect, the same analysis was done as was done for the orthography SOM (see figure 5B). Again, no identifiable language-specificity was found.

To quantify the effect of language membership

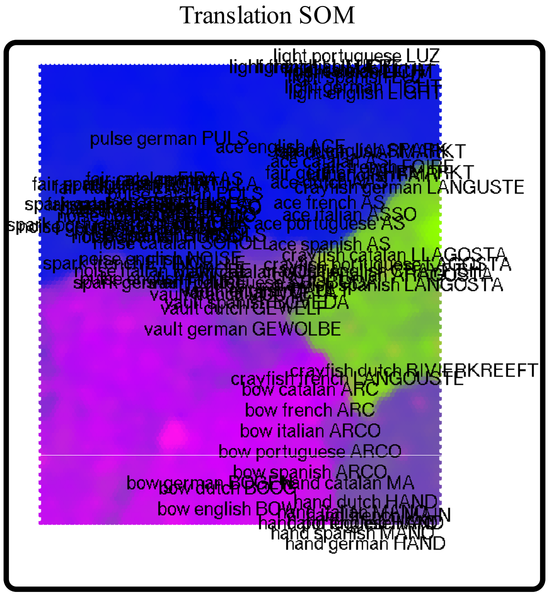


Figure 7. An overview of the translation SOM, with a small number of example translations using the same representation as 3 and 4. Data in the translation SOM consists of both a semantics, language membership, and orthography part. The organization as can be seen on the map is primarily on semantics and orthography. For example, look at the words for bow, of which especially two small groups exist: one with orthography like ARC and one with orthography resembling BOW. The same is true for the words for `crayfish`, with the Dutch translation at a different location than most of the others, except one, which is falsely positioned near the Dutch translation. Note that patterns straight beneath each other mean they are located at the same BMU.

information on the translation layer, we compared the lengths of the language, semantics, and orthography parts of the data vectors in the hidden layer, as calculated by the dot product of the data vector. The lengths of the three parts of the data vectors are directly related to the average distance these parts have to the model vectors. To see why, first notice that the coordinates of model vectors are within the ranges defined by the data. This means the part of the coordinates of a model vector which depend on orthography are within the range of the orthography values, the part depending on language are within the range of the language values, and the part depending on semantics are within the range of the semantics values. Next, take into account that when patterns are presented to the translation layer, the distance between the data and model vectors over all these three parts is calculated. It is this total distance which subsequently determines BMU selection. The larger the distance for a specific part, the more influence it has. The critical point is that the smaller the data and model vectors for a particular part are, the smaller the possible distance between the two and hence the smaller their influence on the calculation of the total distance. The language layer part of the hidden data vectors was expected to have the shortest distance and thus the smallest effect. It turned out the ratio between the three parts, semantics, language and orthography, was 13/1/12, confirming the minor role language played compared to the other two factors.

4.3 Homograph encoding

The last qualitative aspect we tested was the representation of homographs and non-homographs. The representation of these word types is interesting, because they have both language-specific and language aspecific properties. Cognates share both meaning and orthography (e.g. the English-Dutch homograph BED), implying they should be represented in the same way across languages in both the semantic and orthographic SOMs, while a small language-specific difference can be expected in the translation layer due to the effect of the language membership signal. False friends should differ in the semantic layer, while being equal in the orthography layer and somewhat divergent in the translation layer (e.g. the English-Dutch homograph ROOM). Non-homographs on the other hand should have zero distance in the semantic SOM and an on average large distance in the orthography SOM.

To test the representation of different item types, we first calculated the distance between words in the data. The distance was calculated for semantics, translations (semantics, language and orthography) and orthography. On the basis of these distances, three groups were defined: non-homographs (larger than 0 distance for orthography and 0 distance for semantics), cognates (0 distance for orthography and semantics), and false friends (0 distance for orthography and larger than 0 distance for semantics). Next, the distances for the items in these groups were also calculated in the model, but this time not the distance between data vectors was calculated, but the distance between the BMUs that were activated after applying the data to the model. A distance of zero reflected an integrated representation, while non-zero distances represented distinct representations, allowing to test the representation the model used, compared to the actual distances as present in the data. The expectation was that the model would develop efficient representations, using as few nodes as possible and thus use shared representations (distance zero) where possible. Moreover, if the model successfully captured the properties of the different types of homographs, the distances in the model should mimic the distances in the data.

The result of this analysis is shown in figures 8A and B. Comparing the two panels makes it apparent that the model correctly captures the properties of the three word types as present in the data. Cognates turn out to have integrated representations for both orthography and semantics, reflected by zero distances between the representations, while false friends only share their representation for orthography. In the translation layer, cognates hardly differ, while false friends differ to a degree directly related to the size of the semantic difference. Non-homographs show the opposite pattern of false friends, with shared semantics, but different orthographic representations. Homographs thus have no

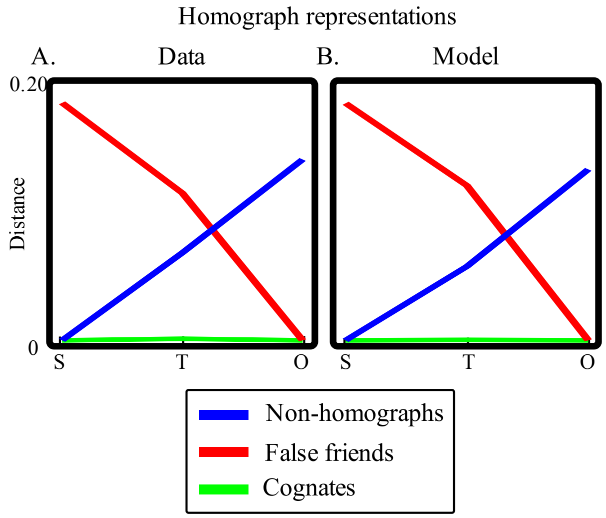


Figure 8. Average distances between semantics (S), orthography (O) and the complete translation (T) in both the data (**A**) and the model (**B**). For the data, these distances were calculated between all data vectors. For the network on the other hand, the distances between the activated units for each pattern were used. The mean distances are shown for non-homographs (blue), false friends (red) and cognates (green) respectively.

distinct status in the model, but are represented quantitatively different from non-homographs. The bottom line is that the model used shared representations where possible and the distance between representations reflected the similarity in the data.

4.4 Summary: an integrated lexicon

In this chapter the qualitative properties of the SOMMUP model were analysed. It turned out the model was able to translate to some extent between the eight languages, making sensible errors and activating related neighbors in the process. Moreover, it did so using a totally integrated lexicon, as nowhere in the model language-specificity could be found: The concepts were shared across languages, the orthography for all languages ended up intermixed on one map, and the language effect on the translation map proved to be only minor. This also resulted in shared representations where possible, meaning that cognates used shared representations for both orthography and semantics, while false friends used a shared orthography representation across languages. Combined, even though the model needs further training, it already incorporates the most important aspects of the human lexicon, as described in section 2.1. The implications of these findings and a comparison to the human lexicon and another model is described in the general discussion (section 7.2).

5. Quantitative test of model validity: experimental findings

As discussed in section 2.1, there are three lines of experimental evidence that together form the basis of the integrated view of the human lexicon: homograph studies, neighborhood studies, and the effect of prior language membership information. The series of tests considered in this chapter was aimed at a comparison between the performance of the model and these experimental findings. These tests essentially tried to determine whether the structural properties of the model as reported in the previous chapter also leads to the behavioral properties thought to be induced by these same structural properties in humans.

As mentioned in section 3.4, two additional networks were trained for these tests. The first represented a monolingual speaker and contained only the 490 English words of which it correctly understood 271 and could 'produce' 274 after 2000 trials. Quantization errors were 0.52, 1.63 and 2.48 in the semantics, translation, and orthography SOM respectively. This version of the model was used to test the general activation and reaction time effects (see section 5.1). The second model was intended to represent a bilingual lexicon with an asymmetric proficiency: English and Dutch words were trained with a relative frequency of 1 to 5, which resulted in 53 correct translations from Dutch to English and 88 from English to Dutch after 2000 trials. The quantization error values ended up to be 0.42, 2.63 and 6.08 for the three SOMs. This model was used to test for differential neighborhood effects in specific and generalized lexical decision tasks (see section 5.3). The complete, eight language model was used to test the effect of multiple times cognates (see section 5.2).

For all of these tests, first a measurement of reaction time had to be defined, which was subsequently used with the aim to replicate the experimental findings as found in humans.

5.1 Definition of reaction time

Three properties of reaction times in human language understanding are important for this purpose (Grainger, 1990) and relevant for the intra- and interlingual effects we try to incorporate in our model. First of all, human reaction times show a clear frequency effect: High frequent words are recognized faster than low frequent words. Secondly, the neighborhood density of a word has an influence on the reaction times: The more neighbors and the closer the neighbors are, the more influence. The direction of this influence is task-dependent. For example, for a language-specific lexical decision task, intralingual neighbors lead to facilitation and interlingual neighbors to inhibition, while in a generalized lexical decision task both lead to facilitation. Thirdly, the neighborhood effect is modulated by frequency: High frequent neighbors have more influence than low frequent ones.

We incorporated these aspects into a reaction time measure in a three-step process. First, we defined a measure for the activation of patterns (word, concept or language) and their neighbors within a SOM. Secondly, we defined how the flow of activation progressed and finally we converted these activation measures to reaction times. The complete procedure used to determine the reaction times is shown in figure 9.

5.1.1 Activation of within-SOM alternatives. Representations involving activations of words, concepts and languages are often used in models of the lexicon and are thought to represent the likelihood that a word is selected (Dijkstra & De Smedt, 1996). Hence, the more evidence there is, for example, for a certain word, the higher its activation. This evidence can be direct, meaning that a specific word is activated, or indirect, due to similarity/neighborhood effects from another pattern that is activated. Often, frequency effects are linked to the activation metaphor as well: The more frequent words are, the less evidence is needed for words to become activated. This implies that both the neighborhood and frequency effects on reaction times should already be incorporated in the definition of the activation measure. To model activations we thus needed to determine a measure for both similarity and frequency in the model.

Similarity is encoded automatically in SOMs, because the distance between nodes represents the similarity between the associated patterns: Words or concepts near the presented pattern on the map should receive more activation than words and concepts further away. To quantify this, the distances between the BMU of a word and the BMUs of all possible alternatives were determined. These distances should correctly reflect the degree of similarity *as represented by the network*. Alternatively, the measure could have been based on the distances between words *in the data*. However, if modeled in this way, the model and its 'proficiency' would not matter, because the distances could be calculated without reference to the model. With sufficient training, the similarity in the model should however converge to the similarity in the data.

To represent the frequency effect, we chose to use the distance of a pattern to its BMU, because the more often a pattern is presented, the closer it ends up to its BMU. To see the rationale behind this approach, view a SOM as there being strings between a pattern and the nodes, which patterns can use to pull nodes towards them. Now, if a pattern is presented more frequently, it gets to pull more often, resulting in the BMU ending up close to the pattern.

Mathematically, the effect of similarity and frequency can be combined into a measure of the activation of pattern j given the presentation of pattern i as follows:

$$\alpha_{i,j}^{freq} = 1 - norm(log(d_{x_j,BMU_j})) \tag{4}$$

$$\alpha_{i,j}^{sim} = \exp\left(\frac{-(1 - norm(d_{BMU_i,BMU_j})^2}{2\sigma^2}\right) \tag{5}$$

$$\alpha_{i,j} = \alpha_{i,j}^{freq} * \alpha_{i,j}^{sim} \tag{6}$$

The first function calculates the effect of frequency, the second of similarity and the third combines the other two into the total activation. Both the frequency effect and similarity effect depend on the function d, representing the Euclidean distance either between the neighboring pattern (x_j) and its BMU or between the BMUs of two patterns. For the frequency effect, as the distance of a pattern to its BMU is small and the value approaches 0 like an inverse logarithmic function, we converted it to a more linear range by applying a logarithmic function. Afterwards, to make sure both effects would have about the same influence, both distance measures were normalized to a $[0 - 1]$ range and inverted, as shown in the formula. For the neighborhood function, we afterwards applied a Gaussian function to the similarity measure with a σ value of 0.05, because only close neighbors are thought to be activated. Note that the activation measure is also applied for the actually presented target ($j = i$). In this case, $\alpha_{i,j}^{sim}$ will be 1, because the BMUs are equal, but $\alpha_{i,j}^{freq}$ still has an effect. The further the target is away from its BMU, the less activation it will receive, representing a frequency effect for the target pattern.

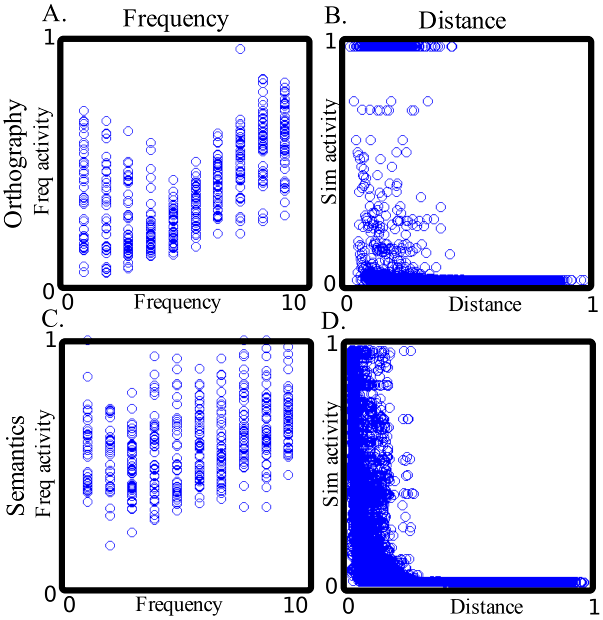


Figure 10. The relation between frequency of the target pattern and its activation (panels **A** and **C**), as well as the relation between the similarity of a neighbor in comparison to the presented pattern and the activation (panels **B** and **D**), both for orthography and semantics (upper and lower panels respectively). The similarity is based on the similarity *in the data*. The results shown are for the monolingual network. These patterns are representative for the other versions of the model and other languages.

To verify that the activations were indeed related to target frequency and the similarity between target and neighbors, we plotted the relation between frequency and α^{freq}, as well as between the distance from a pattern to its neighbors and α^{sim} in the monolingual version of the model (see figure 10). This indeed showed the expected effects in both the orthography and semantics layer. There was a positive effect

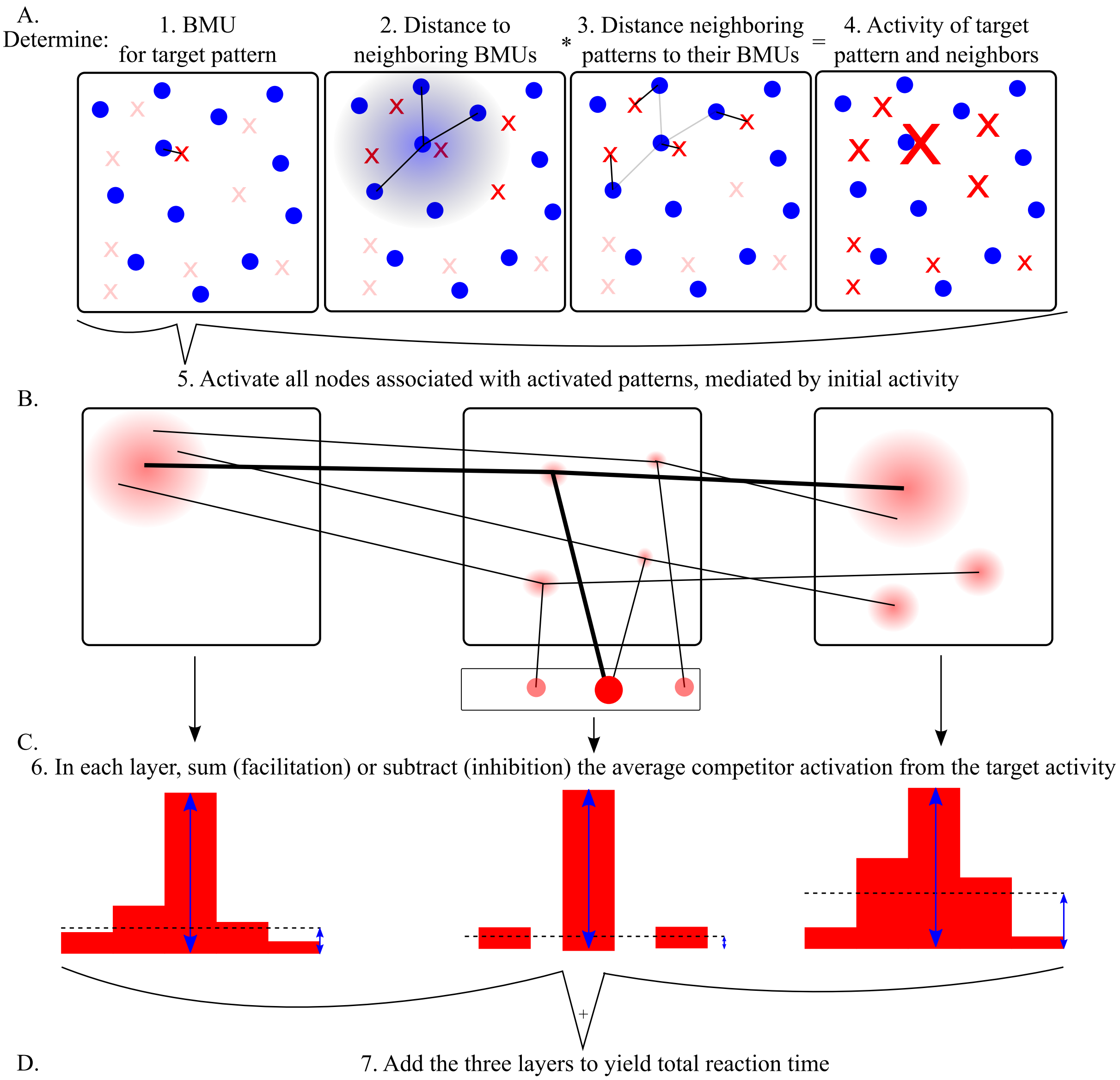

A.
Determine:
1. BMU for target pattern
2. Distance to neighboring BMUs
*
3. Distance neighboring patterns to their BMUs
=
4. Activity of target pattern and neighbors
5. Activate all nodes associated with activated patterns, mediated by initial activity
B.
C.
6. In each layer, sum (facilitation) or subtract (inhibition) the average competitor activation from the target activity
+
D.
7. Add the three layers to yield total reaction time

Figure 9. A schematic overview of the calculation of reaction times. Red crosses indicate pattern, blue dots depict model vectors which are possible BMUs. **A.** In each layer to which a pattern was presented, the within-SOM activations were calculated. This was done in three steps. First the BMU for the pattern was found. Next, the distance from the BMU to the BMUs of neighboring patterns was determined and, using a Gaussian function, converted to a similarity measure. In addition, the distance between both the target pattern and neighbors to their BMUs was calculated and converted to a measure of the frequency effect. The product of frequency- and distance-based activations determined the total activation of neighboring patterns. **B.** All patterns which received activation in this way, also activated the corresponding node in the translation layer and subsequent layers, though the change in activation was modulated by the initial activation. This way, activation of both the target pattern and the neighbors reached all layers. **C.** For each layer, the result of the activation flow determined an activation profile. The activation of the target pattern was combined, using either summation (facilitation) or subtraction (inhibition), with the activation of its neighbors to result in a reaction time measure per layer. **D.** The sum over the reaction time measures per layer resulted in the total reaction time.

of frequency on α^{freq} (slope: .37, $F(1, 488) = 262.27$, $p < .0001$, $R^2 = .35$) and a decrease (slope: $-.21$, $F(1, 488) = 33.15$, $p < .0001$, $R^2 = .064$)) in activation α^{sim} the further away neighboring patterns were, which on visual inspection had a Gaussian-like shape. In the orthography layer results, the Gaussian is less well-shaped and more words are seen as equal, shown by an activation of 1, than should be the case. This is probably due to the less well evolved structure of the orthography layer. The product of the frequency and similarity-dependent activation as shown in figure 10 defined the total activation α in one layer for each pattern in response to any pattern presented.

5.1.2 Flow of activation. Only calculating activations within SOMs is not enough however, as there are effects which are dependent on combinations of factors, which can be modeled by a flow of activation from one layer to another. The most prominent example is the language-dependent orthographic neighborhood effect, such as the differential effect of intra- and interlingual neighbors in a language-specific lexical decision task (Gainger & Dijkstra, 1992; Heuven et al., 1998). Because there is no information on language membership within the orthography layer, language membership had to be determined elsewhere. One possible way is to determine the language of neighbors in the same way as was done for the target. This means passing the activation of both target and neighbors from the orthography layer to the translation layer and subsequently to the language layer, which then represents language membership activation of both targets and neighbors.

The flow of activation was processed as follows. For each pattern on all involved input SOMs and the translation SOM, the within-SOM activations were calculated. These values were used as initial activation values $\alpha^{i}_{j,i}$ and sender activation $\alpha^{s}_{j,i}$ for the activated input layers. For the hidden layer, the within-SOM activation formed the initial recipient activation $\alpha^{r}_{j,i}$. Next, the pattern itself and the neighbors activated the translation layer. This was calculated by multiplying the recipient and sender activations, which meant the activations in the translation layer were modulated by the activations in the source layer. In other words, the less activation a neighbor received, the less activation it got to send to its corresponding BMU in the translation layer. The sum of activations induced by a pattern and all its neighbors in this way, over all inputs involved, resulted in a total activation in the translation layer. As it was then the turn of translation layer to send activation to the other layers, this total activation formed the sender activation $\alpha^{s}_{j,i}$ of the translation layer. The nodes in the translation layer activated their corresponding nodes in the other layers, for which a recipient activation $\alpha^{r}_{j,j}$ was calculated on the basis of the output location. This output recipient activation was multiplied by the total activation in the translation layer, resulting in the total activation in the output layer. This process can be described by the following equation:

$$\alpha^{r'}_{j,i} = \alpha^{i}_{j,i} + \sum_{s}^{S} \sum_{j}^{N} (\alpha^{s}_{i,j} * \alpha^{r}_{j,j}) \tag{7}$$

Where $\alpha^{r'}_{j,j}$ represents the total activation of the pattern j in the layer which is in this case recipient activation. Variable $\alpha^{r}_{j,j}$ represents its old activation, $\alpha^{i}_{j,i}$ its initial activation and $\alpha^{s}_{j,j}$ the activation of the sending layer. Note that only $\alpha^{r}_{j,j}$ is used, meaning that the neighbors a pattern could activate in the receiving layer are not included in the calculation of the activation. For each pattern, the sum over all sending layers and over the product of activations of all patterns in the receiving and sending layer (where N denotes the number of patterns), results in the total activation in the receiving layer. This was done first for the translation layer, taking the activations in all three layers (numbers of layers S is 3) into account, and then for the output layers, adding the newly found activations in the translation layer to the already determined initial within-SOM activations.

5.1.3 Converting activations to reaction times. Next, using the activation measure, the reaction times could be determined in a way comparable to the BIA+ model (Dijkstra & Heuven, 2002). Actually, the activations as calculated by formula 4 to 7 should be interpretable by the task module of the BIA+, because that module interprets the same kind of activations as produced by our model. Two possible ways of converting activations to reaction times are defined in BIA+ and other models: subtraction and addition, representing inhibition and facilitation respectively. Inhibition is relevant when the alternatives are actually competitors to the target pattern; while facilitation is applicable when the other patterns could also be regarded as correct targets for the task. Compare for example a specific lexical decision task, in which all languages different than the target language are competitors, with a generalized lexical decision task, where all languages are equally correct. In-between, it is also possible to model no effect of neighbors by not taking the activations of neighbors into account.

To calculate task-dependent reaction times, we applied the following formula:

$$RT^{T}_{i} = 1 - norm\left(\alpha_{i,i} + \sum_{\substack{j \\ i \neq j}}^{N} sim^{T}_{i,j} * \alpha_{i,j}\right) \tag{8}$$

In words, this formula states that the sum of the activation of the neighbors is combined with the target activation. Subsequently it is normalized and inverted in order to make all reaction times fall within a $[0 - 1]$ range. Whether the combination is facilitation or inhibition, represented by addition and subtraction, depends upon the similarity variable $sim^{T}_{i,j}$, which represents the *task T dependent* similarity of the neighbor, being any value between -1 and 1. A *sim* score of -1 results in inhibition from the specific neighbor, while a 1 results in facilitation. In-between values are appropriate when a neighbor can have a certain degree of task-dependent similarity, for example in semantic priming where semantic primes have a positive or negative influence depending on the degree of similarity (e.g. Neely, 1976, for a review, see Neely, 1991). For the purposes of this thesis, we only used the extremes, the −1 and 1 values, because for the difference between specific and generalized lexical decision no gradual similarity is required.

This measure of reaction time could be applied to each layer, after which the sum over the reaction times for the semantics, orthography and language layer yielded the total reaction time. All layers were used, instead of specific layers depending on the task, because we expect lexical processing to be automatized to such a degree that all factors play a role in all tasks.

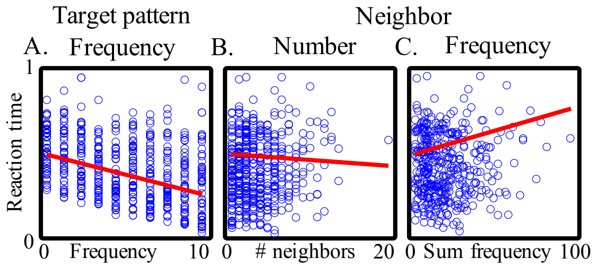


Figure 11. Mean reaction times over all words depending on the frequency of a word (**A**), distance to all neighbors within range (**B**), and frequency of these neighbors (**C**). Each blue circle represents one word, the red lines show the fits using all three factors in a multiple regression analysis.

Before continuing to the actual tests of experimental results, we first tested whether the measure of reaction times we defined was correct in general. This was done by testing the relation between frequency and neighbors on the one hand and the reaction times on the other. We did so in the monolingual version of the model, tested on a lexical decision task, in which neighbors are expected to facilitate. To quantify the effect, we tested the relation between the three components in the data on the one hand and the reaction times generated in the network on the other. Only the effect of orthographic neighbors is discussed, though the effect of semantic neighbors is comparable. These tests confirmed that frequency of the target, frequency of the neighbors, and the number of neighbors influenced reaction time significantly ($F(3,386) = 33.93$, $p < .0001$, $R^2 = .17$). Reaction time decreased with increasing frequency (see figure 11A) (slope: −.25, $p < .0001$). In addition, reaction times increased with the number of neighbors (slope: .29, $p < .1$), as it should (see figure 11B and C). There was no significant effect of the frequency of neighbors however (slope: −.07, $p > .1$).

Though the total explained variance, .17, is rather small, please note the intrinsic non-linearity involved in the reaction times. This non-linearity is caused by the Gaussian neighborhood function and taking the product of neighbor frequency and distance. In addition, the total reaction time is also influenced by the effects in other layers due to the summation over the three layers. More importantly, the similarity relation is based on the similarity in the data. It turned out the model had not yet developed the same order in neighbors as present in the data. If we did the same analysis using the neighbors as found by the model, the R^2 increased to .26 ($F(2,487) = 82.42$, $p < 0001$) and the slopes all became significant with values in the expected directions (−.24 for frequency and .16 for neighbor activation, both $p < .0001$). This means that with increased learning, which causes the model to learn the correct neighborhoods, the result would become better.

5.2 Homograph effect

We tested the homograph effect in the bilingual version of the model trained on two languages to a different extent, representing asymmetric proficiencies. More specifically, we tested both specific and generalized lexical decision, because the effect of homographs are different for the two tasks (see section 2.1). To mimic a language-specific lexical decision task, the words were presented together with information on language membership. For the generalized lexical decision, no language information was given. The important difference was that in the specific lexical decision task, activations in the language layer were assumed to be competitive, and thus inhibitory, while in the generalized lexical decision task they were considered facilitatory. The expectation was that on average, in both tasks, L2 words would be recognized slower than L1 words. In the specific lexical decision task, false friends were expected to be in-between the reaction times for L2 and L1 and cognates should be comparable to L1. In the generalized lexical decision on the other hand, false friends should be closer to L1 and cognates should be recognized even faster.

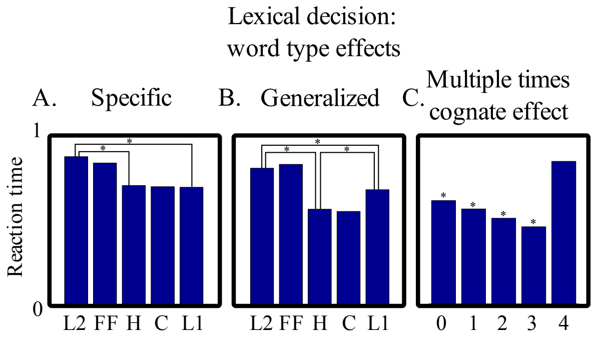


Figure 12. Average reaction times of the model for L2 controls (L2), false friends (FF), homographs (H), cognates (C) and L1 controls (L1), in both a specific and generalized lexical decision task (**A** and **B**), showing faster reaction times for homographs and L1 controls compared to L2 controls. Panel **C** shows the effect of the number of languages a cognate is present in. 0 indicates words being no cognate at all, 1 indicates one time cognates, hence present in two languages, etc. * indicates significance at least at the .005 level.

The relation between, on the one hand, the number of times a word is a cognate, number of times it is a false friend and the proficiency, and on the other hand, reaction times, turned out to be significant in both the specific lexical decision ($F(3,976) = 64.78$, $p < .0001$, $R^2 = .17$) and generalized lexical decision case ($F(3, 976) = 142.04$, $p < .0001$, $R^2 = .30$). Further analysis of this significant effect showed a number of interesting results. First of all, there was a clear proficiency effect in the reaction times: reaction times for L2 words were significantly higher than for L1 words in both the generalized and specific lexical decision tasks (proficiency dependent slope: $-.12$ and $-.17$, $p < .0001$). Secondly, there was a mean reaction time difference between generalized and specific lexical decision tasks (.72 and .76, $p < .0001$). Thirdly, there was a significant difference in reaction times for homographs (including cognates) and L2 words in the specific lexical decision task (slope: $-.19$, $p < .0001$), but not with L1 words ($p > .1$). In contrast, in the generalized lexical decision task, also the difference with the L1 words became significant as the reaction times for cognates further decreased ($p < .0001$). Lastly, there was no significant effect of a word being a false friend in both the specific and generalized lexical decision task ($p > .1$) due to the small number of false friends in this subset of the data. In total, this means the effect of cognates actually became smaller between the two tasks ($-.18$ to $-.14$), but the effect of proficiency increased (from $-.12$ to $-.17$). The pattern in results, with a reaction time difference between L1 and L2 and cognates being recognized about as fast as L1 words in a specific lexical decision task, but faster in a generalized lexical decision task, confirms the results as reported by Lemhöfer and Dijkstra (2004).

As a second analysis of the homograph effect, we tested the model trained on eight languages on the cognate effect to see whether it mattered whether a word was a cognate in more than one language pair, as has been found experimentally (Lemhofer, Dijkstra, & Michel, 2004). We confirmed that being a cognate also lead to facilitation in the eight language network. More importantly, the two times and three times cognates showed an even greater facilitation (slope: $-.19$, $F(1, 3918) = 131.2777$, $p < .0001$, $R^2 = .04$). No conclusion can be drawn for the four times cognates, as only 1 was present, compared to 235 one time, 70 two time and 17 three time cognates. When tested for significance using a t-test for differences between reaction times for the subset of cognates and the total dataset, it indeed turned out the results for one, two and three times cognates were significantly different ($p < .001$), which was not the case for the four times cognates ($p > .1$). The explanation why the reaction times appear to rise for the four times cognates is the fact that the frequency of that specific cognate is low: The word was ANANAS (meaning `pineapple`), which had a CELEX frequency of 3, implying that it belonged to the category with the lowest number of presentations.

5.3 Neighborhood effect

In order to see whether the model could also explain neighborhood effects as found in human participants, we tested the effect of both intra- and interlingual neighbors in the bilingual model. First, we looked at the general effect of both types of neighbors, assuming competition in the orthography and semantics layer but no effect of neighbors in the language layer. Next, we tried to specifically test the difference in effects of neighbors in a specific versus generalized lexical decision task.

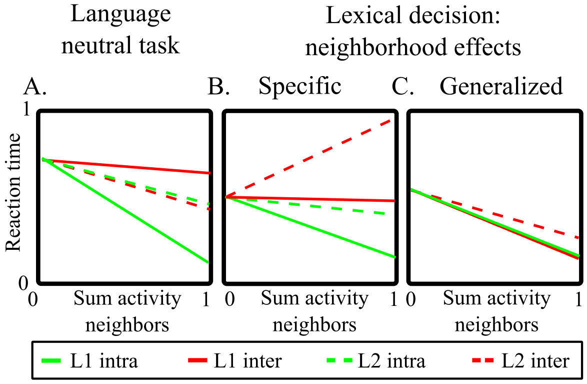


Figure 13. Graphs showing the effect of neighbors in general, without language competition (**A**), as well as the differential effect of neighbors in the specific lexical decision task (**B**) compared to the generalized lexical decision task (**C**). The x-axis shows the normalized sum of activation of neighbors, the y-axis the reaction times. Separate lines indicate the results for L1 (Dutch, straight) and L2 (English, dashed) and intra- and interlingual neighbors (green and red respectively). The sum of A and B corresponds to an approximation of the reaction times for a specific lexical decision task and the sum of A and C to the reaction times for a generalized lexical decision task.

Overall, both types of neighbors had an effect ($F(5, 484) \geq 1248.31$, $p < .0001$, $R^2 \geq .93$) (see figure 13A) and mostly in the same direction ($-.59$ for intralingual neighbors on L1 recognition, .056 for L2 neighbors on L1 recognition, $-.16$ of L2 neighbors on L2 recognition and $-.59$ of L1 neighbors on L2 recognition, $p < .05$ for all). The direction and magnitude of the effect of neighbors depended on the language proficiency, as L1 neighbors had a larger effect than L2 neighbors both within and between languages. This is line with experimental results (Heuven et al., 1998).

When we zoomed in on the differential effect of intra- versus extralingual neighbors in the specific versus generalized lexical decision task, we were confronted with the wrong neighbors being selected by the model, just as mentioned earlier for the reaction times. To counter this, we again used the neighbors as found by the model instead of the neighbors in the data, which should be the result after sufficient training of the model. Using this approach, the reaction times could be explained with a R^2 of at least .62 in the L1 specific, L1 generalized, L2 specific and L2 generalized lexical decision case ($F(5, 484) \geq 158.58$, $p < .0001$). More specifically, on the one hand we found intralingual neighbors to facilitate in the specific lexical decision task, with a slope of $-.36$ for L1 and $-.099$ for L2 (both $p < .0001$). Interlingual neighbors on the other hand induced no change in the L1 case (slope: -.02, $p < .05$) and inhibition in the L2 case (slope: .47, $p < .0001$). In contrast, both types of neighbors were found to mostly facilitate in the generalized lexical decision task: slopes of −.40 and −.41 for L1 intralingual and interlingual neighbors and −.36 for L2 interlingual neighbors (all $p < .0001$). In all cases there was proficiency modulated the neighborhood effect, leading to no facilitation effect of L2 intralingual neighbors in the generalized lexical decision task ($p > .1$)

5.4 Effect of language information

The effect of information on language membership was assessed by testing whether the model interpreted false friends differently with or without knowledge on what language to expect. More specifically, we wanted to see whether the model preferred the low frequent meaning of a false friend in the cued language, or the high frequent meaning in the non-target language. For example, if the model needs to determine whether ANGEL is a Dutch word, it could fail to do so successfully due to the fact that the English reading is more frequent and thus more easily activated, irrespective of the availability of language information. To do so, we tested the model on the 65 false friends differing in frequency only, with and without input on the language membership layer. As quantitative measure of correctness, we used the distance between the output and the actual meaning in the semantic layer for the two languages: The smaller the distance for one meaning, the more likely the model was to select that meaning instead of the other.

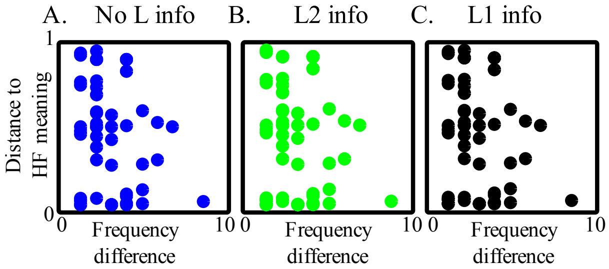


Figure 14. The effect of language information on the meaning detected for false friends. Each circle represents one false friend with differing frequency, with the actual difference between the high and low frequent meaning depicted on the x-axis. The y-axis shows the distance to the high frequent meaning, with 0 implying correct recognition of the high frequent meaning and 1 meaning correct recognition of the low frequent meaning. Panel **A** shows the results without language membership information, **B** with L2 membership information, and panel **C** with L1 membership information.

It turned out there was no significant effect of language information ($p > .1$): With and without language information, the same meaning of false friends was activated (see figure 14A). Concerning the effect of frequency, the number of false friends with large enough differences in frequency proved to be too small to detect a significant difference ($p > .1$) (see figure 14B and C). The model should probably be trained further and ideally with more false friends and real frequencies for the latter test to have real value.

5.5 Summary: replication of experimental findings

To summarize, even though the performance of the model requires improvement, the model captured almost all the tested patterns in reaction times. The first pattern replicated in the model was that reaction times decrease with word frequency and proficiency. Secondly, both intra- and interlingual neighbors influenced the reaction times, depending on the number and language proficiency of the neighbors. This effect of neighbors depended on the kind of tasks, with opposite effects for intra- and interlingual neighbors in a specific lexical decision task and only facilitatory effects in the generalized version of the task. Thirdly, also word type effects were explained, with lower reaction times for cognates compared to non-cognates, decreasing even further for multiple times cognates. The fourth kind of effects, related to false friends, proved not be testable due to the small number of false friends in the dataset and the high error count in the model. In total, the results are promising and further training and testing should be done. What the next steps could be specifically, is described in the General Discussion (section 7.2).

6. The effect of learning scheme

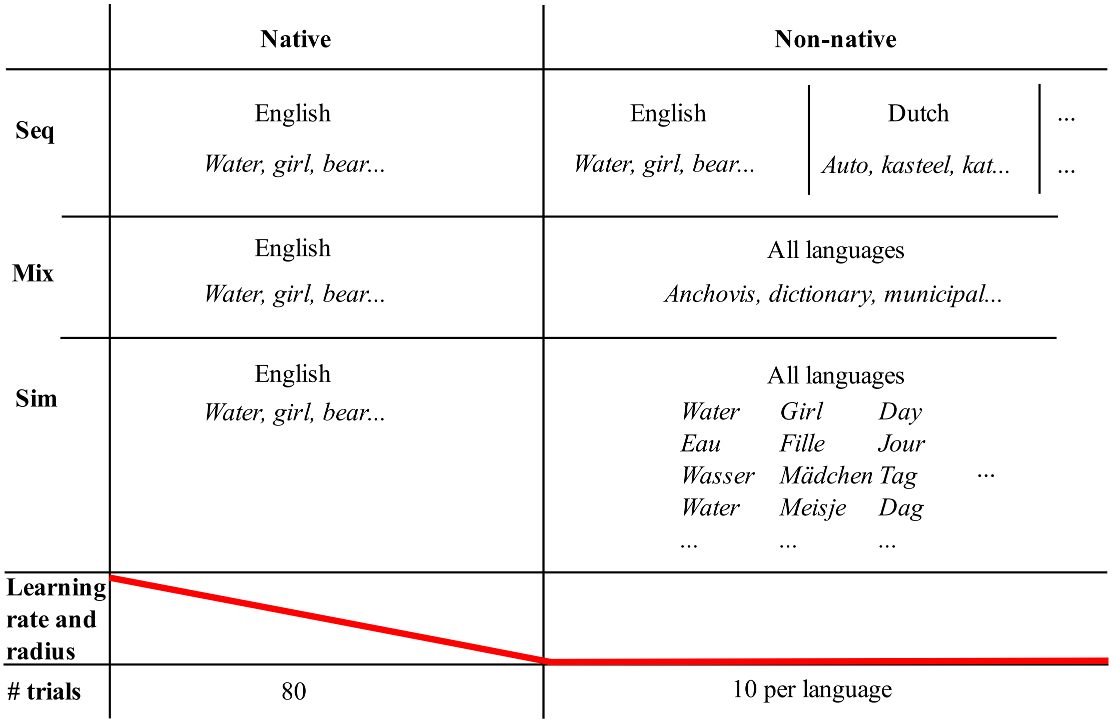
With the validity of the model confirmed, we continued by testing the effect of learning scheme, in order to determine which hypothesis concerning simultaneous language learning is the most likely to be correct (see table 1).

6.1 Definition of learning schemes

To test the learning schemes in the model, the three different ways of late language learning were operationalized as follows:

- Sequential learning: First the native language was trained, then the first foreign language, followed by the next, until each language was trained.
- Mixed learning: Words from all foreign languages were presented intermixed in a random fashion.
- Simultaneous learning: Words from all languages for one concept were grouped and presented together. The groups were presented randomly.

As the focus is on late learning, prior to the learning schemes the models were trained on a 'native' language, which was taken to be English. This training was done for 80 trials using a high learning rate and radius. Next, the late learning phase consisted of 10 trials per language using a low learning rate and radius, as in the finetuning phase as mentioned in section 3.4. As such, the number of presentations of each word and each language was balanced over the three schemes. The difference in learning rate and radius was done to model the difference between early and late learning, where in late learning the learning rate is thought to be lower due to decreased plasticity and the presence of already known language(s) (O'Reilly & Munakata, 2000).



	Native	Non-native		
Seq	English *Water, girl, bear...*	English *Water, girl, bear...*	Dutch *Auto, kasteel, kat...*	... ...
Mix	English *Water, girl, bear...*	All languages *Anchovis, dictionary, municipal...*		
Sim	English *Water, girl, bear...*	All languages *Water Girl Day* *Eau Fille Jour* *Wasser Mädchen Tag* ... *Water Meisje Dag*		
Learning rate and radius				
# trials	80	10 per language		

Figure 15. The three tested learning schemes ('Seq': sequential scheme, 'Mix': mixed, 'Sim': simultaneous), including examples of trials. The native part is the same for all schemes. Only English words are learned in the native phase and it starts with a high learning rate and radius, which decreases over time (depicted by the red line). This phase was trained for 80 trials. The nonnative, late learning part differed between the schemes. In the sequential scheme, the native language was repeated and then followed by the other languages. In the mixed scheme, words from all languages were presented intermixed. In the simultaneous scheme, words for the same concept were presented together. The nonnative phase took 10 trials per language and used a low learning rate and neighborhood radius. '. . . ' indicates the same type of trials followed as shown as examples.

To quantify the results, total quantization error and map-specific quantization error were measured. This provided an indication of the rate of acquisition, both overall and on the different aspects of language learning: semantics, translations, and orthography. The latter was done because for the different schemes, also different effects could arise on the three parts. Simultaneous learning for example is expected to facilitate orthographic discriminations, while semantically it does not necessarily have to be beneficial. To consider the acquisition process in greater detail, the model was tested after each 5 trials on the number of correct translations from native language to foreign languages and vice versa. Again, we expected possible differences between the models on the proficiency for the two directions of translation. The combination of overall measures and more specific measures allowed us to study both the overall and specific proficiency effects of the learning schemes.

When comparing the sequential learning scheme with the other schemes, two important aspects turned out to be unbalanced which would probably influence results. The first unbalance was that the sequential learning scheme was more likely to forget previously learned languages. This was caused by the fact that the sequential scheme presented languages only during a certain period, after which the languages were not repeated anymore, while the other learning schemes repeated all the languages all the time. To counter this possible unbalance, we took the maximum number of correct translations over the training sequence for each language as a second, corrected, performance measure. This meant that for the sequential learning scheme, the last trial within a language-specific block was selected, while in the other schemes one of the latter trials overall would be chosen. This performance measure actually favored the sequential learning scheme, because using that scheme it was possible for the model to focus on one language during a block, while forgetting the others. Still, we regarded this direction of unbalance to be fair in light of our hypothesis that simultaneous language learning is beneficial compared to sequential learning.

The second imbalance however worked against the sequential type of learning. This imbalance was cause by the fact that the native language, which was used as source or target language in all translations, was only learned during the native phase and the first sequential phase and afterwards not anymore. Also training the model on the native language during the sequential steps would have corrected the problem, at the same time increasing comparability with the natural situation. It would however also further facilitate the sequential learning scheme and make the number of presentations for the native language unbalanced over the three schemes. In the end, this imbalance was not corrected, which turned out to induce some difficulties.

6.2 Performance on different learning schemes

The results of the tests described in the previous section, focused on the effect of the different learning schemes, are shown in figure 16. Before considering the general and specific performance differences between the three schemes, some general effects that can be seen in panel A and B are noteworthy. Firstly, performance started near zero (on average on 2.13, SD: 2.83), even though the native training period was performed beforehand. This means at least native language performance was expected to be better, while it actually showed on average only 8.67 correct translations at the start. This low starting performance turned out to be due to the finetuning phase being necessary for actual correct performance. The fast rise in proficiency of the native language directly after the start supported this view. Secondly, for all three learning schemes, it was clear the native language was learned best: On average over the entire time course it showed 58.00 correct translations, compared to 19.03 for the other languages ($p < .0001$). Using sequential learning however, the proficiency of the native language did drop over time (from 104 to 51), due to the fact it was not trained anymore after the first epoch, as was mentioned as a possibly unbalance. Thirdly, there was a clear difference in the pattern in translation performance over time between the sequential scheme on the one side and mixed and simultaneous schemes on the other side. The results for the sequential scheme were marked by ups and downs of performance corresponding to the moments of presentation of the languages in the sequence. In the other schemes, there was hardly a difference in performance between the non-native languages. Finally, related languages seemed to 'help' each other in the sequential learning process, which makes sense given the larger overlap between these languages. This can for example be seen when Spanish was presented, which also induced an increase in proficiency for Portuguese and the same was true for English and French (refer to tables 5 and 9 and figure 6 for an overview of the language similarities).

Next, we turned to the comparison between the learning schemes performance-wise, of which the results are shown in the right column of figure 16. What was immediately striking was the large discrepancy between the corrected and uncorrected performance scores of the sequentially trained network: The corrected score for translations from the native language to the foreign languages was 626, compared to 152 uncorrected, and for the opposite translations the values were 589 and 132 respectively. If we compared the corrected performance values, the performance of the sequentially trained network was far better than the other two schemes (on average 607.5, compared to 383.5 and 438.5 for the mixed and simultaneous trained network respectively). If one instead focused on the uncorrected values, the sequentially trained network performed the worst of all three (142, compared to 360.5 and 438.5). For the other two schemes, there was not such a large difference between the schemes, though overall the simultaneous language learning seemed to perform better than mixed learning.

A comparable pattern, though inverted, was present in the quantization errors (see figure 16C). A small difference was the fact that the simultaneously trained network had a higher quantization error value than the mixed network (99.59 and 91.02 respectively), while its translation performance was actually better. Still, the problem of opposite rankings due to the large difference between corrected and uncorrected scores for the sequential learning scheme was also present in the quantization errors, making it hard to answer the question on the overall effect of learning scheme.

The second question was whether also specific proficiency difference arose for the different learning schemes. This indeed seemed to be the case, both on the basis of the performance scores and the quantization errors. The first difference was a difference in performance between the two directions of translation for the three learning schemes. The sequentially trained network performed better in translations from the native language to the foreign languages (626 and 589), while the simultaneously trained network preferred translations from foreign languages to the native language (395 and 482). A smaller difference in the same direction was present for the mixed network (378 and 389). The meant the outcome of the learning process was different in the three learning schemes, though this difference was small. The difference in effect for the sequentially trained network might well be due to the lack of presentation of the native language. Further tests are thus needed to really draw conclusions on the basis of these results.

A second specific difference in proficiency was present in the quantization error in the three parts of the network. Overall, the pattern was the same for all three learning schemes, with the lowest error for the semantic SOM, the second lowest for the translation SOM, and the highest value for the orthography SOM. If we compared the specific error scores between the networks, small differences were visible. The sequentially trained network for example showed the lowest error scores on the semantics (4.29, compared to 6.00 and 9.22 for the mixed and simultaneous network), while the other two error scores for the sequentially trained network were equal or higher than the scores in the other versions of the model (63.91 for semantics, compared to 68.42 and 67.40, and 127.66 for orthography, compared to 106.64 and 122.84). The simultaneously trained network did not show the expected facilitatory effect for the orthography SOM, but rather performed equally or worse on all three aspects compared to the other models.

Where does this leave us on the question of the effect of simultaneous language learning? The answer is, due to the non-interpretability of the results, as of yet undecided. The tests were not balanced enough, resulting in large differences between the corrected and uncorrected error values for the sequentially trained network. These differences were too large to draw any conclusions, because it made the pattern in results opposite. Nevertheless, the analysis did show that mixed and simultaneous language learning can be performed,

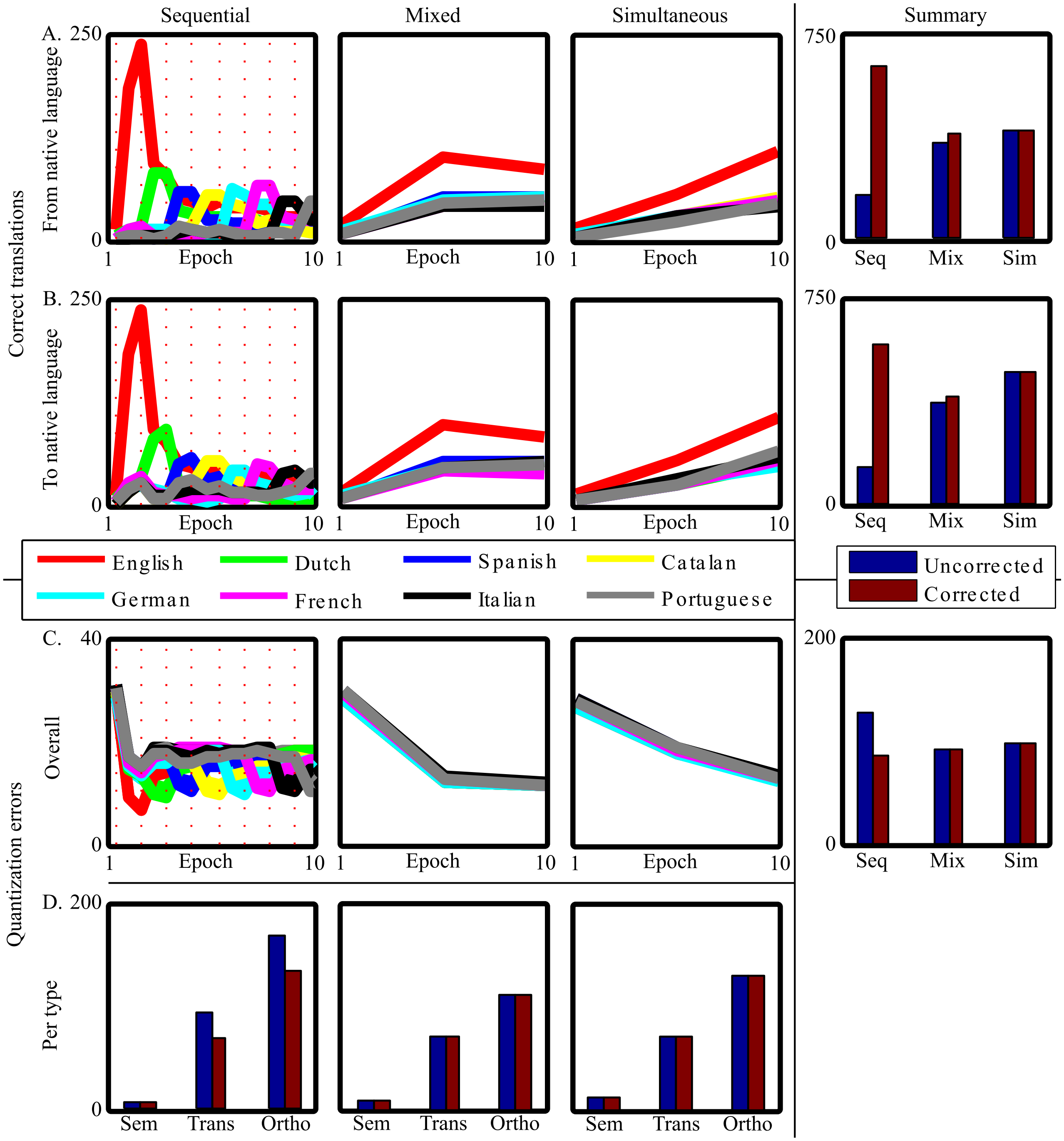


Figure 16. Graphs showing the results for the three different learning schemes for late learning. The first three columns show the results for the sequential, mixed, and simultaneous scheme respectively. The fourth column shows the summary of the first three, comparing the uncorrected (dark blue) and corrected (dark red) performance scores for the three different learning schemes. For the line graphs, the x-axis depicts the number of epochs per language, starting at the end of the native learning period. In the sequential case, the number of epochs per language actually corresponds to 80 epochs, as all languages were presented sequentially. Due to the difference in number of epochs, less errors values were recorded for the mixed and simultaneous learning scheme. The red dashed vertical lines indicate the start of the presentation of a new language. The colors of the other lines indicate the performance for translations to and from specific languages. Performance on translations from the native language to the other languages is depicted in panel **A** and from the foreign languages to the native language in panel **B**. In panel **C**, the quantization errors summed over all SOMs for each language are shown. For a more detailed analysis on the level of the separate SOMs, panel **D** shows both the corrected and uncorrected quantization errors summed over the languages in the three different parts of the model, orthography (ortho), translation (trans), and semantics (sem).

at least in the model. Moreover, the results showed hints of differences between the three learning schemes in what is learned, both in translation direction preference and the relative proficiency on the main constituents of vocabulary learning as concerned by the model: orthography, translations, and semantics.

6.3 Summary

In this chapter we focused on the effect of three different learning schemes on the rate of language acquisition in the model. This implied that the model was trained in three ways: sequentially, presenting one language after the other; mixed, presenting all languages intermixed; and simultaneously, presenting words from all languages grouped by concept. Two imbalances in the learning schemes proved to complicate accurate comparison. The first imbalance was that in the sequential learning scheme previously learned languages were forgotten, which was not the case in the other schemes. The second imbalance was the fact that the native language was not trained anymore after the first few epochs in again the sequential learning scheme. When corrected for the first imbalance, the sequential learning scheme fared best of all schemes. Without the correction however, performance in the sequential scheme was the worst of all three schemes. It was not clear which of the two measures can be regarded the correct one. In addition to the differences in overall performance, there were also signs of differences on specific aspects of the performance. More specifically, the three schemes seemed to differ in the performance for forward versus backward translations and for the performance on semantics, translations, and orthography separately. In conclusion, all three learning schemes were feasible in the model and showed effects on the rate of acquisition, but no well-founded and convincing comparison could be made. What the implications are and what the next steps should be in order to more effectively show the effect of simultaneous language learning is discussed in the next chapter, specifically in section 7.1.

7. General Discussion

The purpose of this thesis was twofold. The main objective was to show whether it is beneficial to learn languages simultaneously instead of sequentially. We tested this using a modeling approach, for which a valid model of the multilingual lexicon and language learning was required. This constituted the second objective. In the end, only the second objective was reached, while more work is required in order to reach the first objective. In short, the proposed SOMMUP model as proposed in this thesis showed both the same kind of structural properties as thought to be true for the human lexicon, as well as output comparable to experimental data in humans. However, tests of the effect of learning scheme in the validated model failed to show convincing results.

What the implications of the results are and what still can or needs to be done is discussed in this chapter. First the prime focus, the effect of simultaneous language learning, is considered, followed by model validity. After this discussion of the main results, extensions and improvements are proposed, as well as a possible practical application in the domain of Computer Assisted Language Learning.

7.1 The effect of simultaneous language learning

The first goal, to determine the effect of simultaneous language learning on speed of language acquisition, was tested in the model after it had been validated (see next section). As put forward in the introduction, three possible hypotheses can be formalized on the effect of simultaneous language learning (see table 1). Of these hypotheses, we expected the 'Facilitation due to comparison' to be correct, implying simultaneous language learning should be beneficial. To find out whether this is likely to be correct, we trained the SOMMUP model in three ways: sequentially, presenting the languages after each other; mixed, presenting all languages intermixed; and simultaneous, grouping the presented words by concept. It was up to the model to show how it performed after being trained with the different language learning schemes, which should subsequently be generalizable to human language learning.

7.1.1 Effect of learning schemes in the model. It turned out the choice of learning scheme had a significant effect on the rate of acquisition in the model, for the better and the worse. Actually, whether it was for the better or the worse was not easy to determine. This was caused by an in imbalance in the learning schemes, which made the sequential learning scheme incomparable to the other schemes. When we corrected for a part of this imbalance by changing the final error scores, the sequential learning scheme performed best of the three schemes. When we did not correct however, the pattern reversed and simultaneous instead of sequential language learning showed the best performance, followed with only a small difference by the mixed learning scheme. Due to this opposite pattern, we are reluctant to draw hard conclusions on the basis of these results, but a possible solution, some interesting patterns, and possible improvements can be discussed.

The main unbalance between the different learning schemes was due to the forgetting of previously learned languages in the sequential approach. Partly this is because of the interference present in most neural networks, partly forgetting could also be expected in humans if languages are no longer used. To counter the interference, representations for words from already learned languages could be made less prone to change. More precise, this could mean that the more often the same BMU is selected for a word, the less likely the BMU is to move to a different word. This would have the interesting side-effect that the model would try, even harder than in the current setup, to represent new languages using the old representations, just as humans seem to do (MacWhinney, 2005). This solution would also solve the second imbalance, the lack of sufficient training on the native language in the sequential learning scheme. Due to the fact that words from the native language are encountered the most often, they would be the least likely to change or be forgotten in this alternative setup. This compares well to the pattern of language development as found in humans, where long or early learned languages are not easily forgotten. The change would thus probably increase model validity. Moreover, this change would apply equally to all three learning schemes, not introducing new imbalances such as the current post-hoc correction of the performance scores actually did.

Even though it is impossible to draw hard conclusions on the basis of the current results, it is possible to draw some tentative conclusions. The first tentative conclusion is that mixed and simultaneous learning can at least lead to reasonable translation results. The mixed and simultaneously trained models did learn the translations to some extent, without severe detrimental effects of the alternative ways of presentation. A second tentative conclusion is that there might be differences in the pattern in subsequent performance depending on the learning scheme used. Sequential learning seems best for translations from the native languages to other languages, while simultaneous learning seems better for the opposite translations. An interesting follow-up question regarding these differences would be to see whether differences in performance are based on significant differences in the underlying structure of the lexicon. I.e., will the different learning schemes lead to different organizations in the lexicon? This was not tested yet, but could well be in the future. The third tentative conclusion is that this kind of modeling of acquisition processes in psycholinguistically plausible models could be valuable, but has its problems, at least in the current setup. Some improvements in addition to the one already mentioned should be able to increase applicability of the model for the current purposes. Firstly, the model should be trained longer, both on the native language and on the specific learning schemes. Not only should training time be increased, also repetitions of training sessions should be done to test reproducibility. Secondly, more analyses needs to be done on the developmental aspects of the model to be sure the ways of learning of the model and humans are actually comparable. This can show whether learning results of the model are actually generalizable, which requires more than only sharing general underlying principles. For example, it would be relevant to test rate and order of language acquisition, for early as well as late learners, to see in greater detail whether the model acquires the same words first and makes the same mistakes in the process of learning. Now, we focused our analysis of model validity mainly on the structural properties and not so much on the developmental properties.

A third possible improvement is related to one of the reasons to expect simultaneous learning to be beneficial compared to sequential learning. Specifically, it is related to the fact that simultaneous learning is expected to lead to better discrimination of alternative words, an expectation based on results in associative learning (see section 2.2). This expectation can be explicitly tested in the model by a small extension. This extension would require that never the same BMUs are activated for different patterns presented in a short time of time. When the words for a concept in different languages are then presented to the network, a number of BMUs equal to the number of unique words is determined. Next, each word is linked to the BMU which is relatively the closest compared to the others, without linking more than one word per BMU. This should force the model to make distinctions fast. This extension can also be applied to the other learning schemes, as they can also have similar words after each other which would otherwise activate the same BMUs. A comparison of the model with and without this extension is interesting for future tests of learning schemes in the model.

In short, the current findings did not lead to the expected results, but there are sufficient ideas for improvement to expect better results in the future.

7.1.2 Effect of learning schemes in humans. Even though the current tests did not work out as expected, using a psycholinguistic model such as the SOMMUP model to predict psycholinguistically and educational scientifically learning effects remains an interesting possibility. For psycholinguistics, it is a test of the implications of a certain view on the lexicon, such as the integrated lexicon in this case. For educational sciences, the reverse is the case, as relating to psycholinguistic models can help to determine what it is that causes the effects found in human learning. The latter is not yet the case for the effect of simultaneous language learning, because no human studies have been done yet. As put forward in the introduction, this is understandable from the commonsense perspective that simultaneous language learning will not work. As put forward in chapter 2 however, there is no strong basis for this commonsense notion, warranting research into the topic. This research could be done in parallel to further development of the model, in which case the model acts as theoretical framework for the actual experiment and the experiment as a test of the model.

Such an experiment on the effect of simultaneous learning could be setup in the following way. A selection of languages should be made. Known and unknown languages should be included to see whether the presentation of translational equivalents from known languages can facilitate the learning of words from new languages, due to the active comparison process. Both related and unrelated languages should be included to be able to check for the effects of similarity between languages and between individual words, as proposed to be a factor by the 'Similarity-dependent facilitation hypothesis'. For a more controlled test, fictitious languages could also be used. Using these languages, the procedure would be to show a concept for a certain period, either using pictures or the word in the native language. The participant would then be asked to translate the word in a number of languages. Different numbers of simultaneously presented languages should be tested to check for possible attention and memory related restrictions, which were not explicitly taken into account in the model. After the user has given his answer, feedback should be given and the user should have a set time to review the translational equivalents before the next concept is presented. During learning, the number of errors for each language should be recorded as well as the total time it takes. If fictitious languages are used, the test of sequential learning could be done within-subject with a new set of languages, which removes the potential effect of differences in language learning skills between participants, because the comparison is done within-participants.

After the controlled experimental phase, computer-assisted learning methods could be used (see section 7.4) to test the effectivity of simultaneous language learning in practice. The advantage of such a test is that the participants are intrinsically motivated to learn the words well and data gathering can be automatized, allowing for a test of actual applicability. The uncontrolledness limits the possible scientific conclusions one can draw however. Both approaches should ideally be tested, to answer both the questions from the psycholinguistic and educational sciences point of view.

7.2 The validity of the model

As stated previously, the second goal was to make a plausible model of the human lexicon. The validity of the model is considered in the next sections, in direct comparison with the human lexicon itself, as well as relatively compared to the most direct competitor: the SOMBIP model. Finally, we analysed which specific properties of the SOMMUP model were responsible for the current results in order to get an impression of the robustness of the findings and of possible alternative setups.

7.2.1 Comparison to the human lexicon. To recite briefly, the human lexicon shows the following four properties (Dijkstra, 2005). Firstly, orthography is organized in an integrated manner, meaning that words from all languages are stored together, without any organization on the basis of language membership. Secondly, semantics are shared across languages: Conceptual information is not language-specific, but only one instance of each concept is stored. Thirdly, the influence of language information on lexical processing is only minor and mainly contextual. Lastly, special word types, such as homographs, do not have special representations, but use a representation quantitatively different from non-homographs. Now, the question is how the SOMMUP model compares on these properties.

The answer is that our model seems to incorporate all four aspects. On the one side, words from all languages are located next to each other on one common map and can interact: Neighborhoods for orthography are both intra- and interlingual, resulting in errors due to similar orthography within and across languages. The semantic representation on the other side is ordered by semantic similarity and is totally language aspecific. This similarity-based structure results in similar concepts to be considered candidates for selection, as well as errors based on semantic similarity, in the same way as experimentally found for humans (Damian, Vigliocco, & Levelt, 2001; Vigliocco, Lauer, Damian, & Levelt, 2002; Schnur, Brecher, Rossi, & Schwartz, 2004). Language membership plays a minor role in the relation between orthography and semantics, just as it does in humans (see section 2.1) and it was intended to do (see section 3.1). It allows to separate translations to such extent that they are not confused. Most variations in orthography or meaning have a larger impact than a change in language membership information however. Regarding the fourth aspect, the representation of special word types, it was found homographs are not considered special in the model. Rather, they are considered quantitatively different: Shared representations are used where possible, meaning that cognates use shared representations for both semantics and orthography, while false friends only share the orthographic representations. Combined, our model captures the most important qualitative properties of the human lexicon.

On top of incorporating the correct qualitative properties of the human lexicon, the model was also able to predict the direction of change in reaction times in a range of tasks. We constructed a reaction time measure based on frequency and similarity, formalized by taking the distance from data to model vectors and between model vectors respectively. Using this measure we were able to explain not only frequency and neighborhood effects, but also homograph, proficiency, task and language effects. Importantly, the model showed both intra- and interlingual effects, confirming the integrated nature of the underlying artificial lexicon. Language-dependent effects on reaction times did require a language representation in the current setup though, of which the validity is subject of debate (French & Jacquet, 2004). We do think alternative setups are possible (see section 7.2.3). Also, we want to stress that we expect language information to be a contextual cue, related to the context of frequent usage of words, and no module representing 'language membership' is thought to be required. It is regarded mere one of the many properties words can be more or less similar on.

The combination of the qualitative and quantitative results leads to the most important implication of the current findings. This is the fact that the reaction times as found experimentally can indeed be the result of an integrated lexicon, as also confirmed in other models such as the BIA+ model (Dijkstra & Heuven, 2002). What is special about this model is its developmental and unrestrained nature: The model developed through training and only few assumptions were needed to constrain the training process. The main assumption behind the entire model was that lexical learning and representation is based on similarity, being similarity in orthography, semantics, or the relation between the two, with language membership being a non-dominant aspect words can be similar on. Depending on their similarity, words and concepts could next facilitate or inhibit each other. For the reaction times, this assumption was combined with the assumption that representations improve with additional learning, represented by the frequency effect. Together, these two assumptions allowed for the explanation of a wide variety of effects, without the need of qualitative distinction between word types, or other significant assumptions.

This makes for a quite parsimonious explanation of the experimental data and could be regarded a minimalist framework for models of multilingualism in general.

7.2.2 Comparison to the SOMBIP model. The SOMBIP model (Li & Farkas, 2002) is the only model of multilingualism that is comparable to the model we proposed. The SOMMUP and SOMBIP model use a similar structure to explain the workings of the multilingual lexicon. The SOMMUP model does so for a lexicon containing eight similar European languages, while the SOMBIP incorporates two more distinctive languages, Chinese and English. But even though structurally the two models are similar, the results and implications differ.

First, a closer inspection of the structural features of both models is useful in order to see the similarities and differences. Just as the current model, the SOMBIP model is primarily based on two SOMs: one for phonology, instead of orthography, and one for semantics. The data for the phonological map is based on vectors of consonants and vowels. Each consonant and each vowel is represented by five feature units representing articulatory features of the sound. Semantics are represented by co-occurrence in native texts, resulting in language-specific semantic representations. The semantic representations in our model, in contrast, are language aspecific, which means our model needs a language signal to guide translations. Another difference is that the SOMBIP model used Hebbian learning to link the two SOMs, correlating the most activated model vectors in both maps, while we applied another SOM for this purpose.

The difference in structure and data also results in differences in the developed maps. The maps we found were totally integrated across languages; words from all languages were situated next to each other and concepts were shared. In contrast, the SOMBIP model developed language-*specific* representations on both the phonology and semantics map. I.e., even though the words from both languages were situated on the same map, as did the concepts on the other map, each language was confined to a specific part of the map, instead of totally integrated. This is not in line with the currently dominant view of shared semantics and an integrated orthography across languages (Dijkstra, 2005).

On closer inspection, these differences in the maps do not necessarily contradict the current findings for two reasons. For one, the way Li and Farkas modeled semantics is more related to the functional role of a word, instead of pure meaning. It could be that semantics and role are distinct modules in cognition and hence are organized in a different manner. If there are two stores, one for the functional role and one for the meaning, also different experimental results should be found. For example, it could mean that in a sentence context, where usage is more important, neighborhood effects between conceptually related words from different languages are less likely to occur than in concept naming. Orthographic neighborhood effects are found to be the same in sentence processing as in single word recognition (Perea & Pollatsek, 1998; Pollatsek, Perea, & Binder, 1999; Rüschemeyer, Nojack, & Limbach, 2008), though differences have also been reported (Mulatti, Reynolds, & Besner, 2006). No study is known to us which compares the effect of interlingual semantic neighbors between sentence and single word processing. A second cause of differences between the models might be the included languages. The language-dependent way phonology developed in the SOMBIP model could be due to the large differences between the languages used, while more related languages (such as the European languages in our model) would result in more integrated maps. The latter view is partly confirmed by the fact that less similar languages have fewer common hits on the orthographic map, which means they are represented further apart (see table 9): Compare, for example, the common hits of Spanish and Catalan, as well as German and Dutch, to the common hits of Spanish and English and German and Italian.

However, if there is only one conceptual system and the difference in organization of the lexicon is not due to differences in language similarity, the two models can be regarded as two different views on the language system. The difference between the two views has an important implication: Interlingual neighborhood effects are predicted to arise at different levels. In the SOMBIP model on the one hand, there are no within-map interlingual neighborhood effects, because both the phonology and semantics for the two languages are essentially separate. The only neighborhood effects that are possible are in the associations *between* the maps: I.e., the phonology can activate interlingual neighbors on the semantic map and the semantic map can activate phonological neighbors on the phonological map. As a result, the SOMBIP model predicts that multiple levels of the language system need to be involved for neighborhood effect to occur. Our model on the other hand predicts that neighborhood effects will also occur within-orthography or -semantics only, in addition to neighborhood effects in the mapping between the two.

Experimental evidence should be able to clarify which of the two accounts is correct. For such an experiment, it is important that semantics and orthography are manipulated separately. This could be done using a priming task, employing primes with neighbors that are either orthographically or semantically related to the target word. A comparable task as used by Costa et al (1999) could be used, with the difference that not primes themselves, but neighbors of the primes should be related or unrelated to the target. Ideally, four variables would be manipulated in the task:

- Word or picture targets, thought to be retrieved from the lexicon and semantic memory respectively.
- Word or picture primes, thought to influence either the lexicon or semantic memory.
- Primes with either orthographic or semantic neighbors related to the target.
- The neighbors being either intralingual or interlingual with respect to the prime.

In total this would result in sixteen trial types. If there is any neighborhood priming effect, the intralingual neighbors of the primes should influence the recognition of the target word and picture. The critical question is whether the *inter*lingual neighbors of the primes influence the recognition process within one level (an effect of interlingual orthographic neighbors on word recognition and interlingual semantic neighbors on a picture) or only in the mapping from one level to another (an effect of orthographic neighbors on the recognition of a picture and of semantic neighbors on word recognition). If the first is the case, this would be in line with the predictions of the SOMMUP model. If only neighborhood effects are found for the mappings from orthography to semantics and vice versa however, the account of the SOMBIP model seems more likely. Current evidence makes the SOMMUP account more probable, as there are clearly within-orthography and within-semantics neighborhood effects (e.g., Yates, Locker Jr, & Simpson, 2003; Ferraro & Hansen, 2002; Huntsman & Lima, 2002). Still, this could also be explained by assuming recurrency between orthography and semantics. In that case, a smaller effect of neighborhood in the mappings would be expected, compared to within-level neighborhood effect, as the latter should be faster and more primary. The just mentioned analysis allows to quantify the difference between the two effects and in this way see which of the two effects is more primary. A larger between-level neighborhood effect would favor the SOMBIP model, while no difference between the two at all or a larger within-neighborhood effect could be seen as evidence for the SOMMUP model.

An additional way to disentangle the views as embodied by the SOMMUP and SOMBIP model, is to look at the processing of nonwords, which are not expected to be interpreted up to the semantic level. If the latter is indeed the case, they should not show neighborhood effects according to the SOMBIP model, while the SOMMUP model would predict neighborhood effects to occur. As nonwords seem to induce neighborhood effects and can be part of neighborhoods (Pugh, Rexer, Peter, & Katz, 1994; Siakaluk, Sears, & Lupker, 2002), the account of the SOMMUP model is again more likely.

To conclude, the two models have a slightly different focus and also give different results and predictions (see table 10). Depending on the degree to which it is this difference in focus that causes the development of either language-specific or integrated maps, different experimental results can be expected. If the difference is due to the difference between function and meaning, different results can be expected for sentences and isolated words. If the difference depends on the similarity of languages, also in reaction times less interlingual effects should be found when using distinct languages. If the difference is general however, experiments testing interlingual orthographic and semantic priming effects, or nonword stimuli and neighbors, should be able to solve the contradiction. At the very least the SOMMUP model forms an alternative to the SOMBIP model, showing how the lexicon can also develop. More elaborate tests, both in the models and in behavioral experiments, are needed to shed more light on which of two accounts is more likely to be correct.

7.2.3 Comparison to alternatives. In addition to the question how the current model compares to the human lexicon and how it compares to the best comparable alternative model, one can ask whether all aspects of the current model are required to come to the reported results. The orthography and semantic SOM can be regarded the core of the model, as they are the primary cause of the neighborhood and frequency effects. In addition, they are easy to interpret, giving them a certain 'face validity'. Also the representations used in these SOMs, language-independent orthography and shared semantics, seem appropriate, as noted in the previous sections. But the interpretation of the in-between part, the translation and language layer, is less straightforward. This is also the part which is most different from the SOMBIP model.

The translation layer is intended to represent the similarities between conversions from semantics to orthography and vice versa, mediated by language. This means it features neighborhood effects based on all three aspects. These neighborhood effects are language-dependent, as language information guides the selection of the correct relation. Small language activation or asymmetric proficiencies can lead to the selection of the relation from the wrong language and thus the wrong word or concept. In addition, within a language the neighborhood of relations can also lead to the selection of an incorrect neighboring *relation*, even when orthography and language membership were correctly selected. For example, activating MAN and English, could instead activate the relation 'WOMAN English `woman`' because both semantically and orthographically they are close together. Also in learning, relations having similarities in both semantics and orthography are predicted to transfer more. Finally, the translation layer can be used to represent recurrency effects: If a relation is activated in the translation layer, its neighbors can be activated and their activation can be send back to the input layer. For instance, activating BEAR and English in a Dutch-English bilingual version of the model will possibly also activate the relation 'BIER Dutch `beer`', leading to feedback activation for the word BIER. This would predict that words that are semantically and orthographically similar are more likely to be selected as neighbors than words that are only orthographically similar.

The last aspect of the translation layer, the recurrency, is also its main problem: All properties it can help explain, except for the language effects, can also be explained using recurrency. In that case, the alternative for the first example would state that MAN activates `man`, which activates its neighbor `woman`, which subsequently activates the orthography for WOMAN. Language error effects could be explained without a need of relation neighborhoods by assuming that the wrong word was activated in the first place or the connections between orthography and semantics are noisy. There is no clear way to make a distinction between the two accounts, with and without relation neighborhoods, experimentally. And even though a distinction does not have to be made, the translation layer can just be viewed as representing the (recurrent) interrelation

Table 10
An overview of the differences and shared properties of the SOMMUP and SOMBIP model.

Aspect	SOMMUP	SOMBIP
Algorithm	SOM	SOM and Hebb
Languages	Similar (European Languages)	Different (Chinese and English)
Word types	Nouns	Nouns and verbs
Word representation	Orthography (edit distances)	Phonology (CVC structure)
Concept representation	Meaning (WordNet)	Function representation (word co-occurrence)
Language representation	Present (Languages nodes)	Absent (Language-specific representations)
Neighborhood effects	Both within and between maps	Only between maps
Language-specific lexicons	No, intermixed on one map	Yes, separated on one map
Language-specific semantics	No, shared representations	Yes, separated on one map
Word class representations	No, only nouns	Yes, separate representations for nouns and verbs
Incorporates language similarity	Yes	No
Incorporates reaction times	Yes (both facilitation and inhibition)	No
Explains frequency effect	Yes	No
Detailed developmental analysis	No	Yes
Different learning schemes	Yes	No

between orthography and semantics, parsimony should urge us to look for possible alternatives.

There is at least one alternative, inspired by the SOMBIP model. In this alternative, language membership information is thought to be represented in the orthography, but not prominent enough to drive lexical organization. For the model, this would mean that the small language effect now embedded in the translation SOM, would instead be added to the orthography SOM. Note that this effect is explicitly meant to be small, it should be just one of the features of the words and by far not strong enough to drive map organization. This would lead to words which are sensitive to language information, but are still organized on the basis of orthographic similarity in a predominantly language-independent lexicon. Preactivating the language membership feature would then correspond to giving contextual language information and would facilitate the selection of the correct word. Because the orthography holds the language information, language information is no longer required to guide the conversion from semantics to orthography and the relation between the two SOMs becomes linearly separable again. This means Hebbian learning could be used to learn the relations between the maps, just as in the SOMBIP model.

This account is more parsimonious, as one level of representation less is needed. Still it is expected to yield the same results: None of the effects now featured by the model critically depends on the neighborhoods in the translation layer. There are however two predictions which would differ between the two accounts. Firstly, intralingual neighbors are predicted to be activated more than interlingual neighbors, as the language membership feature brings intralingual neighbors closer together. This indeed could well be the case (Heuven et al., 1998; Lemhöfer et al., 2008). Secondly, this setup predicts the wrong word to be activated when a wrong relation is selected. Testing these predictions, which is at least possible for the first prediction, could show how likely this alternative account is. Note though that the first effect could again also be explained with recurrency, by assuming feedback from the translation layer to the orthography layer.

In summary, the orthography and semantic SOMs are the driving force behind the current results. Therefore, it is possible to explain the current results by an alternative model, differing in the setup of the mapping between the two SOMs. This version of the model would assume a influence of language membership information on the orthography, instead of on an in-between level. Still, most of the mechanisms would remain intact, because the model would still assume separate language-independent stores for semantics and orthography, a small influence of language membership, and development and organization on the basis of similarity. Except for the differences noted, the same results could hence be expected.

7.3 Extensions and improvements

There are a number of extensions and improvements possible to the model, dividable in three categories: data, network, and tasks. These changes should improve the results and generalizability of the results and are described next.

7.3.1 Data. The data as used for the current version of the model was restricted in a number of ways. First of all, only nouns were used, while also other word types could easily be included, without requiring any adjustments to the model. More importantly, the data was confined to orthography and semantics only, while also phonology plays an important role in vocabulary learning. For some experimental effects, phonology even plays a more dominant or sometimes opposite role compared to orthography (e.g., Dijkstra, Grainger, & Heuven, 1999; Jared & Kroll, 2001). This implies that an extended version of the model should ideally include phonology to allow for the explanation of more experimental findings.

Also, there is more to language learning than mere vocabulary. A possible extension would be to also include sentence processing in the model in order to allow the model to learn grammar. In addition, sentence processing could allow the model to recover the meaning of words from the contexts in the sentences (French, 1998), instead of through other means, such as the now used distances in WordNet. Sentence processing does require a modification of the model, as recurrent connections will be needed to represent the information over time. Still, including sentence and grammar processing allows for interesting new questions to test, such as transfer effects in grammar acquisition and the simultaneous acquisition of grammar.

Not only the data itself, also the representation of the data could be improved. The edit distances which were used for orthography and semantics offer high resolution distinctions, but at the cost of little psycholinguistic plausibility and high dimensionality. Instead, it is preferable to use a representation based more strongly on the processing known to be performed in human, for example representing orthography as open bigrams (Dehaene et al., 2005). This will however probably yield even less efficient representations and did not work in our preliminary tests. Reducing the dimensions using principal components analysis could solve the dimensionality problem, but makes the features no longer intuitively interpretable and even less convertible to any cognitive constructs. An optimum between the two goals, plausibility and efficiency, needs to be found, in which the concern for plausibility is the most important from a scientific point of view.

7.3.2 Network. The most obvious and important improvement of the model is an improvement of the results: The model needs more training to reach a higher percentage of successful translations. Especially the orthography SOM clearly required more time to settle, which had a negative influence on especially the neighborhood effects reported. Also the tests of the different learning schemes could be improved by more intensive training.

The next most important improvement of the model is a more efficient way of representation in the network, which could also decrease training time. In the current implementation, an even worse than localist representation, namely three times the number of patterns, was needed for topology preserving autoassociative mapping to be possible. This is far from ideal from a computational perspective, because complexity rises with layer size. This is especially a problem for the translation SOM. For the other SOMs, a more efficient representation can be used. For example, words and concepts can be represented by the distances to a number of reference model vectors, instead of using one BMU per pattern. This would decrease the required number of model vectors. We chose to not use this method to ease comparison with the SOMBIP model and make the model more intuitive.

For the translation layer, a solution is less easily found, but some potential ones exist, in addition to the combination of SOMs and RBF networks as described in section 3.1. For example, a backpropagation network could be used, trained in all directions. This approach loses the topographical properties of the hidden layer, which can be countered by either adding Hebbian learning and lateral connections to the network (O'Reilly & Munakata, 2000) or mapping the activations from the backpropagation units on a SOM to order them topographically. The latter is possible because the locations of units have no role in non-recurrent backpropagation, meaning that they can be ordered by another algorithm to invoke topology. Ideally though, a multidirectional algorithm would be developed for this purpose, either using SOMs or Hebbian learning, which should be able to automatically learn in multiple directions instead of having to train the network in each direction separately. In theory, this seems possible, in practice no such algorithm is known to us.

Another extension to the development in the hidden layer and possibly the other layers is to use a dynamic instead of a fixed number of nodes. This way, the model is less restricted in its development. A number of algorithms exist for this purpose (e.g.,Fritzke, 1995; Dittenbach, Merkl, & Rauber, 2000; Flentge, 2006). Especially the Growing Neural Gas algorithm (Fritzke, 1995) is interesting in this regard, as it makes no assumptions on the dimensionality to which the data needs to be converted. The dynamic Neural Gas algorithm tries to represent the multidimensional data within a multidimensional space with as view nodes as possible, resulting in an efficient representation. The drawback is that it is harder to understand and graphically depict the representations that develop due to the high dimensionality.

A more drastic change would be to not use separate layers for orthography, semantics, and language, but one multidimensional SOM. In this case, the SOM needs to be more than two-dimensional, as early tests done to otest this option showed that using only two dimensions made the problem really hard to learn for the algorithm. This should not come as a surprise, since the dominant eigenvectors normally determine the dimensionality of the map and it is clear there are more large eigenvectors that describe all aspects of the data than mere two. At least two dimensions for both orthography and semantics should be included, as in the current implementation. Using one large SOM has the advantage of a lower computational complexity and less required assumptions concerning the structure of the cognitive system. Still all neighborhood effects can be explained, because partial maps can be viewed by only taking a subset of model vector dimensions into account. However, SOMs are not often used in such a way and many potential problems exist, as we found out in our preliminary tests and the current implementation. For one, the data needs to be balanced to make sure that not certain aspects of the data dominate the developing topology. If there is such an imbalance, the entire SOM could become structured on the basis of the semantics only, with little structure in the orthography, or vice versa.

7.3.3 Tasks. With regard to the comparison to experimental results, much has still to be done. Foremost, the current analyses only looked into the direction of reaction time changes, but not into the actual magnitude of the change or into an otherwise more concise comparison to experimental results. Doing such an analysis would show in greater detail whether the model is indeed comparable to the human language system. Moreover, it could also help to determine the weight of the different factors determining reaction times. For example, what are the relative impacts of the frequency and neighborhood effects, or the specific reaction times for semantics, language membership, and orthography? Further analyzing the comparison to experimental results, forces the model to actually predict reaction times for specific items, instead of only the rough direction of effects. Hence it makes for a stronger test of the model.

In addition, other experimental tasks should be tested in the model. The current tests for instance only focused on word recognition, while the model can actually also be used for picture recognition, as well as production, by applying inputs to the conceptual side only, with or without language. In addition, the model can also be used to model results in translation tasks. The interesting question in this regard is whether, and to what extent, the model actually needs to be adapted to predict reaction times in these tasks. If the modifications are large, one could expect different systems for recognition and production in human cognition. If this is not the case however, a (partly) shared system seems more likely. Previous studies point to partly the same neural regions to be responsible for language production and language understanding (Gernsbacher & Kaschak, 2003) on the one hand, making it likely also the processes involved are partly the same. On the other hand, there are reports of a language-specific lexicon underlying language production (Costa & Caramazza, 1999; Costa et al., 1999), which is in stark contrast with the language-aspecific respresentations found in recognition.

But also for word recognition studies, there are many possible directions of extension. Both semantic and orthographic priming studies could for example be incorporated. This can be done by calculating the activation for the prime and adding the activation for the actual target. This should yield faster reaction times for targets after related versus unrelated primes because of the competition between activations in the SOMs. Using the task-based similarity, task-dependent priming effects should also be explainable, as mentioned in section 5.1.

A topic which received too little focus in the current analyses, is language development. As mentioned, analysis of the developmental aspect of the model is also important for the tests of different learning schemes. Development is tested in greater detail in the monolingual developmental version of the SOMBIP model, the DevLex model (Li et al., 2004). Actually, this is one of the unique strengths of the DevLex, as well as the SOMMUP model: Both are a developmental and structural model of the lexicon at the same time. For the DevLex model for instance, it has been shown how new words are represented over the course of learning. New tests of the SOMBIP model furthermore showed how newly learned languages are added to an already developed monolingual lexicon. It was found new languages developed their own confined regions in the course of learning. Also comparing the SOMMUP model with the SOMBIP model on these aspects, should at least show the generalisability of the results found in the latter model. Based on the difference in the structure which developed for orthography in our model, compared to phonology in the SOMBIP model, we actually expect different developmental results. More precisely, we expect new words and concepts to be added dispersed over the maps, instead of in a language-specific confined region. An additional advantage of the SOMMUP model over the SOMBIP model is that it combines the possibility of structural and developmental analysis with the ability to generate reaction time predictions. This should also allow the prediction of experimental results in language learners.

To summarize, the fact that the SOMMUP model is a learning model, incorporating multidirectional mappings between orthography and semantics, neighborhood and frequency effects, and appropriate reaction time predictions allows the test of a wide variety of experimental tasks. This ability can and should be used to put the model to a more scrutinized test and check whether the current framework can be generalized to explain, and hopefully predict, more experimental outcomes.

7.4 Possible practical application: Computer Assisted Language Learning

The goal of Computer Assisted Language Learning (CALL) is to facilitate language learning with intelligent software. Often, such software is based on models of both the domain and the learner (Beatty, 2003). These models are hardly ever based on knowledge from psycholinguistics though, let alone on models already developed on the basis of such scientific knowledge. Still, it is reasonable to expect something is to be gained from turning towards scientific insights to improve models of the language learner (Ellis, 1995), or conversely look how a model such as proposed here can be applied in practice.

For an effective learner model for a CALL implementation, at least four factors are important:

- The ability to determine the cause of errors a learner makes.
- An adjustable representation of the language proficiency of a learner.
- A way to determine what the next steps in learning should be.
- A computationally efficient implementation to allow real-time processing during learning.

Our model should adhere to all these factors. We think it is possible for our model to do so and as such convert it to a working user model. In the following, we will briefly sketch how. This is followed by a short description of the specific advantages a CALL implementation would have for simultaneous language learning. The proposals done here need to be tested to determine the actual applicability.

7.4.1 Interpretation of errors. With respect to the ability to determine the cause of errors, it should first be clear what sources of errors there possibly are. A number of different error classifications exist (see e.g., Burt, 1959; Duskova, 1969; Odlin, 1989). Of these, the distinction between intra- versus interlingual errors and semantic versus orthographic errors is most important, because these are the main factors in vocabulary learning and hence in our model. The maps for orthography and semantics allow for an intuitive representation of the kinds of errors, namely by determining the distance between the target word or concept and the actually selected word or concept. I.e., the larger the distance between the correct and actual answer in a certain layer, the larger the error. Using the pattern in the errors, one can then predict the most likely source, as shown in table 11.

To clarify this point, a short example is useful. A learner is asked to translate the word BIKE from English to Dutch (correct answer: FIETS). The answer of the user happens to be BIETS. To interpret this answer, it is encoded and presented to the orthography layer, without any language information. Next, the language layer tells us what language the word most likely is (Dutch in this case, as BIETS does not resemble any high frequent English words) and what the most likely concept is, which turns out to be the correct concept, `fiets`, followed by the concept `biet`. In addition, the correct answer, FIETS for the orthography map, and `fiets` for the concept map, are activated. The shortest distance in the three layers (BIETS versus FIETS, Dutch versus Dutch, and `fiets` versus `fiets`) then determines the most likely type of error, in this case an orthographic, within-language error. If the user had answered AUTO (meaning CAR in English), it is clear the user made a conceptual error, while FAHRAD (BIKE in German) would point to a clear language error.

This error detection scheme should at least be applicable to the rough distinctions as considered in the example. We expect the distances to also give a reasonable approximation of the most likely source of a error in the case of more complex or multiple sources of error. Using information on language proficiency, the pattern of errors from other trials, and earlier presentations of the same trial, this approximation can probably be further improved. How this will work and what improvements will be needed in practice remains to be investigated.

7.4.2 Representation of proficiency. To be able to predict the errors correctly, the model should also have a notion of the language proficiency of the learner. For example, it should not point to a confusion with German as the source of an error when the learner does not know any German.

Proficiency could be represented by training the network real-time during language learning. If the model learns in a comparable way to the way humans do, this is the preferred way of proficiency representation and updating. There are two reasons to expect this not to work though. First, even though the model is thought to learn in a comparable way at the level of languages over time, this not to say it learns at the same speed or in the same way at the level of individual words. For instance, the model probably needs far more repetitions than humans to learn the same translation. Secondly, the computational complexity of learning the model alongside the user is quite high, as considered in section 7.4.4.

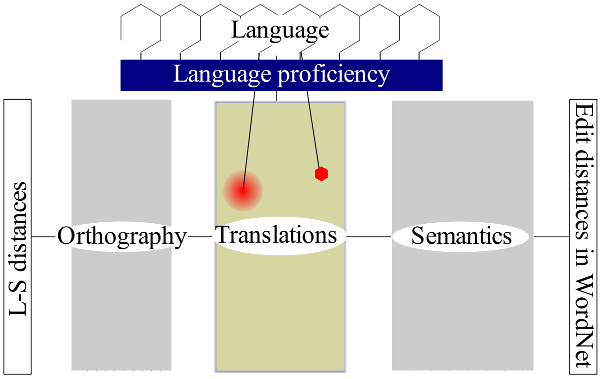


Figure 17. A possible way to convert the SOMMUP model to a user model by adding a proficiency module between the language and translation layer. This layer modulates the effect of each language node on the translation SOM for individual translations (shown by the red hexagon). These individual proficiencies combine to the overall proficiency for a language. The smallest possible proficiency change is at the level of these individual translations, but can be as large as the entire translation layer (represented by the yellow rectangle), representing overall language proficiency, or anywhere in-between (shown by the red Gaussian shape).

An alternative way is to test whether the fully learned model can be adapted to incorporate proficiency. Because language proficiency is directly related to language membership, the language layer is the best candidate to represent proficiency. In comparison, language proficiency cannot be represented in the orthography and conceptual layers, as these are both language-independent. Moreover, it are not primarily orthography and semantics that have to be learned in foreign language learning, but the mapping between the two, mediated by language. Using the language layer to represent proficiency, should make it possible to use a fully learned model of multiple languages and only manipulate the weights from the language layer to the translation layer to represent proficiency. No knowledge of German can, for example, be represented by setting the weights from the German node to the translation layer to zero, making sure no German nodes in the translation layer are activated. For more fine grained distinctions, this effect can be made node specific: if the German word for bike is not yet known, the weight of German to the FAHRAD translation node can be set to zero. This way, the proficiency can be modulated on both a general level, by modulating the overall influence of a language node on the translation layer, and on specific translations, through modulation of specific associations in the translation layer.

Table 11
An overview of the way distances in the maps could be used to determine the most likely cause of an error.

Orthography	Language	Semantics	Interpretation error	Example
-	-	-	Correct	BIKE - BIKE
Small	-	-	Orthography	BYKE - BIKE
Large	-	Small	Semantics (small)	CAR - BIKE
Large	-	Large	Semantics (large)	ELEPHANT - BIKE
Large	Large	-	Language	FIETS - BIKE
Large	Large	Large	Total	ROOS - BIKE

For large vocabularies, modulating one association at a time seems rather cumbersome to represent proficiency, while adapting all relations for one language seems rather crude. Here the topographical properties of the model come into play: If we assume that the proficiency of a user generalizes on the basis of word and concept similarities, proficiency can be modulated for larger, topographical related regions. In other words, if one knows a certain translation, one is expected to also know similar translations. The starting point of learning a new language for the user model would be the already known languages, with no influence of not yet known languages, as their weights are set to zero. Even with the foreign language nodes switched off, the user is expected to have some proficiency in foreign translations from the start, the ones close to the known language(s). The most apparent example is the translation of cognates, which should indeed be learned the fastest (Lotto & Groot, 1998). Next, when these most close neighbors are learned, some proficiency is expected to generalize to again neighboring translations, etcetera.

If a user is consistent in the correctness of his answers, the rate of proficiency change can be increased by including a wider range of translations surrounding the performed translations. Consistent correct answers should lead to an increase and consistent incorrect answers to a decrease in proficiency in this enlarged region. In contrast, if a user is inconstant in the correctness of his answers, which means that the model can be less certain of the proficiency changes, the region of proficiency change remains small or even shrinks. This way, proficiency can increase and decrease in a potentially fast, though controllable, way, which should allow for an efficient representation of proficiency.

7.4.3 Trial selection to facilitate learning. The topographical properties of the model are also useful for another purpose: the selection of new trials to remediate deficits in proficiency. As the first of three ways, trials can be ordered by the distance to the already known translations: First cognates and near cognates, slowly moving to more difficult words, less resembling known translations. Secondly, the trials can be selected on the basis of proficiency to influence the difficulty of learning. Recall, proficiency is proposed to be represented by the weights from the language layer to the translation layer. These weights could be used to select the translations with a particular proficiency. Trials can be chosen to fit the current proficiency, leading to a kind of 'Zone of proximal development' (Vygotsky, 1978) or be chosen with a certain difference to the current proficiency to challenge the learner. Thirdly, if a user makes many semantic, orthographic, or language errors, the topographical properties of the three SOMs can be used to select words with similar properties (to force the user to learn specific distinctions) or different properties (to make the learner first learn the more general differences). For example, if the learner keeps confusing Spanish and Italian words, orthographically close words from the two languages could be used as the next trials to make the user learn to discriminate the hard words, or first the more distinctive words could be presented to let the learner focus more on the general differences.

7.4.4 Feasibility and complexity. The complexity of the current model is its major drawback for actual application as a user model. Especially for the large vocabularies one would want for real world applications, the current implementation is too inefficient. There are two sources of this inefficiency, as already mentioned: the representation of the data and the worse than localist representation in the model. The representation of the data we already tried to make more efficient by selecting a subset of distances. Still the representation was rather high dimensional. In addition, the representation does not allow to precisely predict which words someone selected if it is not a word from the dataset; The representation does not include information on individual letters which would allow such generalization. More efficient representations of data, which do contain letter information for orthography, are hence needed. The second source of inefficiency, the representation in the SOMS, is due to the localist requirements of the autoassociative mapping. As discussed in section 3.1 and 7.3.2, this should be remediable by combining the SOM and RBF approach or one of the alternative possibilities. The current approach is at least too inefficient to allow real-time processing for large datasets and is thus not yet ready for implementation in actual intelligent tutoring systems. Still, the ideas behind a possible implementation, as just proposed, could already be tested.

7.4.5 Advantages for simultaneous learning. A CALL-based implementation as just described can specifically facilitate simultaneous language learning, especially due to the help it can offer to learn to discriminate

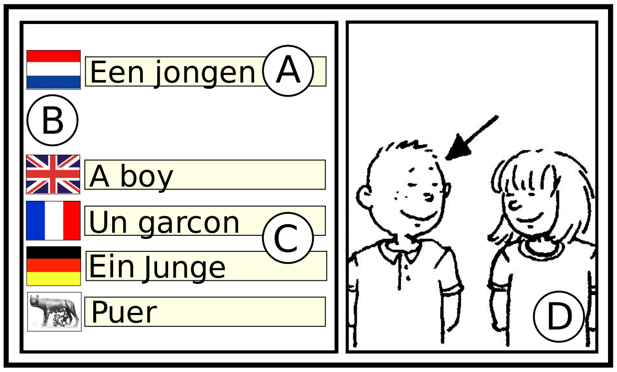


Figure 18. A schematic overview of a possible interface for a CALL-implementation aimed at simultaneous language learning. The word in the native language (**A**) is shown, together with the words in the foreign languages (**C**) and their language cues (**B**). The user would be required to fill in the words for the foreign languages. To facilitate activation of the underlying concept, a graphical depiction of the concept is presented (**D**).

languages. In addition to the ways already described based on the user model, also the interface itself can be an advantage. One way it can be, is by presenting different kinds of tasks. For example, explicitly asking the learner to select the appropriate languages for a set of words helps learners to detect the discriminative properties of languages. Another way is specifically focused on the importance of concepts in the kind of simultaneous language learning we consider here: presenting translational equivalents for one concept together. We expect this kind of learning to be facilitated by activating the concept through the use of media such as pictures, sounds, and movies. Moreover, to help the learner to discriminate the languages, it is also possible to add cues on language membership. Using such a multimodal approach is thought to facilitate learning in general (Hede, 2002; Levy & Stockwell, 2006). More specifically we think it can help learners to make the most of learning multiple languages simultaneously. An schematic overview of a possible user interface is shown in figure 18.

7.5 General Conclusion

The SOMMUP model, as proposed in this thesis, has the potential to become a well-rounded model. It does not only feature the structural properties of the human lexicon, but it can also predict patterns in reaction times and allows to study developmental properties of this lexicon. The latter did not work out yet for the main question of this thesis, on the effect of simultaneous language learning, due to conflicting results. However, the model did show that simultaneous language learning is possible. To determine whether it is actually beneficial, more modeling and experimental work needs to be done. The current work gave sufficient leads to follow up on, hopefully resulting in more research into this scientifically and practically relevant topic. It essentially boils down to the question whether we can do something as late learners, namely simultaneous language learning, which we once all could do as early learners.

References

Acha, J., & Perea, M. (2008, July). The effect of neighborhood frequency in reading: Evidence with transposed-letter neighbors. *Cognition*, *108*(1), 290–300.

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation of search? *Journal of experimental psychology: learning, memory and cognition*, *15*, 802-814.

Aupetit, M., Couturier, P., & Massotte, P. (2000). Function approximation with continuous self-organizing maps using neighboring influence interpolation. In *Proc. of neural computation.* Berlin, Germany.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). *The celex lexical database (release 2) [cd-rom].* Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Beatty, K. (2003). *Teaching and researching computer-assisted language learning*. Pearson Education.

Birdsong, D. (2005). Handbook of bilingualism: Psycholinguistic approaches. In J. F. Kroll & A. M. B. de Groot (Eds.), (p. 109-127). Oxford University Press US.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer-Verlag New York Inc.

Bruijn, E. de, Dijkstra, T., Chwilla, D., & Schriefers, H. (2001). Language context effects on interlingual homograph recognition: evidence from event-related potentials and response times in semantic priming. *Bilingualism: Language and Cognition*, *4*(02), 155-168.

Burt, C. (1959). *A psychological study of typography*. Cambridge University Press.

Campos, M. M., & Carpenter, G. A. (2000). Building adaptive basis functions with a continous self-organizing maps. *Neural processing letters*, *11*, 59-78.

Castagne, E. (2001). « intercompréhension et inférences: de l'expérience eurom4 au projet ice». In *Actes du colloque pour une modélisation de l'apprentissage simultanée de plusieurs langues voisines ou apparentées* (Vol. 13). Université de Nice-Sophia Antipolis.

Costa, A., & Caramazza, A. (1999). Is lexical selection in bilingual speech production language-specific? further evidence from spanish-english and english-spanish bilinguals. *Bilingualism: Language and Cognition*, *2*(3), 231-244.

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal of Memory and Language*, *41*(3), 365-397.

Crosson, B., Rao, S., Woodley, S., Rosen, A., Bobholz, J., Mayer, A., et al. (1999). Mapping of semantic, phonological, and orthographic verbal working memory in normal adults with functional magnetic resonance imaging. *Neuropsychology*, *13*(2), 17–187.

Cuvo, A. J., Klevans, L., Borakove, S., Borakove, L. S., Landuyt, J. van, & Lutzker, J. R. (1980). A comparison of three strategies for teaching object names. *Journal of applied behavior analysis*, *2*, 249-257.

Damerau, F. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171-176.

Damian, M., Vigliocco, G., & Levelt, W. (2001). Effects of semantic context in the naming of pictures and words. *Cognition*, *81*(3), 77-86.

Degache, C. (2003). Romance cross-comprehension and language teaching : a new trend towards linguistic integration in europe. the galanet project solution. In *The international conference. teaching and learning in higher education: new trends and innovation.* Universidade de Aveiro, Portugal.

De Groot, A. M. B., & Van Hell, J. (2005). Handbook of bilingualism. In J. F. Kroll & A. M. B. de Groot (Eds.), (p. 9-29). Oxford University Press US.

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: a proposal. *Trends in Cognitive Sciences*, *9*(7), 335-341.

Deneve, S., Latham, P., & Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, *4*, 826-831.

Dijkstra, A. (2005). Handbook of bilingualism. In J. F. Kroll & A. M. B. de Groot (Eds.), (p. 179-201). Oxford University Press US.

Dijkstra, A., & De Smedt, K. (Eds.). (1996). *Computational psycholinguistics*. London: Taylor & Francis.

Dijkstra, A., Grainger, J., & Heuven, W. van. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language*, *41*(4), 496-518.

Dijkstra, A., & Heuven, W. van. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(03), 175-197.

Dijkstra, A., Timmermans, M., & Schriefers, H. (2000). On being blinded by your other language: Effects of task demands on interlingual homograph recognition. *Journal of Memory and Language*, *42*(4), 445-464.

Dijkstra, A., & Van Heuven, W. (1998). Localist connectionist approaches to human cognition. In J. Grainger & A. M. Jacobs (Eds.), (p. 189-225). Lawrence Erlbaum Associates.

Dittenbach, M., Merkl, D., & Rauber, A. (2000). The growing hierarchical self-organizing map. In *Proceedings of the international joint conference on neural networks (ijcnn 2000)* (Vol. 6, p. 15-19).

Duskova, L. (1969). On sources of errors in foreign language learning. *International review of applied language learning*, *7*, 11-36.

Ellis, N. C. (1995). The psychology of foriegn language vocabulary acquisition: Implications for call. *Computer Assisted Language Learning*, *2&3*, 103-128.

Fellbaum, C. (1998). *Wordnet: an electronic lexical database*. MIT Press USA.

Fellbaum, C., & Vossen, P. (2007, January 25-26). Connecting the universal to the specific: Towards the global grid. In *Proceedings of the first international workshop on intercultural collaboration (iwic 2007).* Kyoto, Japan.

Ferraro, F. R., & Hansen, C. L. (2002). Orthographic neighborhood size, number of word meanings, and number of higher frequency neighbors. *Brain and Language*, *82*(2), 200 - 205.

Flentge, F. (2006). Locally weighted interpolating growing neural gas. *IEEE transactions on neural networks*, *17*(6), 1382-1393.

French, R. (1998). A simple recurrent network model of bilingual memory. In *Proceedings of the twentieth annual conference of the cognitive science society: August 1-4, 1998, university of wisconsin-madison.*

French, R. M., & Jacquet, M. (2004). Understanding bilingual memory: models and data. *Trends in Cognitive Sciences*, *8*(2), 87-93.

Fritzke, B. (1995). A growing neural gas network learns topologies. *Advances in neural information processing systems*, *7*, 625-632.

Gainger, J., & Dijkstra, A. (1992). On the representation and use of language information in bilinguals. *Advances in psychology*, *83*, 207-220.

Gathercole, S., & Thorn, A. (1998). Foreign language learning: Psycholinguistic studies on training and retention. In A. Healy & L. Bourne (Eds.), (p. 141-158). Mahwah, NJ: Erlbaum.

Gernsbacher, M. A., & Kaschak, M. P. (2003). Neuroimaging studies of language production and comprehension. *Annual Review of Psychology*, *54*(1), 91-114. (PMID: 12359916)

Göppert, J., & Rosenstiel, W. (1993). Topology-preserving interpolation in self-organizing maps. In *Proceedings of neuronimes 1993.*

Göppert, J., & Rosenstiel, W. (1995). Interpolation in som: Improved generalization by interactive methods. In *Proceedings of icann'95.* Paris, France.

Göppert, J., & Rosenstiel, W. (1997). The continous interpolating self-organizing map. *Neural processing letters*, *5*, 185-192.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of memory and language*, *29*, 228-244.

Greenwood, M. A. (2007). *Pure java wordnet similarity library.* Available from `http://nlp.shef.ac.uk/result/software.html`

Groot, A. de. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, *56*(3), 463-506.

Hede, A. (2002). An integrated model of multimedia effects on learning. *Journal of Educational Multimedia and Hypermedia*, *11*(2), 177-192.

Heuven, W. van, Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language*, *39*(3), 458-483.

Huntsman, L., & Lima, S. (2002). Orthographic neighbors and visual word recognition. *Journal of psycholinguistic research*, *31*(3), 289-306.

Jared, D., & Kroll, J. (2001). Do bilinguals activate phonological representations in one or both of their languages when naming words? *Journal of Memory and Language*, *44*(1), 2-31.

Kohonen, T. (2001). *Self-organizing maps* (T. Kohonen, Ed.). Springer.

Lee, S.-S. (1982). Acquisition of inductive biconditional reasoning skills: Training of simultaneous and sequential processing. *Comtemporary educational psychology*, *7*, 371-383.

Lemhöfer, K., & Dijkstra, T. (2004). Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, *32*(4), 533-550.

Lemhofer, K., Dijkstra, T., & Michel, M. (2004). Three languages, one echo: Cognate effects in trilingual word recognition. *Language and Cognitive Processes*, *19*(5), 585-612.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Learning, Memory*, *34*(1), 12-31.

Lenneberg, E. (1964). The structure of language: Reading in the philosophy of language. In F. J. Fodor & J. J. Katz (Eds.), (p. 579-603). Englewoord Cliffs, N.J.: Prentice-Hall.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*, 707.

Levy, M., & Stockwell, G. (2006). *Call dimensions: Options and issues in computer assisted language learning*. Lawrence Erlbaum Assoc Inc.

Li, P. (1999). Generalization, representation, and recovery in a self-organizing feature-map model of language acquisition. In N. L. E. Mahwah (Ed.), *Proceedings of the 21st annual conference of the cognitive science society* (p. 308-313).

Li, P. (2000). The acquisition of lexical and grammatical aspect in a self-organizing feature-map model. In *Proceedings of the twenty-second annual conference cognitive science.*

Li, P. (2001). A self-organizing neural network model of the acquisition of word meaning. In *In proc. of the 4th int. conf. on cogn. modeling* (p. 67-72).

Li, P., & Farkas, I. (2002). Bilingual sentence processing. In R. Heredia & J.Alterriba (Eds.), (chap. A self-organizing connectionist model of bilingual processing). North Holland: Elsevier Science Publisher.

Li, P., Farkas, I., Zhao, X., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*(8-9), 1345-1362.

Li, P., Zhao, X., & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science*, *31*(4), 581-612.

Lieberman, D. A. (2000). *Learning: Behavior and cognition*. Belmont, CA: Wadsworth Thomson Learning.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the fifteenth international conference on machine learning* (p. 296-304).

Lotto, L., & Groot, A. M. de. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, *48*, 31-69.

MacWhinney, B. (2005). Handbook of bilingualism. In J. F. Kroll & A. M. B. de Groot (Eds.), (p. 49-67). Oxford University Press US.

Mathworks. (2008). *Matlab 2008b for linux.* Available from `http://www.mathworks.com`

McCann, W., Klein, H., & Stegmann, T. (2003). *Eurocomrom–the seven sieves: how to read all the romance languages right away, ed. eurocom vol 5.* Shaker Verlag, Aachen.

Mondahl, M. (2002, November). Across the germanic language borders – text selection and the learner in the iglo-project. In *Eurocom – mehrsprachiges europa durch interkomprehension in sprachfamilien, eurocom – une europe plurilingue par l'intercompréhension dans les familles de langues* (p. 246-254). Fernuniversität, Hagen (D).

Mulatti, C., Reynolds, M., & Besner, D. (2006). Neighborhood effects in reading aloud: New findings and new challenges for computational models. *Journal of experimental psychology. Human perception and performance*, *32*(4), 799-810.

Münte, T. F., Heinze, H.-J., & Mangun, G. R. (1993). Dissociation of brain activity related to syntactic and semantic aspects of language. *Journal of Cognitive Neuroscience*, *5*(3), 335-344.

Neely, J. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648-654.

Neely, J. H. (1991). Basic processes in reading: Visual word recognition. In D. Besner & G. W. Humphreys (Eds.), (p. 264-336). Lawrence Erlbaum Associates.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. MIT press.

Pavlenko, A., & Jarvis, S. (2002). Bidirectional transfer. *Journal of applied linguistics*, *23*(2), 190-214.

Perea, M., & Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of experimental psychology: Human perception and performance*, *24*, 767-779.

Pienemann, M., Biase, B. D., Kawaguchi, S., & Håkansson, G. (2005). Handbook of bilingualism. In J. F. Kroll & A. M. B. de Groot (Eds.), (p. 128-153). Oxford University Press US.

Pollatsek, A., Perea, M., & Binder, K. (1999). The effects of neighborhood size in reading and lexical decision. *Journal of experimental psychology. Human perception and performance*, *25*(4), 1142-1158.

Pugh, K., Rexer, K., Peter, M., & Katz, L. (1994). Neighborhood effects in visual word recognition: Effects of letter delay and nonword context difficulty. *Journal of experimental psychology. Learning, memory, and cognition*, *20*(3), 639-648.

Ruhlen, M. (1991). *A guide to the world's languages: Classification*. Stanford University Press.

Rüschemeyer, S.-A., Nojack, A., & Limbach, M. (2008). A mouse with a roof effects of phonological neighbors on processing of words in sentences in a non-native language. *Brain and Language*, *104*(2), 132 - 144.

Schepens, J. (2008). *Distributions of cognates in europe based on the levenshtein distance.* (Bachelor thesis)

Schnur, T., Brecher, A., Rossi, N., & Schwartz, M. (2004). Errors of lexical selection during high and low semantic competition. *Brain and Language*, *91*(1), 7-8.

Siakaluk, P., Sears, C., & Lupker, S. (2002). Orthographic neighborhood effects in lexical decision: The effects of nonword orthographic neighborhood size. *Journal of experimental psychology. Human perception and performance*, *28*(3), 661-681.

Skinner, B. (1953). *Science and human behaviour*. MacMillan.

Snow. (1993). Psycholinguistics. In J. Gleason & N. Ratner (Eds.), (p. 391-416). Fort Worth: Harcourt Brace Jovanovich.

Studnitz, R. von, & Green, D. (1997). Lexical decision and language switching. *International Journal of Bilingualism*, *1*(1), 3-24.

Studnitz, R. von, & Green, D. (2002). Interlingual homograph interference in german-english bilinguals: Its modulation and locus of control. *Bilingualism: Language and Cognition*, *5*(1), 1-23.

Tagamets, M.-A., Novick, J. M., Chalmers, M. L., & Friedman, R. B. (2000). A parametric approach to orthographic processing in the brain: An fmri study. *J. Cogn. Neurosci.*, *12*(2), 281-297.

Tennyson, C. L., Tennyson, R. D., & Rothen, W. (1980). Content structure and instructional control strategies as design variables in concept acquisition. *Journal of educational psychology*, *4*, 499-505.

Thomas, M., & Allport, A. (2000). Language switching costs inbilingual visual word recognition. *Journal of memory and language*, *43*, 44-66.

Vamvakos, T. (2006). *Panorame of the european words.* Available from `http://www.users.otenet.gr/~vamvakos/multilingual.htm`

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). *Som toolbox for matlab.*

Vigliocco, G., Lauer, M., Damian, M., & Levelt, W. (2002). Semantic and syntactic forces in noun phrase production. *Learning, Memory*, *28*(1), 46-58.

Vossen, P. (Ed.). (1998). *Eurowordnet*. Dordrecht, Holland: Holland: Kluwer.

Vygotsky, L. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Yates, M., Locker Jr, L., & Simpson, G. (2003). Semantic and phonological influences on the processing of words and pseudohomophones. *Memory & Cognition*, *31*(6), 856-866.

Appendix A
Vocabulary

Throughout this thesis a number of concepts from the psycholinguistic literature are used. For clarity, the most relevant concepts are defined here, in order to facilitate understanding of the text.

Word properties

Different properties and types of words are reported in the literature, such as:

Word frequency The number of times the words is used in written and/or spoken language. Frequency values for English and Dutch were gathered from the CELEX database.

Translational equivalent Words that share meaning, but not necessarily orthography or phonology.

Interlingual homograph [3] Words that share orthography between languages, irrespective of a common meaning. Examples of Dutch-English homographs are ROOM, BED, and BANK.

Cognate [3] Interlingual homographs that not only share orthography, but also meaning between languages. BED and STOP are examples of Dutch-English cognates.

False friend [3] Refers to homographs that only share orthography across languages, but have a different meaning. An example of a Dutch-English false friend is ROOM, meaning `cream` in Dutch and `room` in English.

Neighbor A neighbor of a word is a word that is similar in orthography or semantics, within or between languages. An orthographic neighbor has a similar spelling, such as SHOE and SHOW (intralingual) or BOS and BOW (Dutch-English interlingual). A semantic neighbor on the other hand has a comparable or related meaning. The concepts `father` and `mother` are semantic neighbors.

Neighborhood density The number of orthographic neighbors a word has is called the neighborhood density. In English for example, LAKE has a large number of neighbors (TAKE, FAKE, MAKE, STAKE, CAKE, etc.), while fewer neighbors can be found for QUIET: the neighborhood density for LAKE is said to be higher than the density for QUIET.

Neighborhood frequency The neighborhood frequency describes the word frequency of the orthographic neighbors. LAKE for example has a high neighborhood frequency, as on average the neighbors are high frequent words. QUIET on the other hand has a low neighborhood frequency.

WordNet word properties

For each English word in its database, WordNet contains the following properties:

Hyponem : A word more specific then the target word.

Hypernem : A word more general then the target word.

Holonym : A word that names the whole of which the target word is a part.

Meronym : A word that names a part of the concept denoted by the target word.

Experimental tasks

The following types of experimental tasks are often employed in psycholinguistic research into multilingualism:

Language decision task Experimental task in which participants have to judge the language of a presented word. Often it is a forced choice between two languages. For example, a question in a lexical task could be: “Is the letter string LAAT a word in English or Dutch?”.

Lexical decision task In a lexical decision task participants have to judge whether a letter string is a word or not in a particular language. For example, a question could be: “Is the letter string WATER a word in English?”.

Generalized lexical decision task In the generalized version of the lexical decision task, participants have to answer whether a letter string is a word in any of the languages studied, not specifically in one. A question for the participants could be: “Is the letter string WATER a word in either English or Dutch?”.

Progressive demasking task In a progressive demasking task, a participant has to a respond as soon a word is recognized. The recognition of the letter string is complicated by the presentation of a mask directly following the stimulus. In the beginning of a trial, mask presentation is longer than stimulus presentation, but with each successive stimulus presentation mask duration is shortened and stimulus duration lengthened, until the word is recognized.

Priming studies In priming studies, a non-target stimulus, related or unrelated to the target stimulus is presented, which is intended to influence the performance on the target. For example, presentation of the word WING

[3] Note that there are multiple conventions using the notions homographs, cognates and false friends. In part of the literature, homographs for example refer to false friends. We instead used the term homograph to refer to the combined group of cognates and false friends.

is thought to facilitate recognition of the word PLANE because of the semantic relation, but no facilitation effect is expected for BOTER.

Go/no go task a task in which the participant has to respond to words from one language only, not to words from another language.

Appendix B Conventions

Because the distinction between orthography and semantics is an important one, the following convention is used:

1. The orthography of a word is denoted in capitals, like this: HOME.

2. Concepts are represented by the typewriter notation of the word, such as `home`.

To summarize, HOME denotes the English word for the concept `home`.

In addition, language proficiency is denoted by referring to the native or first language (L1), second language (L2), third language (L3) etc., with the order of languages representing the order of proficiency.