LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

CONCEPTUAL REVIEW ARTICLE

Toward Computational Models of Multilingual Sentence Processing

Stefan L. Frank
Radboud University, The Netherlands

Abstract: Although computational models can simulate aspects of human sentence processing, research on this topic has remained almost exclusively limited to the single language case. The current review presents an overview of the state of the art in computational cognitive models of sentence processing, and discusses how recent sentence-processing models can be used to study bi- and multilingualism. Recent results from cognitive modeling and computational linguistics suggest that phenomena specific to bilingualism can emerge from systems that have no dedicated components for handling multiple languages. Hence, accounting for human bi-/multilingualism may not require models that are much more sophisticated than those for the monolingual case.

Keywords multilingualism; sentence processing; computational models; probabilistic grammars; neural networks

Introduction

Computational modeling has been fundamental to the cognitive sciences and continues to be one of the most valuable methods for studying human cog-

I am thankful to Sol Lago, Xavier Hinaut, Jan Vanhove, Raphael Berthele, and three anonymous reviewers for their comments on earlier versions of this paper. The work presented here was funded by the Netherlands Organisation for Scientific Research (grant number 024.001.006 awarded to the Language in Interaction Consortium).

Correspondence concerning this article should be addressed to Stefan L. Frank, Centre for Language Studies, Radboud University. P.O. Box 9103, 6500 HD Nijmegen, the Netherlands. E-mail: s.frank@let.ru.nl

DOI: 10.1111/lang.12406

nition, including language processing. Research in the psycholinguistics of bilingualism, too, has been advanced by computational models; some of the best known or most recent examples being the BIA+ model of bilingual word recognition (Dijkstra & Van Heuven, 2002), the Multilink model of word translation (Dijkstra, Wahl, Buytenhuijs, & Van Halem, 2019), the SOMBIP model of mapping between phonology and semantics in the bilingual mental lexicon (Li & Farkaš, 2002), and its successor, the DevLex-II model of crosslinguistic lexical priming (Zhao & Li, 2013).

In stark contrast to this wide array of models of bilingual lexical development, representation, or processing, there are only very few sentence-level models of bilingualism. Although formal or verbal sentence-level theories do exist (e.g., Amaral & Roeper, 2014; Hartsuiker & Bernolet, 2017; Sharwood Smith & Truscott, 2014) there is a profound lack of computationally specified and fully implementable (let alone implemented) models.

Because of the immense complexity of the human language system, details about its functioning can only be understood by capturing them in implemented computational models. Unlike "mere" verbal theories, specifying a model forces one to bring hidden assumptions out in the open, and to do away with any ambiguity or vagueness in the theory. Successful models can reveal unexpected emergent phenomena and thereby help make sense of empirical findings. Hence, developing such models is critical to advance the field.

This is especially true for the bilingual case because, as bilingualism researchers are well aware, "the bilingual is not two monolinguals in one person" (Grosjean, 1989, p. 3): The bilingual system is even more complex than the monolingual system because of the ways languages can interact, not only with each other but also with other aspects of the cognitive system. However, little is understood of the mechanisms underlying these interactions. A successful computational model of bilingual or multilingual sentence processing can help fill this gap in our understanding and thereby provide answers to many concrete questions at the frontier of current multilingualism research: What is the effect of bilingualism on native-language processing (Cop, Keuleers, Drieghe, & Duyck, 2015)? Do predictive processes differ between L1 and L2, and if so, how (Kaan, 2014)? How does increasing L2 proficiency or exposure affect interaction between languages (Morett & MacWhinney, 2013; Whitford & Titone, 2012)? How are individuals' cognitive characteristics borne out in their L2 processing (Hopp, 2015; Linck, Osthus, Koeth, & Bunting, 2014)? Why are code-switches not distributed randomly over utterances (Green & Wei, 2014)?

The near nonexistence of computational models of bi-/multilingual sentence processing is particularly remarkable considering the current success of

computational psycholinguistic models in the monolingual domain. In what follows, I will first review some of the current work in single-language sentence processing models. Next, I discuss how these models could be (and, occasionally, have been) extended to handle two languages simultaneously. The final section speculates about what it would take to create truly multilingual sentence-processing models and what these might teach us about human multilingualism.

Computational Models of Sentence Processing

Probabilistic Processing and Word Surprisal

In the field of psycholinguistics, it is no longer a controversial idea that statistical patterns from the linguistic environment come to play an important role in the cognitive system. A considerable part of language acquisition comes down to learning (co-)occurrence frequencies of linguistic units,[1] and much of language processing is the application of the learned statistics. This view dovetails well with recent thinking about prediction during language comprehension because knowledge of language statistics can give rise to probabilistic expectations of upcoming language input. Although there is considerable debate about the precise role of linguistic prediction (Huettig, 2015; Kuperberg & Jaeger, 2016) it is clear that, at least sometimes and to some extent, people anticipate future input.

If linguistic prediction is probabilistic (i.e., statistical), it can be formalized and quantified using concepts from information theory. The most successful of these information-theoretic measures is *surprisal*—the negative logarithm of word's occurrence probability given the (linguistic) context. The linguistic unit of analysis here does not need to be a word but can be of any size. However, surprisal has most often been investigated at the word level. A word's surprisal is lower if the word was more likely to occur in the current context. To give a simple example: The word "gentlemen" is the most likely continuation of "ladies and" but not of "women and"; consequently, the surprisal of "gentlemen" is lower in the context of "ladies and" than after "women and".

A word's surprisal has been argued to correspond to the amount of cognitive effort required to process the word (Hale, 2001; Levy, 2008), a claim that has seen ample validation: Word reading times increase linearly with word surprisal (Goodkind & Bicknell, 2018; Smith & Levy, 2013); surprisal predicts neural activity as measured by electroencephalography (EEG; Frank, Otten, Galli, & Vigliocco, 2015), magnetoencephalography (MEG; Wehbe, Vaswani, Knight, & Mitchell, 2014), and functional Magnetic Resonance Imaging (fMRI; Willems, Frank, Nijhof, Hagoort, & Van den Bosch, 2016); and, crucially, learning more accurate language statistics generally results in stronger fit to

human processing data (Aurnhammer & Frank, 2019; Goodkind & Bicknell, 2018). Word surprisal has also shown to affect sentence production: When different syntactic structures are possible, writers tend to choose the one with lower surprisal (Rajkumar, Van Schijndel, White, & Schuler, 2016); and in speech, words with higher surprisal are pronounced more slowly (Demberg, Sayeed, Gorinski, & Engonopoulos, 2012) and are more likely to be preceded by a disfluency (Dammalapati, Rajkumar, & Agarwal, 2019).[2]

Surprisal theory is specified at the so-called "computational level of analysis" (Marr, 1982), which is to say that it is merely a function that relates probability to cognitive effort, without specifying any *mechanism* that gives rise to the probabilities (or to cognitive effort, for that matter). However, to apply surprisal theory, one needs an implemented language model that estimates the word occurrence probabilities. Such a model will be specified at the "algorithmic level," where representations and mechanisms are assigned that are appropriate to the task at hand. Probabilistic language models can differ widely in their algorithmic-level descriptions. The two general model classes that have been most influential in psycholinguistics are probabilistic phrase-structure grammars and neural networks, which will be discussed in turn below.[3]

Probabilistic Grammars and Incremental Parsing

It is traditionally assumed that the first (or, at least, an important) step toward understanding a sentence is incremental parsing: the word-by-word construction of its syntactic tree structure (Frazier & Rayner, 1982, among many others). When the sentence is not yet complete (and sometimes even when it is), it is usually consistent with a range of possible tree structures. In the probabilistic view on parsing, we do not select just one of these but generate many (if not all) possible structures, together with their probabilities. When the next word of the sentence comes in, it needs to be incorporated into the set of structures being considered, leading to changes in (the probabilities of) the structures. Levy (2008) showed how this amount of change gives rise to the word's surprisal. Hence, surprisal provides a direct connection between probabilistic incremental parsing and probabilistic predictive processing.

The question remains where all these probabilities come from. Put simply, the probability of a tree structure follows from the probabilities of all the production rules (also known as rewrite rules) that were involved in the structure's derivation. The rules' probabilities (and, most often, the rules themselves) are extracted from a corpus of sentences annotated with their syntactic tree structures. This collection of rules and their probabilities forms a probabilistic phrase-structure grammar.

Hale (2001) and Levy (2008) substantiate their claim that surprisal is a measure for cognitive processing effort by demonstrating that, under a probabilistic grammar, surprisal predicts reading times that match certain psycholinguistic phenomena such as garden-path effects. A garden-path effect occurs when a critical word is read more slowly if it syntactically disambiguates (i.e., its context is locally ambiguous) then when it appears in an unambiguous context. For example, it is well known that the main verb ("lost") is read more slowly in sentence (2) than in sentence (1), because in (2) it disambiguates between the two possible structures of "The doctor sued for damages" (Rayner, Carlson, & Frazier, 1983).

1. The doctor who was sued for damages lost the lawsuit.
2. The doctor sued for damages lost the lawsuit.

Simulations using probabilistic grammars in English (Hale, 2001; Levy, 2013; Van Schijndel & Linzen, 2018) and Dutch (Brouwer, Fitz, & Hoeks, 2010) have shown that a word's surprisal is indeed higher (corresponding to longer reading times) when it disambiguates than when it does not. Moreover, the success of probabilistic grammars goes beyond accounting for the reading time effects of particular experimental manipulations. Boston, Hale, Patil, Kliegl, and Vasishth (2008) and Demberg and Keller (2008) were the first to correlate word surprisal values (estimated by probabilistic grammars) to word reading times across sentences in German and English, respectively, that were not constructed for any particular experiment. Higher surprisal correlated with longer reading times, after factoring out more superficial variables such as word length, frequency, and probability given the previous word. This finding has been replicated many times since, at least for English, using different grammars and sets of reading time data (e.g., Monsalve, Frank, & Vigliocco, 2012; Roark, Bachrach, Cardenas, & Pallier, 2009; Van Schijndel & Schuler, 2015). More recent work on grammar-based surprisal has mostly focused on relating it to neural activity (Brennan & Hale, 2019; Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016; Frank et al., 2015; Shain, Blank, Van Schijndel, Schuler, & Fedorenko, 2020), often resulting in novel insights into the different brain processes and areas involved in language comprehension.

Note that this very short exposition barely scratches the surface of a large and multifaceted research field. For one, there exist many different types of grammar as well as many ways to relate aspects of sentence structure to observed cognitive processing difficulty (for a recent and more comprehensive review, see Demberg & Keller, 2019). As I will discuss next, several of the

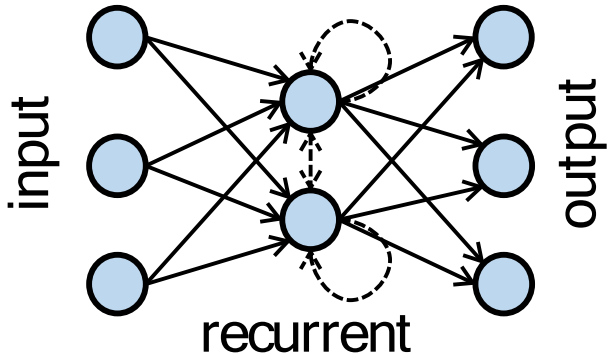

Figure 1 Schematic depiction of a small recurrent neural network, comprising an input layer, an output layer, and a layer in between. Solid arrows are connections from one layer to the next. Dashed arrows are recurrent (i.e., feedback) connections from a layer to itself. [Color figure can be viewed at wileyonlinelibrary.com]

psycholinguistic results from probabilistic grammars can also be obtained using models that are not based on syntactic parsing.

Recurrent Neural Networks

Inspired by biological neural networks, artificial neural networks simulate information processing as the flow of activation through a large number of, simple, numerical processing units (or "neurons") that are connected to one other into a network (as shown in Figure 1). A unit's activation is a simple function of the total activation it receives from the units it is connected with. Each connection has a weight that determines how strongly the "sending" unit affects the activation of the "receiving" unit. Weights may be negative, so that one unit deactivates another.

The connection weights are adapted to make the network perform a particular task. This is most often done by so-called supervised training, in which the network receives examples of inputs and the corresponding desired (target) outputs. For instance, a sentence comprehension network would receive sentences as input and their semantic representations as target output. Its connection weights are then adapted such that the network (approximately) generates the desired semantic representation for each example sentence. To the extent that it successfully learned a general sentence-to-meaning mapping, the network is then able to generate semantic representations of novel sentences.

The most popular neural network architecture for cognitive models of incremental language processing is the recurrent neural network (RNN; Elman, 1990). As can be seen in Figure 1, an RNN is fitted with feedback connections through which it receives its own previous internal activations, giving the model a decaying "short-term memory." This allows it to process input that comes in over time. If the model simulates word-by-word sentence comprehension, for example, each incoming word is interpreted in the context of the network's memory of the entire sentence so far.

Neural networks offer a perspective on the human language system that is very different from, and perhaps less linguistically informed than, grammar-based theories (see also Frank, Monaghan, & Tsoukala, 2019). Nevertheless, currently the most successful practical applications of language processing by machines (such as automatic translation) are based on neural network technology.

RNNs for Next-Word Prediction

The most common use of RNNs is as next-word prediction models, as originally proposed by Elman (1990). In this case, the network receives a collection of sentences or texts as training material, and at each point it is trained to predict the upcoming word. The network can only perform well on this task to the extent that it has discovered relevant aspects of the language's statistical structure. However, even with perfect knowledge of the language, predicting the correct next word is usually impossible. The network's output is thus not a single predicted word but a probability distribution over word types, that is, each word type has an estimated probability that it will be the next word. This provides an easy computation of the upcoming word's surprisal, which equals the negative logarithm of its probability estimate.

Early RNN models could only be trained on small, artificial languages, but recent technological developments allow training on millions of sentences from text corpora, using RNN variants that are more sensitive to long-range linguistic dependencies. The resulting surprisal values predict certain psycholinguistic phenomena that are traditionally thought to reveal structural syntactic processing, such as garden-path effects in Dutch and English (Frank & Hoeks, 2019; Futrell et al., 2019; Van Schijndel & Linzen, 2018). Also, like grammar-based surprisal, RNN surprisal accounts for word-reading times (Goodkind & Bicknell, 2018; Monsalve et al., 2012) and brain activity (Brennan & Hale, 2019; Frank et al., 2015; Wehbe et al., 2014) across naturalistic sentences in English.

RNNs for Sentence Comprehension

Compared to next-word prediction, the construction of a semantic representation is more central to the goal of sentence comprehension. Only a few RNN models have recently been proposed for mapping sentences to representations of the meaning expressed. Because of the difficulty of representing realistic semantics, all these RNN comprehension models are limited to hand-crafted, miniature languages, usually modeled on English. These models take as input a sentence, presented one word at a time, and generate as output some representation of the sentence's meaning. Where they differ is in how they represent meaning:

- *Propositional structures* (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Hinaut & Dominey, 2013) identify the agent, patient, and action of a given sentence, that is, they represent the semantic roles and concepts that fill those roles.
- *Situation vectors* (Frank & Vigliocco, 2011; Venhuizen, Crocker, & Brouwer, 2019) represent the state-of-affairs in the world as described by the sentence, without any internal role-concept structure.
- *Sentence gestalts* (Rabovsky, Hansen, & McClelland, 2018; based on a classical model by McClelland, St.John, & Taraban, 1989) are developed by the neural network itself during training. Unlike propositional structures and situation vectors, they are not designed in advance and, consequently, a sentence gestalt is not directly interpretable but can only be used as part of the network in which it arose.

Most, if not all, RNN sentence comprehension models relate some measure of the "amount of change" in their semantic representation to cognitive processing difficulty. The propositional structure models by Brouwer et al. (2017) and Hinaut and Dominey (2013) take this measure to correspond to the well-known P600 EEG component,[4] which is often viewed as indicative of a sentence reinterpretation process.[5] The situation vector models by Frank and Vigliocco (2011) and Venhuizen et al. (2019) show that the amount of change in the network's output can be expressed in terms of word surprisal. Frank and Vigliocco further demonstrate that this predicts simulated word-processing time, that is, their model provides a mechanistic account of why higher surprisal leads to longer reading time. In Rabovsky et al.'s (2018) sentence gestalt model, the amount of change in the gestalt representation explains a wide range of N400 EEG effects from the literature, which are usually interpreted as indexing a word's unexpectedness or the difficulty of semantically integrating it with the earlier context.

RNNs for Sentence Production

There exist relatively few cognitive models of sentence production. By far the most influential is the Dual-Path RNN model by Chang (2002). As its name suggests, it assumes there are two processing pathways: The syntactic path takes care of sequencing words in the correct order, and the semantic path encodes the propositional structure (i.e., the role-concept pairs such as AGENT is WOMAN) of the to-be-expressed semantics. The Dual-Path neural network also has a layer of units with recurrent connections (as in Figure 1). Both the syntactic and the semantic pathway pass through this single layer, so that syntax and semantics can interact. As input, the Dual-Path model receives the proposition it is to convey as well as information about the desired verb tense and aspect. It then produces the corresponding sentence one word at a time. Each word that is produced is fed back into the network as input, so that subsequent words depend on what has been produced so far.

To learn a mapping from semantic representations to word sequences, the model is provided with a training set of target semantics paired with corresponding sentences. The target semantics is given to the network, which then starts to produce words. Each word is compared to the target output sentence, and if the produced word is incorrect, the network's connection weights are updated such that its output is more likely to be correct next time. This makes network training very similar to the RNN word prediction models of Section *RNNs for Next-Word Prediction*.

As was the case for the comprehension models discussed above, the Dual-Path model has only been applied to miniature languages, although these were based on a wider range of languages (e.g., English, German, and Japanese) than for other computational cognitive models. Moreover, the model's validity has been extensively evaluated against human language data, for example, from structural priming experiments (Chang, Dell, & Bock, 2006) and aphasic patients (Dell & Chang, 2014). The model's success in accounting for human sentence production behavior once again highlights the importance of statistical learning and processing to the language system. Hence, it makes sense to also take the statistical approach when moving from the monolingual to the bilingual case.

From Mono- to Bilingual Models

What are the desiderata for a bilingual sentence processing model? The statistician George Box famously said that "all models are wrong but some are useful" (Box, 1979, p. 202), although of course a model can be too wrong to be useful.

If our bilingual model is to be of any use and suggest answers to questions of interest, such as those mentioned in the introduction, it needs to:

- simulate aspects of processing in at least two languages, possibly incorporating differences in proficiency and exposure between languages;
- display one or more phenomena unique to bilingual processing, such as code switching, language transfer, or crosslinguistic structural priming;
- account for the relevant empirical data (e.g., a code-switching model should display code-switching patterns similar to those observed in bilinguals);
- be cognitively and/or neurobiologically reasonable (e.g., the long-term memory capacity required for learning two languages should not exceed the sum of memory capacities for the individual languages).

A model that successfully mimics a relevant phenomenon and accounts for the associated empirical data (i.e., that meets the second and third criteria above) displays what Jacobs and Grainger (1994) call *descriptive adequacy*: It describes some aspects of reality. Any model can be made descriptively adequate to some extent by introducing additional free parameters or mechanisms that serve no other purpose than to make the model display the desired behavior. In that case, descriptive adequacy is merely due to ad hoc assumptions and the model is said to have no *explanatory adequacy* (in Jacobs & Grainger's terminology) which is to say that it fails to explain the phenomenon. It goes without saying that models that lack explanatory adequacy are not very useful or interesting from a scientific perspective.

One way to reduce the risk that a bilingual model lacks explanatory adequacy is to construct it by building upon a monolingual model, such that the monolingual model is subsumed under the bilingual one. Such a model will reduce to the monolingual case if the L2 is absent,[6] which means that everything we have learned from the original monolingual model remains valid. For example, removing one language from the Bilingual Interactive-Activation (BIA) model of word recognition (Van Heuven, Dijkstra, & Grainger, 1998) reverts it to the foundational, monolingual IA model (McClelland & Rumelhart, 1981). Hence, to the extent that the BIA model simulates specifically bilingual phenomena, it actually provides an explanation of these phenomena because nothing was added to the IA model apart from a second language.

Bilingual Statistical Processing

As we have seen, currently the most successful monolingual sentence processing models are those that exploit the language's statistical properties. This may not be true for second language processing in cases where L2 proficiency is

very low and the language is acquired by explicit instruction, so that comprehension or production proceeds via rote-learned translations into/from the L1 and morphosyntactic (de)composition is based on the conscious application of nonprobabilistic rules. In what follows, we focus on more fluent and proficient L2 processing, when knowledge and use of *both* languages are statistical in nature.

However, even when a bilingual's individual languages can be captured in statistical language models, this approach may not immediately seem suitable for modeling the bilingualism itself because keeping the two languages separate requires something more strict and categorical than a probabilistic system can offer. How is a statistical system able to tell languages apart? There are at least two answers to this question. First, co-occurrences within a language are much more frequent than co-occurrences between languages. A statistical learner can quite easily pick up on this and thereby approximately categorize the languages, as also demonstrated in the very simple RNN model by French (1998), discussed briefly in Section *Bilingual Recurrent Neural Networks* below.

Second, the cognitive system is sensitive not only to *linguistic* statistical patterns but also to their interaction with the statistics of the world in general. Languages tend to be separated in the real world (by speakers, locations, situations, etc.), enhancing their separability and identifiability. Along the same lines, Pajak, Fine, Kleinschmidt, and Jaeger (2016) propose a framework according to which learning a second (or later) language is a process of probabilistic inference based not only on (perceived) similarities between languages but, crucially, also on "socio-indexical structure." That is, the distribution of speech/language varieties in the learner's environment provides information about similarities and differences between the varieties. This theory describes how learners develop a hierarchy of language, dialect, and idiolect knowledge (all with their unique, but also shared, statistics) and may account for patterns of transfer between languages.

Bilingual Probabilistic Grammars

Research on technologies for automatic natural language processing has included a fair amount of work on inducing bilingual grammars. Many of these studies (e.g., Burkett & Klein, 2008; Saers, Addanki, & Wu, 2012) require a so-called "parallel corpus," which contains pairs of translation equivalent sentences in both languages. Although this is clearly not a realistic way to learn two languages from a cognitive perspective, the resulting bilingual grammars could still form useful models of a bilingual adult's language knowledge. Other proposals do not depend on parallel corpora but can learn from corpora that

differ not only in language but also in content (Cohen, Das, & Smith, 2011; Iwata, Mochihashi, & Sawada, 2010). These models can therefore more realistically capture a bilingual's language exposure and form more promising cognitive models, although in practice they are not developed or presented as such. It thus remains to be seen whether these models have much psycholinguistic import. They are currently only evaluated in terms of the adequacy of the induced grammars, and not as human language processing models. However, it should be possible to use the grammars in an incremental probabilistic parser in order to compute word surprisal values, which could then be compared to human data from (bilingual) sentence comprehension experiments.

Perhaps a more fundamental issue for these models is the additional syntactic machinery they require in order to handle more than one language. Unlike many of the bilingual neural network models that will be discussed later, the bilingual grammar models are not simply monolingual models exposed to two languages. For instance, when the Burkett and Klein (2008) model induces the grammars of English and Chinese, it also learns explicit links between those parts of the languages that are believed to be translation equivalents. Likewise, in the Cohen et al. (2011) model, language-specific syntactic categories are replaced by language-general categories before grammar learning can begin. Reliance on an architecture that is especially designed for allowing bilingualism substantially lowers these models' cognitive plausibility as well as its explanatory adequacy.

It should of course be kept in mind that these bilingual grammar models were designed for technological applications and therefore never intended to be cognitively realistic or evaluated against human processing data, so judging them by these standards may be a bit unfair. An alternative approach to obtaining a bilingual, incremental, probabilistic parser would be to start from an existing cognitive theory and embodying this in a computational implementation. For example, Multiple Grammar theory (Amaral & Roeper, 2014) claims that the language system has access to a large number of mutually incompatible "sub-grammars" to deal with (apparent) irregularities in the language. Bilinguals have sub-grammars for both languages, with tags that indicate the language to which a sub-grammar belongs. Hence, there is no qualitative difference between mono- and bilinguals. This theory is not defined in probabilistic terms but, as the authors suggest, the sub-grammars could have frequency-based activation levels (along the lines proposed by Truscott, 2006), which would be a first step toward a full, probabilistic implementation. If the probabilistic sub-grammars can then be used by an incremental parser to obtain word surprisal values, the resulting model can be evaluated against human processing data in the same manner as

is routinely done with standard, monolingual models. Currently, however, there are no implemented and cognitively validated bilingual probabilistic grammars.

Bilingual Recurrent Neural Networks

Bilingual Next-Word Prediction

Recurrent neural networks are routinely applied to next-word prediction in a single language. What happens when an RNN is trained on two languages simultaneously? Interestingly, this requires no architectural changes; the network does not even need to be told that its input comprises two different languages. Rather, it learns to distinguish them from the simple fact that words of the same language co-occur much more often than words from different languages. A notion of language identity (even if only approximate and implicit) is both learnable from the training data and very useful to perform next-word prediction, hence, the network will develop a separation between the languages.

An early demonstration of this principle was provided by French (1998) who trained an RNN model on two simple, artificial languages that differed only in their vocabularies. All sentences had a three-word SVO structure, with all three words coming from the same language. During training, the input language would occasionally switch from one sentence to the next, but never within a sentence. The (rather predictable) outcome was that the network's internally developed word representations showed a clear separation between the two languages.

More recently, Frank, Trompenaars, and Vasishth (2016) used an RNN that was trained for next-word prediction on large corpora of Dutch and English text. The network computed surprisal values over Dutch and English sentences from a reading experiment with Dutch-English bilinguals. The experiment manipulated grammaticality of sentences with double-embedded structures and showed a difference in reading-time effects between the two languages, which the RNN-based surprisal approximately matched. Other instances of large-scale, bilingual word-prediction RNNs were developed for certain applications, in particular automatic translation (e.g., Auli, Galley, Quirk, & Zweig, 2013, for English–French and English–German), rather than for cognitive modeling.

Bilingual RNN Sentence Comprehension

There has been even less work on bilingual neural network models of sentence comprehension than next-word prediction. Hinaut, Twiefel, Petit, Dominey, and Wermter (2015) show that the RNN model by Hinaut and Dominey (2013; see Section *RNNs for Sentence Comprehension*) can successfully be trained simultaneously on two miniature languages modeled on English and French,

without receiving any explicit information that the input comes from two languages. In fact, performance of the bilingual model in either language was similar to performance of each of two monolingual networks. This result is particularly impressive because the English, French, and bilingual RNN models did not differ in their input and recurrent connection weights; only the weights going to the output units were trained on sentences from one or both languages. However, the authors do not present any evaluation against human data or interpretation in terms of human bilingual processing.

Although research on bilingual comprehension by neural networks is clearly still in its infancy, the Hinaut et al. (2015) result suggest that there is no principled problem for an RNN model to learn bilingual comprehension. There also seems to be no a priori reason to believe it will be particularly difficult for models based on other semantic representations than the propositional structures used by Hinaut and colleagues. The question remains, however, whether such models account for any human data that is relevant to bilingualism research and what, if anything, can be learned about human bilingualism from such models.

Bilingual RNN Sentence Production

Currently the most successful bilingual sentence processing models are from the domain of sentence production. Chang's (2002) Dual-Path model (see Section *RNNs for Sentence Production*) can quite easily be turned into a bilingual production model by simply training it on two languages simultaneously. It requires no architectural adaptations, apart from the minimal addition of language-control units that steer the network toward producing a sentence in one of the two languages.

Janciauskas and Chang (2018) used the bilingual Dual-Path model to study the effects of L2 age of acquisition (AoA) and length of L2 exposure. They first trained the model on Korean-like sentences. After an amount of training that was varied to model differences in AoA, they then trained it to process sentences in an English-like language. The model mimicked human learners in that its L2 English production had more errors with later AoA. However, the empirical finding that L2 never reaches native levels (except for very early AoA) could only be simulated when a "critical period" was built into the model by switching off learning in the syntactic pathway after some time, an addition that was not present in the original, monolingual model. Hence, when it comes to this particular finding, the bilingual Dual-Path model displays descriptive but not explanatory adequacy.

Tsoukala, Frank, and Broersma (2017) used a similar approach to investigate why L1 Spanish speakers sometimes produce the wrong gender for L2 English pronouns, that is, they confuse "he" and "she" (Antón-Méndez, 2010). A Spanish–English version of the bilingual Dual-Path model showed the same behavior, but when the Spanish-like language was adapted to lose its pro-drop feature, the L2 English pronoun gender error disappeared. This confirmed that the error may be caused by a transfer effect from Spanish pro-drop. The Spanish-English bilingual Dual-Path model can also produce code-switched sentences, even though it was never exposed to code switches during training (Tsoukala, Frank, Van den Bosch, Valdés Kroff, & Broersma, 2019). The patterns of generated code switches show some resemblance to those of Spanish-English bilinguals. In particular, the model behaves like Spanish-English bilinguals in that it is much more likely to code-switch into English after the auxiliary verb in progressive than in perfect-tense structures, while code-switches on the auxiliary occur equally often in both structures. However, it remains unclear why, exactly, the model shows these typically bilingual and somewhat human-like behaviors. Even though the model did not require ad hoc assumptions (it displays some explanatory adequacy), it currently fails to provide much in terms of a true understanding of the cognitive processes that underlie the behavior.

Beyond Bilingualism: Is There Anything Special About Multilingual Models?

We have seen several examples of bilingual grammar induction or sentence processing models. How can these be extended to handle more than two languages? This question has not yet been tackled in cognitive models but is a research focus in the field of computational linguistics, which is to say that these multilingual models are not intended to be models of human multilingualism.

Cohen et al. (2011) demonstrate that, for a typologically diverse range of languages, single-language grammar induction from unannotated sentences is often more successful when, at the same time, grammars of four other languages are learned from syntactically annotated sentences. Even more impressively, the model by Iwata et al. (2010) can learn simultaneously, without supervision or parallel corpora, a shared grammar, and language-specific sub-grammars for ten Indo-European languages plus Finnish. Apparently, going from bi- to multilingualism is unproblematic for models of grammar induction but, as mentioned in Section *Bilingual Probabilistic Grammars*, they do require specific machinery to deal with more than one language, and their psycholinguistic relevance is questionable and has not been evaluated.

More recent work on computational linguistic models that are massively multilingual (i.e., can handle more languages than most individual humans) applies neural networks instead of probabilistic grammars. This generally does not require much (if any) architectural adaptation specific to bi- or multilingualism. Johnson, Schuster, Le, and Krikun (2017) took an RNN model for translation between two languages and trained it on up to twelve language pairs including several European languages, Korean, and Japanese. It was then also able to translate between pairs it was never trained on, to handle code-switched input, and to engage in code-switched production. The only required addition to the original model (apart from the additional languages) was a set of extra input units to signal which of its languages the network should translate into.

Moving toward massive multilingualism does not imply that such language control gets out of control. Östling and Tiedemann (2017) managed to train a text-prediction RNN on Bible translations in up to 990 languages simultaneously, although increasing the number of languages did lead to worse performance. Language identity was not represented by a discrete symbol but as a continuous-valued, high-dimensional vector. This allows for the representation of an unbounded number of languages, as well as of their interrelatedness by assigning similar vectors to similar languages. Consequently, the model could interpolate between languages, and, for example, generate sentences from a continuous range of languages from Middle English to Modern English. In effect, the "number of languages" known by the model has become a meaningless concept (see Berthele, this issue, for a discussion of how the same can be argued for human multilingualism).

What is the potential for cognitive models of multilingualism? The results on bilingual RNNs (see Section *Bilingual Recurrent Neural Networks*) suggest that multilingualism is relatively straightforward to model: Next-word prediction and sentence-production RNNs can, in principle, be trained on one, two, or many languages. Although performance may degrade as the number of languages increases, as it did in the Östling and Tiedemann (2017) model, neural networks for translation (Zoph, Yuret, May, & Knight, 2016) or image-caption retrieval (Kádár, Elliott, Côté, Chrupała, & Alishahi, 2018) have demonstrated that training the model on additional languages can actually improve performance on languages for which little training data is available.

We have already seen that surprising phenomena can emerge when monolingual models are turned bilingual or when bilingual models become multilingual: The bilingual Dual-Path model displays unexpected transfer and code-switching behavior (Tsoukala et al., 2017, 2019), and RNN translation systems can translate between new language pairs (Johnson et al., 2017).

As more languages are added, more opportunities for such emergence arise. These uniquely multilingual abilities thus emerge from simply presenting multiple languages to models that have no unique multilingual design features. Therefore, observing "special" behavior of multilingual models (or, by extension, people) does not imply the presence of any "special" underlying mechanism. Rather, the increased complexity of the bilingual or multilingual system can arise without increasing the complexity of the underlying cognitive architecture. The view from the state-of-the-art in computational language models thus appears to be that there is nothing special about multilingualism.

Final revised version accepted 30 January 2020

Notes

1 This does not preclude a role for abstract linguistic units. Rather, the statistical learning view holds that the abstractions themselves are discovered from the statistical regularities in the language data, and that learning extends to the (co-)occurrence frequencies of the learned abstract units.
2 All these studies used data from English, with the exception of Willems, Frank, Nijhof, Hagoort, and Van den Bosch (2016) who used Dutch, so it remains to be demonstrated that these findings generalize to a typologically diverse set of languages. However, the information-theoretical basis of surprisal is language independent so there is no reason to assume that its validity is restricted to Indo-European (or even just Germanic) languages.
3 There also exist hybrid models that explicitly include syntactic structure in neural networks (e.g., Dyer, Kuncoro, Ballesteros, & Smith, 2016; Sturt, Costa, Lombardo, & Frasconi, 2003) but these have not been very influential in psycholinguistics.
4 To be precise, Brouwer, Crocker, Venhuizen, and Hoeks (2017) take as a prediction for P600 size the amount of change in a network layer just before the semantic output.
5 But see Fitz and Chang (2019) for an alternative, learning-based account of the P600, supported by RNN simulations.
6 In much the same sense that a monolingual person who will (in the future) learn a L2 is not a priori different from someone who will not.

References

Amaral, L., & Roeper, T. (2014). Multiple grammars and second language representation. *Second Language Research*, *30*, 3–36. https://doi.org/10.1177/0267658313519017

Antón-Méndez, I. (2010). Gender bender: Gender errors in L2 pronoun production. *Journal of Psycholinguistic Research*, *39*, 119–139. https://doi.org/10.1007/s10936-009-9129-z

Auli, M., Galley, M., Quirk, C., & Zweig, G. (2013). Joint language and translation modeling with recurrent neural networks. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1044–1054). Stroudsburg, PA: Association for Computational Linguistics.

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198

Boston, M. F., Hale, J., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*, 1–12. https://doi.org/10.16910/jemr.2.1.1

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkonson (Eds.), *Robustness in statistics* (pp. 201–236). New York, NY: Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS ONE*, *14*, e0207741. https://doi.org/10.1371/journal.pone.0207741

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in semantic processing. *Cognitive Science*, *41*, 1318–1352. https://doi.org/10.1111/cogs.12461

Brouwer, H., Fitz, H., & Hoeks, J. (2010). Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory. In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics* (pp. 72–80). Stroudsburg, PA: Association for Computational Linguistics.

Burkett, D., & Klein, D. (2008). Two languages are better than one (for syntactic parsing). In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 877–886). Stroudsburg, PA: Association for Computational Linguistics.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*, 609–651. https://doi.org/10.1207/s15516709cog2605_3

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Cohen, S. B., Das, D., & Smith, N. A. (2011). Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 50–61). Stroudsburg, PA: Association for Computational Linguistics.

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychological Bulletin & Review*, *22*, 1216–1234. https://doi.org/10.3758/s13423-015-0819-2

Dammalapati, S., Rajkumar, R., & Agarwal, S. (2019). Expectation and locality effects in the prediction of disfluent fillers and repairs in English speech. In *Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: Student research workshop* (pp. 103–109). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-3015

Dell, G. S., & Chang, F. (2014). Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B*, *369*, 20120394. https://doi.org/10.1098/rstb.2012.0394

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

Demberg, V., & Keller, F. (2019). Cognitive models of syntax and sentence processing. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 293–312). Cambridge, MA: The MIT Press.

Demberg, V., Sayeed, A. B., Gorinski, P. J., & Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 356–367). Stroudsburg, PA: Association for Computational Linguistics.

Dijkstra, T., & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*, 175–197. Retrieved from https://doi.org/10.1017/S1366728902003012

Dijkstra, T., Wahl, A., Buytenhuijs, F., & Van Halem, N. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, *22*, 657–679. https://doi.org/10.1017/S1366728918000287

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 199–209). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1024

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211. https://doi.org/10.1207/s15516709cog1402_1

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Frank, S. L., & Hoeks, J. C. J. (2019). The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. In *Proceedings of the 41st annual conference of the cognitive science society* (pp. 337–343). Montreal, QB: Cognitive Science Society.

Frank, S. L., Monaghan, P., & Tsoukala, C. (2019). Neural network models of language acquisition and processing. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 277–291). Cambridge, MA: The MIT Press.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Frank, S. L., Trompenaars, T., & Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, *40*, 554–578. https://doi.org/10.1111/cogs.12247

Frank, S. L., & Vigliocco, G. (2011). Sentence comprehension as mental simulation: An information-theoretic perspective. *Information*, *2*, 672–696. https://doi.org/10.3390/info2040672

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye-movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

French, R. M. (1998). A simple recurrent network model of bilingual memory. In *Proceedings of the 20th annual conference of the cognitive science society* (pp. 368–373). Mahwah, NJ: Erlbaum.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1004

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics*. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0102

Green, D. W., & Wei, L. (2014). A control process model of code-switching. *Language, Cognition and Neuroscience*, *29*, 499–511. https://doi.org/10.1080/23273798.2014.882515

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*, 3–15. https://doi.org/10.1016/0093-934X(89)90048-5

Hale, J. T. (2001). A probabilistic early parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the association*

for computational linguistics (Vol. 2, pp. 159–166). Stroudsburg, PA: Association for Computational Linguistics.

Hartsuiker, R. J., & Bernolet, S. (2017). The development of shared syntax in second language learning. *Bilingualism: Language and Cognition*, *20*, 219–234. https://doi.org/10.1017/S1366728915000164

Hinaut, X., & Dominey, P. F. (2013). Real-time parallel processing of grammatical structure in the fronto-striatal system: A recurrent network simulation study using reservoir computing. *PLoS ONE*, *8*, e52946. https://doi.org/10.1371/journal.pone.0052946

Hinaut, X., Twiefel, J., Petit, M., Dominey, P. F., & Wermter, S. (2015). A recurrent neural network for multiple language acquisition: Starting with English and French. In T. R. Besold, A. d'Avila Garcez, G. F. Marcus, & R. Miikkulainen (eds.), *Pre-proceedings of the workshop on cognitive computation: Integrating neural and symbolic approaches (CoCO @ NIPS 2015)*.

Hopp, H. (2015). Individual differences in the second language processing of object—Subject ambiguities. *Applied Psycholinguistics*, *36*, 129–173. https://doi.org/10.1017/S0142716413000180

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Iwata, T., Mochihashi, D., & Sawada, H. (2010). Learning common grammar from multilingual corpus. In *Proceedings of the ACL 2010 conference short papers* (pp. 184–188). Stroudsburg, PA: Association for Computational Linguistics.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1311–1334. https://doi.org/10.1037/0096-1523.20.6.1311

Janciauskas, M., & Chang, F. (2018). Input and age-dependent variation in second language learning: A connectionist account. *Cognitive Science*, *42*, 519–554. https://doi.org/10.1111/cogs.12519

Johnson, M., Schuster, M., Le, Q. V., & Krikun, M. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351. https://doi.org/10.1162/tacl_a_00065

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*, 257–282. https://doi.org/10.1075/lab.4.2.05kaa

Kádár, Á., Elliott, D., Côté, M. A., Chrupała, G., & Alishahi, A. (2018). Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 402–412). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-1039

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*, 32–59. https://doi.org/10.1080/23273798.2015.1102299

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. Van Gompel (Ed.), *Sentence processing* (pp. 78–114). Hove, UK: Psychology Press.

Li, P., & Farkaš, I. (2002). A self-organizing connectionist model of bilingual processing. *Advances in Psychology*, *134*, 59–85. https://doi.org/10.1016/S0166-4115(02)80006-1

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin and Review*, *21*, 861–883. https://doi.org/10.3758/s13423-013-0565-2

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman and Company.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*, 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McClelland, J. L., St.John, M. F., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, *4*, 287–335. https://doi.org/10.1080/01690968908406371

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In W. Daelemans (Ed.), *Proceedings of the 13th conference of the European chapter of the association for computational linguistics* (pp. 398–408). Stroudsburg, PA: Association for Computational Linguistics.

Morett, L. M., & MacWhinney, B. (2013). Syntactic transfer in English-speaking Spanish learners. *Bilingualism: Language and Cognition*, *16*, 132–151. https://doi.org/10.1017/S1366728912000107

Östling, R., & Tiedemann, J. (2017). Continuous multilinguality with language vectors. In *15th Conference of the European chapter of the association for computational linguistics* (pp. 644–649). Stroudsburg, PA: Association for Computational Linguistics.

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning additional languages as hierarchical probabilistic inference: Insights from L1 processing. *Language Learning*, *66*, 900–944. https://doi.org/10.1111/lang.12168

Rabovsky, M., Hansen, S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behavior*, *2*, 693–705. https://doi.org/10.1038/s41562-018-0406-4

Rajkumar, R., Van Schijndel, M., White, M., & Schuler, W. (2016). Investigating locality effects and surprisal in written English syntactic choice phenomena. *Cognition*, *155*, 204–232. https://doi.org/10.1016/j.cognition.2016.06.008

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased

sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*, 358–374. https://doi.org/10.1016/S0022-5371(83)90236-0

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 324–333). Stroudsburg, PA: Association for Computational Linguistics.

Saers, M., Addanki, K., & Wu, D. (2012). From finite-state to inversion transductions: Toward unsupervised bilingual grammar induction. In M. Kay & C. Boitet (Eds.), *Proceedings of COLING 2012: Technical papers* (pp. 2325–2340). Mumbai, India: Indian Institute of Technology Bombay.

Shain, C., Blank, I. A., Van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Sharwood Smith, M., & Truscott, J. (2014). *The multilingual mind: A modular processing perspective*. Cambridge, UK: Cambridge University Press.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sturt, P., Costa, F., Lombardo, V., & Frasconi, P. (2003). Learning first-pass structural attachment preferences with dynamic grammars and recursive neural networks. *Cognition*, *88*, 133–169. https://doi.org/10.1016/S0010-0277(03)00026-X

Truscott, J. (2006). Optionality in second language acquisition: A generative, processing-oriented account. *International Review of Applied Linguistics in Language Teaching*, *44*, 311–330. https://doi.org/10.1515/IRAL.2006.014

Tsoukala, C., Frank, S. L., & Broersma, M. (2017). 'He's pregnant': Simulating the confusing case of gender pronoun errors in L2 English. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 3392–3397). Austin, TX: Cognitive Science Society.

Tsoukala, C., Frank, S. L., Van den Bosch, A., Valdés Kroff, J., & Broersma, M. (2019). Simulating Spanish-English code-switching: El modelo está generating code switches. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 20–29). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2903

Van Heuven, W. J. B., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language*, *39*, 458–483. https://doi.org/10.1006/jmla.1998.2584

Van Schijndel, M., & Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.),

Proceedings of the 40th annual conference of the cognitive science society (pp. 2603–2608). Austin, TX: Cognitive Science Society.

Van Schijndel, M., & Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In *Proceedings of the conference of the North American chapter of the association for computational linguistics* (pp. 1597–1605). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1183

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, *56*, 229–255. https://doi.org/10.1080/0163853X.2018.1448677

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 233–243). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1030

Whitford, V., & Titone, D. (2012). Second-language experience modulates first- and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic Bulletin and Review*, *19*, 73–80. https://doi.org/10.3758/s13423-011-0179-5

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*, 2506–2516. https://doi.org/10.1093/cercor/bhv075

Zhao, X., & Li, P. (2013). Simulating cross-language priming with a dynamic computational model of the lexicon. *Bilingualism: Language and Cognition*, *16*, 288–303. https://doi.org/10.1017/S1366728912000624

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1568–1575). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1163

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

Creating Multilingual Computer Models of Human Language

What This Research Was About and Why It Is Important

One way to study how people understand language is by creating computer models that simulate some part of the mental processes. This research method has been very valuable for investigating understanding in a single language, but is not often used to learn about people who know two or more languages (bilinguals and multilinguals). It is important to understand bilingualism and multilingualism better because, worldwide, most people speak more than one language. For this reason, we looked at current computer models of language and discuss how these can be (and, occasionally, have been) turned into bilingual and multilingual models.

What the Researchers Did

The most successful computer models of human language understanding assume that people try to match what they hear or read to the patterns of language that they have experienced before. We review which models of this type have been proposed, which of these have also been used for bilingualism and multilingualism, and what their strengths and weaknesses are.

What the Researchers Found

- Only very few models have been applied to two languages simultaneously.
- In some cases, a bilingual model can be created by simply providing two languages to a model that was designed for dealing with one language.
- Researchers who developed automatic language processing systems, without the goal to learn about human language understanding, have created systems that can handle many languages at the same time, sometimes more than a hundred. Again, this often does not require more than providing many languages to a system that was designed for dealing with just one.
- Surprising, and sometimes human-like, behavior can emerge from models that are exposed to multiple languages. For example, they can switch language midsentence, or aspects of one language can influence how another language is processed.

Things to Consider

- Creating multilingual computer models of language might not be very difficult, if a suitable single-language model is already available.

- Behavior that is typical for bilinguals (such as language switching) can emerge without the presence of a specifically bilingual system. All that is needed is a system for language processing that has been exposed to two languages. This would imply that there is nothing special about multilingualism, in machines or in people.

How to cite this summary: Frank, S. L. (2020). Creating multilingual computer models of human language. *OASIS Summary* of Frank in *Language Learning*. https://oasis-database.org