Learning constructions from bilingual exposure

Matusevych, Yevgen

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Matusevych, Y. (2016). *Learning constructions from bilingual exposure: Computational studies of argument structure acquisition*. LOT Netherlands Graduate School of Linguistics.

Learning constructions from bilingual exposure

Computational studies of argument structure acquisition

TiCC PhD Series no. 49

Published by

LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6111
e-mail: lot@uu.nl
http://www.lotschool.nl

ISBN: 978-94-6093-223-6
NUR: 616

Learning constructions from bilingual exposure

Computational studies of argument structure acquisition

Proefschrift

ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus, prof. dr. E. H. L. Aarts,
in het openbaar te verdedigen ten overstaan van een
door het college voor promoties aangewezen commissie
in de aula van de Universiteit
op maandag 19 december 2016
om 10.00 uur

door

Yevgen Eduardovych Matusevych

geboren op 7 november 1987
te Leningrad, Sovjet-Unie

Promotiecommissie

Promotor:	Prof. Dr. A. M. Backus
Copromotor:	Dr. A. Alishahi
Overige leden:	Dr. B. Ambridge
	Prof. Dr. A. P. J. van den Bosch
	Dr. M. E. P. Flecken
	Prof. Dr. F. Keller
	Dr. M. B. J. Mos

Das Pergament, ist das der heil'ge Bronnen,
Woraus ein Trunk den Durst auf ewig stillt?
Erquickung hast du nicht gewonnen,
Wenn sie dir nicht aus eigner Seele quillt.

Johann Wolfgang von Goethe

Contents

Acknowledgments

This thesis is the result of over four years of intense work, which would have never even started without my academic advisors – Afra Alishahi and Ad Backus, to whom I owe my deepest gratitude. It was they who initiated the project, it was they who gave me an opportunity to work on it, it was they who made sure everything was running smoothly, and it was they who helped me to finish the thesis successfully on time. During these years I've had the privilege of being able to ask them for help nearly at any moment, whether they were working in their offices, or staying at their homes, or traveling on vacation, or even having sleepless nights during a maternity leave. Afra and Ad became to me much more than supervisors, and I am grateful for that too.

I would like to thank everyone whose professional advice helped me to shape my thoughts at different points during these years: both colleagues (Grzegorz, Seza, Véronique, etc.) and many other professionals who reviewed my manuscripts, commented on my conference presentations, shared with me their ideas and opinions. It is a pleasure to thank my dissertation committee members – Ben Ambridge, Antal van den Bosch, Monique Flecken, Frank Keller, and Maria Mos – for their effort and time spent on reading my manuscript, giving feedback, and attending my defense. Special thanks to Peta Baxter, who helped me a lot with the French data annotation.

I am grateful to all my colleagues from the two departments, known as DCI and DCU, for creating a relaxed, yet professional working environment. I am indebted to Fons and Sjaak, who head(ed) these departments and who never refused to fund my conference trips. I would especially like to show my appreciation to all the friends who were around, ready to support me personally: Ákos, Derya, Doug, Ingrid, Jan, Moin, Nanne, Phoebe, Suzanne, Thiago, Yan, Yu, Zsuzsa, and many others.

Speaking about personal support, I am deeply thankful to my parents who actively supported my decision to change my life and come to Tilburg six years ago, just as they always respected my (not always smart) decisions.

Keeping focused on a single project for such a long time was difficult at times. A PhD student's life easily turns into a nightmare when research is their only activity, and I am delighted that mine has never turned out this way – thanks to Adriana, who has shared this long journey with me and has kept me sane over its entire length.

CHAPTER 1

Introduction

Over half of the world's population is able to hold a conversation in more than one language (European Commission, 2012; Ansaldo, Marcotte, Scherer, & Raboyeau, 2008; Tucker, 2002). This majority, however, is not homogeneous: individuals arrive at their knowledge of a second language (L2) in many different ways. In fact, there is so much variability among L2 learners that it is hardly possible to describe an "average" learner. Their age, mother tongue, exposure to additional languages, aptitude, foreign language in question, amount and temporal patterns of first and second language exposure define countless specific populations, such as simultaneous English–French bilinguals in Quebec or heritage speakers of Turkish in the Netherlands. Moreover, speakers within each population may vary in a lot of the input-related details. Despite this diversity, an important goal is to understand general mechanisms of L2 acquisition applicable to all learners.

In this thesis, I focus on a particular mechanism – *statistical learning*. This is a type of inductive, bottom-up, input-driven learning: humans are able to acquire a language by noticing regularities in the linguistic input. The statistical learning account has mostly been developed on the material of child language acquisition, while in second language acquisition (SLA) the relevant theory is not so well-established yet (but see Onnis, 2011). One possible reason is the mentioned variability in L2 learners: isolated studies with particular groups of learners do not provide enough material for the broad picture to emerge. Another reason is that learning in SLA context, as Rebuschat (2013) argues, is often associated with explicit instruction, while statistical learning usually applies to implicit pattern-finding.

The methodology employed in my studies – *cognitive computational modeling* – allows me to eliminate the unwanted sources of between-learner variation, and to

focus on the phenomenon of interest: statistical L2 learning from input data. Another advantage of this method is the explicit control it provides over the input data that the simulated learner is exposed to.

The accounts of statistical learning are situated within a broader paradigm of a usage-based approach to language, which I adopt here. In this theory, language use obtains the central role: it advances an individual's linguistic competence, and guides the emergence of linguistic representations in the mind. Speaking about representations, it is common in usage-based theories to study constructions – units that comprise a form and a meaning. Constructions are positioned on a continuum from fully specific (e.g., a lexeme) to fully abstract (e.g., a syntactic pattern). In the present thesis, I work with a particular type of abstract constructions present in nearly every utterance. These constructions deal with sentential representations of a verb and its arguments, and are referred to as *argument structure constructions*, or *verb argument constructions*. By simulating the process of learning of such constructions in L1 and L2, I investigate how the learning outcomes are affected by various factors often discussed in the literature: variables reflecting distributional and semantic properties of the input, the amount of input and the moment of L2 onset, as well as the quantity of cross-linguistic influence (CLI).

Before I proceed with the description of the key notions, a short terminological note is necessary. First, by "bilingual learners/speakers" I mean any individuals able to use more than one language, without assuming their native-like proficiency in both languages. Second, the terms "bilingual learning" and "L2 learning" are used interchangeably, referring to any setup which involves the acquisition of more than one language.

1.1 Usage-based linguistics, statistical learning, and SLA

Statistical, or frequency-based, accounts of language learning are usually positioned within the general framework of cognitive, or usage-based linguistics. Language acquisition in this framework is directly grounded in speakers' individual experiences with language, and this view has yielded comprehensive theories of first language acquisition (Bybee, 2003; Tomasello, 2003). However, there is no such encompassing account explaining how two or more languages are learned. One reason for this is the variability of bilingual learners: there are at least two distinct groups – early bilinguals and late second language learners. The intra-group variability can be very high, and there is no clear boundary between the two groups (Unsworth & Blom, 2010). This makes it difficult to provide a universal usage-based description that would fit all. At the same time, the usage-based account has been applied both to early bilingualism and second language acquisition (SLA): there has been a great number of theoretical developments in this framework (Paradis & Grüter, 2014; Tyler, 2012; N. C. Ellis & Cadierno, 2009; N. C. Ellis & Larsen-Freeman, 2009; Bybee, 2008; N. C. Ellis, 2006a), as well as empirical studies with various populations of bilinguals (Paradis, Nicoladis,

Crago, & Genesee, 2011; Blom, 2010) and L2 learners of different proficiency, spanning from absolute beginners (Denhovska, Serratrice, & Payne, 2016; Saturno, 2015; N. C. Ellis & Sagarra, 2010) to highly proficient learners (Siyanova-Chanturia, Conklin, & Van Heuven, 2011; Durrant & Schmitt, 2010; Forsberg & Fant, 2010).

One direction towards developing a universal usage-based model is a claim that there is no fundamental difference between first and second language learning. This idea is not new (e.g., Ervin-Tripp, 1974), and it has been promoted within the Unified Competition Model (MacWhinney, 2015, 2012, 2008). As Ortega (2015) puts it:

> ... the usage-based perspectives adhere to the working hypothesis of a fundamental continuity between early and late language learning: Language learning is qualitatively, fundamentally the same complex dynamical adaptive systems phenomenon regardless of starting age. The differential success observed for varying starting ages – and for different contexts or diverse types of learners – can be accounted for, at least in principle, by different initial conditions ... and by differential experience of language ... (p. 370).

This view finds support in recent psycholinguistic and neurolinguistic literature: multiple languages use the common neural resource (e.g., Abutalebi & Green, 2007), linguistic knowledge is shared between languages (e.g., Hartsuiker, Beerts, Loncke, Desmet, & Bernolet, 2016), and various types of CLI[1] are observed in language use (Rothman, 2011; Pavlenko & Jarvis, 2002). Besides, the single-system view is compatible with recent theories in applied linguistics and sociolinguistics, which suggest that multilingual learners should rather be seen as "users" (in line with the usage-based approach) possessing a set of linguistic tools that can be mixed in actual language use (Blommaert & Backus, 2013; Hall, Cheng, & Carlson, 2006; Cook, 2002). This contrasts with theories which describe each language (L1, L2, etc.) as a system of rules that needs to be acquired, and see each particular usage as either correct or wrong.

The usage-based framework brings together several basic ideas: that language has a fundamentally social function; that language is not a set of rules, but rather a network of units; that grammar emerges from use, etc. (Beckner et al., 2009). This last point is important here, because it sets out the background for the statistical account of language learning. This account became influential after it was found that both babies and adults rely on distributional information in the input to segment words in speech (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). More specifically, learners were able to observe the frequency of co-occurrence of different syllables – so-called transitional probabilities – and intuitively make use of this information to make decisions about the locations of word boundaries in an unknown language. Subsequently, similar use of language statistics has been shown in other linguistic tasks: sound discrimination (Maye, Werker, & Gerken, 2002), grammar learning (Gomez & Gerken, 1999), etc. At the moment, it is widely agreed that the effects of language statistics, also known as frequency effects, are "ubiquitous" in acquisition (Ambridge,

[1] Cross-linguistic influence, or cross-linguistic transfer, or interference, although the latter term is usually used in negative sense only, unlike the other two.

Kidd, Rowland, & Theakston, 2015): learners are sensitive to the distribution of various linguistic units across multiple domains, such as phonology, lexis, morphology, syntax (N. C. Ellis & O'Donnell, 2012; Rebuschat & Williams, 2012; Saffran, 2003). Language statistics enables learners to find patterns in the input, often unintentionally, and discover the input structure.

If there is no fundamental difference between L1 and L2 learning, frequency effects must also be manifested in SLA. Indeed, it has been argued that statistical learning plays an important role in L2 learning as well. In particular, Onnis (2011) proposed four principles that guide L2 learning: integrate probabilistic information sources, seek invariant structure, reuse learning mechanisms for different tasks, and learn to predict (p. 204). The studies in this thesis are largely compatible with these principles. To simulate statistical learning, I employ the method of probabilistic (Bayesian) cognitive modeling. In language acquisition studies, this method has mostly been employed for developing experience-based computational models (Poibeau, Villavicencio, Korhonen, & Alishahi, 2013). In this capacity, probabilistic modeling is very close to statistical learning: both of them describe the process of rational inference from data (Perfors & Navarro, 2011), limiting the domain of investigation to a particular bottom-up type of learning.

Statistical learning is often discussed in relation to another notion – implicit learning (Hamrick & Rebuschat, 2011; Perruchet & Pacton, 2006). The latter is usually defined in terms of being unconscious, incidental, as opposed to conscious explicit learning of which the learner is aware. Both terms – statistical and implicit – refer to observation-based, experience-based type of learning. However, the exact link between statistical and implicit learning has been a point of debate. While some claim the two terms address the same phenomenon from slightly different perspectives (Perruchet & Pacton, 2006), others draw the distinction by emphasizing that statistical learning can result in acquiring both conscious (explicit) and unconscious (implicit) knowledge (Hamrick & Rebuschat, 2011). Studies in this thesis deal with bottom-up statistical learning, while characterizing the simulated learning process as implicit or explicit is difficult: measures of (un)awareness and (un)intentionality can hardly be applied to probabilistic computational models.

1.2 Cognitive modeling in SLA and bilingualism

Cognitive computational modeling is a well-established method in cognitive science in general, and in research on language acquisition in particular (Poibeau et al., 2013; MacWhinney, 2010; Chater & Manning, 2006). Computational models have helped scholars to demonstrate how input may enable children to acquire their mother tongue: words (Fazly, Alishahi, & Stevenson, 2010; Frank, Goodman, & Tenenbaum, 2009; Li, Zhao, & MacWhinney, 2007), morphology (Monner, Vatz, Morini, Hwang, & DeKeyser, 2013; Albright & Hayes, 2003), sentence-level constructions (Alishahi & Stevenson, 2008; Chang, 2008), etc. In contrast, there are far fewer models that simulate the learning of two languages, as I explain below (see also Li, 2013).

Although computational modeling cannot provide scholars with conclusive evi-

dence in favor of one or another theory, it is particularly helpful when it comes to studying the contribution of a specific learning mechanism or an input-related variable in isolation. This is often critical for studies in SLA and bilingualism: populations of learners are very heterogeneous, and it is difficult to study how individual factors affect the learning in the long term. Longitudinal studies are usually carried out in natural environments, for example in a classroom (see an overview by Ortega & Iberri-Shea, 2005). Additionally, there are learner corpus studies: corpora may cover long periods of time (e.g., Meunier & Littre, 2013). However, these types of research provide little to no control over many potentially interfering variables. In theory, this problem can be solved by ensuring that each variable is balanced in the population sample, but in reality this can hardly be done, given the number of variables.

Imagine a research group studying how the age of second language onset influences language proficiency in immigrant learners. DeKeyser (2013) recommends to narrow down the immigrant population by controlling such variables as subjects' first language, their age, length of residence in immigration, amount of communication in second language. Following these recommendations, our imaginary research group would have to find middle-aged participants speaking the same mother tongue who have been living in immigration for at least 10 years and communicating mostly in second language during this period. However, even if the researchers succeed to obtain a large enough sample of participants, it may potentially still be very diverse. A number of variables, in fact, remain uncontrolled, such as participants' knowledge of other languages, the amount of time they spend talking to native speakers, the amount of formal language training they have received, their language aptitude, etc.

In order to find out how each variable influences language learning, they should be studied in isolation. Manipulating one variable at a time is the key idea behind experimental research. Experimental SLA studies presuppose control over input and instruction. This can be achieved through exposing learners to a language they have no experience with (Dimroth, Rast, Starren, & Watorek, 2013), or to a (semi)artificial language (see an overview by Hulstijn, 1997). However, it is often problematic to keep variables under control in the long term, because participants normally perceive and produce immense amounts of natural language every day, and this may affect the target language learning through cross-linguistic influence.

Considering these methodological difficulties, the study of bilingualism and SLA can benefit from the use of computational models (Li, 2013). One of the main advantages of computational models is that each variable can be manipulated and studied in isolation. An ideal computational model, then, should give a researcher an opportunity to manipulate each learning factor relevant within the adopted theoretical framework. Within the usage-based approach, a number of factors are believed to potentially influence the learning:

1. Learner variables: mother tongue, history of language learning and use, biological age, moment of onset, learning aptitude, motivation, learning goals, etc.

2. Learning variables: learning setting, type of instruction (if any), type of linguistic input, amount of exposure, etc.

This list is far from complete and can be extended, and an ideal computational model would have to account for all of these variables. In this thesis, I focus on statistical learning and investigate several input-related variables, such as amount of exposure, moment of onset, etc.

There have been multiple overviews of existing computational models of SLA and bilingualism (Li, 2013; Murre, 2005; Thomas & van Heuven, 2005). In one of them, Li (2013) notes that there have been few models of bilingual acquisition, in contrast to models of processing, which account for linguistic representations in mature bilingual speakers. At the same time, it is often difficult to draw a clear line between the modeling of language processing and that of language acquisition. Some models are designed for studying language processing, but they are *learning* to process the language, and thus can potentially be used to study some phenomena of language acquisition. This also follows the reasoning in the usage-based account: on the one hand, language processing always leaves diachronic traces in acquisition, and on the other hand, acquired linguistic knowledge is the long-term outcome of language processing.

Early attempts to model SLA included simple connectionist models employed to simulate cross-linguistic influence in various domains (Broeder & Plunkett, 1994; Gasser, 1990). These were followed by the development of connectionist models of bilingual language comprehension (Dijkstra & Van Heuven, 1998; French, 1998; Thomas, 1998). The model of Li and Farkas (2002), who investigated bilingual processing using self-organizing networks, gave rise to the most fruitful line of studies in this area, which employed similar type of models for investigating bilingual lexical development (Shook & Marian, 2013; Zhao & Li, 2010; Li, 2009), cross-linguistic priming (Zhao & Li, 2013), lexical recovery in bilingual aphasia (Kiran, Grasemann, Sandberg, & Miikkulainen, 2013), bilingual object naming (Fang, Zinszer, Malt, & Li, 2016). The topological representations provided by these models have provided extremely useful insights about the bilingual lexicon. Various aspects of lexical learning were simulated in other recent models (Cuppini, Magosso, & Ursino, 2013; Monner et al., 2013; Yang, Shu, McCandliss, & Zevin, 2013). At the same time, bilingual learning beyond the word level has not been simulated computationally (although see Rappoport & Sheinman, 2005). This is one gap that the present thesis tries to fill in.

1.3 Cross-linguistic approaches to argument structure constructions

My studies focus on the acquisition of abstract constructions. I adopt the construction grammar view, which suggests that constructions are pairings of form and meaning (Goldberg, 1995; Langacker, 1987). More specifically, the focus of the studies is on argument structure constructions (Goldberg, 2006, 1995) – a particular type of constructions present in nearly every utterance we produce. To give an example, the sentence *My mom called me yesterday* contains the verb predicate *called* and two participants of the event, or arguments: *my mom* (AGENT) and *me* (PATIENT). At the same time, *yesterday* is a non-obligatory component of the verb's argument structure.

Although there is no comprehensive account of bilingual argument structure learning, there are a number of studies in this area. For example, it has been shown that constructions emerge in second language learners (Gries & Wulff, 2005), and that second language construction learning is guided by the same input properties as in the first language (N. C. Ellis, O'Donnell, & Römer, 2014a). A more difficult question is to what extent the constructions are shared in bilingual linguistic knowledge. Some studies argue in favor of shared constructional representations (Higby et al., 2016; Bernolet, Hartsuiker, & Pickering, 2013; Salamoura & Williams, 2007; Santesteban & Costa, 2006), in line with the usage-based theory of common learning mechanisms and common neural resources. At the same time, there is an alternative view: based on cross-linguistic typological data, Wasserscheidt (2014) argues that "constructions do not cross languages" (p. 305).

In the studies presented in this thesis, I assume that the learner builds up a unified set of constructions (constructicon) and uses the knowledge of both first and second language in making decisions during language production and comprehension. At the same time, there are no assumptions about whether actual constructions are language-specific or "blended" – that is, based on the evidence from both first and second languages.

1.4 Overview of the studies

This thesis consists of four main chapters: chapter 2 presents the data sets used in my studies – a multilingual corpus of verb usages annotated with argument structure information, while the three studies in chapters 3–5 employ computational modeling for investigating various phenomena in bilingual learning of argument structure constructions.

Although there is a clear overarching theme in the four main chapters, each of them constitutes a standalone study based on a journal article, which can in principle be read separately. For this reason, there is a certain amount of overlap between the studies, in particular when it comes to the methodological sections: three out of the four studies employ the same computational model, although it undergoes certain changes from one study to another.

Computational cognitive modeling is the methodological core of this thesis. More specifically, I use a learning model that acquires linguistic argument structure constructions from input. The model was first presented in the study of Alishahi and Stevenson (2008), and employs a mechanism of unsupervised Bayesian clustering. I choose this particular model thanks to its roots in cognitive science (J. R. Anderson, 1991), as well as its cognitive plausibility in many respects and its successful applications in the study of monolingual learning (Barak, Fazly, & Stevenson, 2013a, 2013b, 2012; Alishahi & Stevenson, 2010, 2008, etc.). In the studies presented here, I adapt the model to bilingual learning scenarios (chapters 3–5), implement new linguistic tasks for testing the model (chapters 3–5), and propose an enhanced learning mechanism better compatible with languages characterized by free word order.

Because the learning process is virtually input-driven, it is important to provide

the model with naturalistic linguistic input. Chapter 2 presents two subcorpora used as input data to the model: a smaller manually annotated corpus and a larger corpus automatically extracted from existing linguistic resources.

The study in chapter 3 does not focus on bilingual learning alone: instead, both monolingual and bilingual simulations of constructions learning are presented. The variables of interest are distributional and semantic input properties: they have been shown to affect construction learning, yet their exact contributions are not known. I investigate the model's performance on a verb selection task, and how this performance can be predicted depending on the values of the three target factors.

In chapter 4, I look into the effects of two basic input-related properties – amount of input and time of onset – on the learning of argument structure constructions. These two properties have been often discussed in the literature, yet their effects are difficult to disentangle in human learners: the amount of input is often confounded with the time of second language onset. In this study, five linguistic tasks are employed to approximate language development of the simulated learners, and I attempt to explain this development by the two properties mentioned above.

Finally, chapter 5 focuses on the phenomenon often believed to be central to bilingual learning – cross-linguistic influence. I propose that the amount of cross-linguistic influence in a computational model can be quantified, and the model's performance in linguistic tasks can be explained in terms of this amount. As a case study, I present computational simulations of two experiments on case marking comprehension, carried out with human participants.

The concluding chapter 6 includes a general summary, a discussion of broad theoretical and methodological implications of my work, and outlines a few directions for future research.

CHAPTER 2

A multilingual corpus of verb usages annotated with argument structure information[1]

2.1 Introduction

A number of cognitive computational models of language learning (Beekhuizen, 2015; Freudenthal, Pine, Jones, & Gobet, 2015; Alishahi & Stevenson, 2008; Chang, 2008, etc.), as well as practical NLP applications, such as semantic role labeling (SRL: Palmer, Gildea, & Xue, 2010) or relation extraction (Nastase, Nakov, Seaghdha, & Szpakowicz, 2013), deal with sentential representations based on the syntactic and semantic relations between the predicate and its arguments – predicate argument structure. These applications require annotated resources for model training and evaluation. Despite the variety of existing resources and the recent developments in this domain (Lopez de Lacalle, Laparra, Aldabe, & Rigau, 2016; Fellbaum & Baker, 2013; Baker, 2012; Palmer, 2009, etc.), there is still a lack of multilingual resources that combine both syntactic and semantic argument structure information. In this chapter, we present a multilingual corpus of verb usages annotated with such information.

2.1.1 Predicate argument structure

Argument structure is a term describing the realization of a predicate (usually a verb) and its arguments. To give an example, the verb *give* often describes a transfer of physical possession:

(1) He gave a toy to his friend yesterday.

[1] This chapter is based on the article of the same name submitted for publication in a journal.

The event described in (1) includes three arguments:

1. *he* – the giving person, or AGENT;
2. *toy* – the given object, or THEME;
3. *friend* – the receiving person, or BENEFICIARY.

Note that the word *yesterday* is an optional part of the sentence, it is not crucial for interpreting the predicate meaning. Such optional elements are not referred to as arguments, but as adjuncts.

In constructionist linguistic theories, multiple features have been proposed to affect the acquisition of verb argument structure. In our corpus, each verb usage is represented as an assembly of multiple features belonging to one of the three groups: lexical, semantic, and syntactic. Lexical features include the predicate head (*give*), its lexical arguments (*boy, toy, mom*), and prepositions (*to*). Syntactic features are represented by the word ordering pattern (e.g., ARG1 VERB ARG2 PREP ARG3). Finally, semantic features describe the semantics of the head predicate (or event), of the lexical arguments, as well as the participant semantic roles such as AGENT or THEME. Note, however, that all the semantic features obtain distributed representations, including semantic roles: e.g., AGENT in the sentence above is described as {ACTING, ANIMATE, CONCRETE, GIVING, VOLITIONAL}.

2.1.2 Related work

Resources which are most commonly used for obtaining the information about predicate argument structure in English are PropBank (Palmer, Gildea, & Kingsbury, 2005), FrameNet (Ruppenhofer, Ellsworth, Petruck, Johnson, & Scheffczyk, 2006), and VerbNet (Kipper Schuler, 2006), which we present in more detail in section 2.3.1. These resources have been used for developing automatic SRL methods (see overviews by Màrquez, Carreras, Litkowski, & Stevenson, 2008; Carreras & Màrquez, 2005; Litkowski, 2004). Such methods are often integrated with syntactic parsing within the same system, and the overall F1-score for such systems reaches 75–85%, depending on the domain (Surdeanu, Johansson, Meyers, Màrquez, & Nivre, 2008).

The situation is more complex with multilingual resources and methods. The mentioned resources have their respective equivalents in other languages (Bai & Xue, 2016; Subirats, 2013; Duran & Aluísio, 2012; Zaghouani, Diab, Mansouri, Pradhan, & Palmer, 2010; Burchardt et al., 2006; Palmer, Ryu, Choi, Yoon, & Jeon, 2006; You & Liu, 2005, etc.). But given the variety of resources, compatibility is one of the problems. The development of mappings and alignments is intended to improve the compatibility across languages and resource types (e.g., Lopez de Lacalle et al., 2016; Wu & Palmer, 2015; Palmer, 2009; Shi & Mihalcea, 2005), however more work needs to be done in this respect. Multilingual SRL and syntactic parsing systems (Hajič et al., 2009) address this problem, and their performance is comparable to the monolingual systems. However, multilingual systems and resources often contain ambiguous PropBank-style semantic labels, while fine-grained FrameNet-style labels may be more useful.

To summarize, multilingual resources that provide both syntactic and fine-grained semantic relations between predicates and their arguments have been rare, while the

existing automatic systems for creating such resources may generate substantial amount of noise.

In what follows, we describe two subcorpora of verb usages consisting of the features described above. The first corpus has been annotated manually – it is rather free of noise, but small due to the tedious nature of the annotation. The second corpus has been compiled by automatically extracting verb usages from existing English and German linguistic resources, which contain the necessary syntactic and semantic annotations. This subcorpus is larger, but noisier than the small corpus. We proceed with the explanation on how each corpus was obtained, followed by the description of the resulting data sets.

2.2 Multilingual manually annotated corpus

2.2.1 Resources

English, German, Russian, and French sentences representing first language (L1) input were obtained from **CHILDES** (MacWhinney, 2000) – a database with various transcripts of conversations between young children and adults. Additionally, we used **the Flensburg English Classroom Corpus** (Jäkel, 2010) to obtain English sentences representing second language (L2) input. This corpus contains classroom transcripts of lessons of English as a foreign language in German schools.

For English, the data of three children from the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001) were used: Anne, Aran, and Dominic. For German, we used the data of Caroline (von Stutterheim, 2004), Kerstin (M. Miller, 1979), and Leo (Behrens, 2006). For Russian, the data from the only two children available in CHILDES were used: Varja (Protassova, 2004) and Tanya (Bar-Shalom & Snyder, 1996). Finally, for French, the data from the Paris corpus were used: Léonard, Madeleine, and Julie (Morgenstern & Parisse, 2012; Morgenstern, Parisse, Sekali, Bourdoux, & Caet, 2004).

2.2.2 Choosing sentences for the annotation

From each corpus, all the child-directed speech was extracted (or learner-directed speech, in the case of L2). For each resulting data set, we compiled a word frequency list to estimate which verbs occurred more frequently. Based on the overall frequency of occurrence, we selected several frequent verbs for the annotation. Verbs whose forms are predominantly used in phrasal verbs (e.g., *go on, come on*) were not considered, as well as verbs such as *think* and *want*, whose arguments tend to be sentential clauses (2) or verb phrases (3):

(2) I think **it's getting dark**.

(3) I want **to come to the picnic**.

The verbs selected for the annotation are given in Table 2.1. Note that for Russian we select more verb types than for other languages, because for some of these verbs

we could only select fewer than 100 usages for annotation. Also, Russian verbs are characterized by either perfective or imperfective aspect: *delat* "to do" and *sdelat* "to have done" are considered different verb types in Russian.

For each selected verb, we sampled an equal number of its usages from each participating subcorpus (e.g., Anne, Dominic, and Aran, for L1 English). The sampled usages were merged into a single set and shuffled. For some Russian verbs, the total number of usages was rather low, in which case the sizes of the subcorpus samples were not balanced.

From the final sample for each verb, we only selected the usages in which the target verb was the head predicate. We eliminated the sentences in which the target verb:

- had no explicit arguments (4), or
- was accompanied with an auxiliary verb, as long as omitting the auxiliary from the utterance would make the sentence ungrammatical or change its meaning (5–6), or
- appeared in a relative clause, which affected the number or the ordering of the verb arguments (7).

(4) Oh look!

(5) What was it made of?

(6) Can I look now?

(7) Just like that one we've been playing with.

In German, the sentences with prefixed/particle verbs (e.g., *zumachen* "to close, shut down") were considered to represent the target verb (in this case, *machen* "to make"), as long as the meaning of the prefixed/particle verb was compositional and the prefix/particle was actually separated (8).

(8) Mach es wieder zu!
make.IMP it again shut
'Close it again!'

Following these guidelines, approximately 100 random verb usages for each language were selected for the annotation.

2.2.3 Annotation guidelines

Each verb usage is annotated with the following features: sentence pragmatic function, head predicate, lexical arguments, predicate semantics, argument semantic roles, syntactic pattern, and case-marking. Details on each feature are given next.

Sentence pragmatic function refers to the speaker's communicative goal: statement, question, or request.

Head predicate is the lemma of target verb (e.g., *give* for *gave* or *gives*).

Lexical arguments denote the lemmas of the verb arguments, except for the pronouns, which retained their actual form in the annotations (9). This is because there are substantial differences between the subject and the object forms of the same

Table 2.1: The verbs selected for annotation, with their average frequencies across the analyzed corpora.

Language	Verb	Frequency
L1 English	*put*	962
	look	677
	play	361
	make	282
	take	248
	give	172
L2 English	*come*	81
	go	79
	read	61
	show	58
	look	48
	put	29
German	*kommen* “to come”	569
	gucken “to look”	599
	gehen “to go”	325
	machen “to make”	282
	geben “to give”	158
Russian	*delat/sdelat* “to do”	420/52
	hotet/zahotet “to want”	237/4
	smotret/posmotret “to look”	105/120
	govorit/skazat “to say”	49/73
	sidet/sest “to sit”	63/20
	idti/poyti “to go”	44/51
	videt/uvidet “to see”	104/3
French	*faire* “to do”	1024
	vouloir “to want”	517
	regarder “to look”	464
	dire “to say”	359
	voir “to see”	290

pronoun (*I* vs. *me*, *on* "he" vs. *ego* "him", etc.). Since all these forms are exceptionally frequent in a language, it is sometimes argued that speakers store both forms of each pronoun in their lexicon, without one form being derived from the other one (Diessel, 2007; Hudson, 1995). Similarly, frequent French question words *qu'est-ce que* and *qu'est-ce qui* are considered to be a single lexeme (10).

(9) Show me your pencil case.
Arguments: *me*, *case*.

(10) Qu'est_que tu dis?
what 2SG say
'What are you saying?
Arguments: *qu'est_que*, *tu*.

Predicate semantics is annotated as a set of semantic primitives describing the verb: ACTION (*look at her hair*), STATE (*that looks interesting*), etc. An example is given in (11) below, while the full list of features is provided in Appendix A.

(11) You play with the trains.
Predicate semantics: {ACTION, PHYSICAL, MANIPULATE, PLAYFUL}

Argument lexical semantics is represented as a set, similar to the predicate semantics. It is obtained differently for nouns and for other parts of speech. For nouns, the semantic features are extracted automatically from WordNet, a large lexical database described in more detail in section 2.3.1 below. We use a method designed by Alishahi and Fazly (2010): for each argument (with its particular sense), its hypernyms are recursively extracted from WordNet up to the root. On each level of hyperonymy, multiple elements are provided – we select the first element and add this element to the lexical semantic representation (see Figure 2.1).[2] For parts of speech other than nouns (e.g., adjectives, pronouns) the semantic representation is compiled manually using the relevant features from those already present in the noun representations, and/or similar ones (12–13).

(12) I: {REFERENCE, SELF, PERSON, ORGANISM, LIVING THING, WHOLE, OBJECT, PHYSICAL ENTITY, ENTITY}

(13) big: {SIZE, ATTRIBUTE, PROPERTY}

Argument semantic roles are encoded as a set of semantic proto-roles each. Traditionally, roles are denoted as single labels, such as AGENT or THEME, however in our data sets the roles receive more event-specific semantic annotations (14). Importantly for the simulated task of construction learning, it has been shown that semantic roles, such as AGENT or THEME can be acquired from proto-roles (Alishahi & Stevenson, 2010).

[2] We do not claim that the resulting primitives are the actual semantic units humans employ in language processing. Instead, the reasoning is that humans are able to estimate the degree of semantic relatedness of different words, and the corpus must contain semantic representations that would enable a computational model to compute the degree of such semantic relatedness. In this sense the described approach based on WordNet relations is not only viable, but also rather common in computational linguistics (e.g., Pedersen, Patwardhan, & Michelizzi, 2004; Budanitsky & Hirst, 2001).

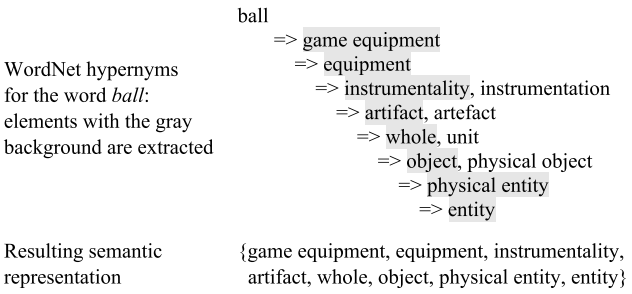


Figure 2.1: Extracting lexical semantics from WordNet.

(14) Cows make milk.
Arg1 proto-roles: {ANIMATE, CONCRETE, PRODUCING}
Arg2 proto-roles: {BECOME, INANIMATE, PRODUCED, SUBSTANCE}

Syntactic pattern denotes the order of arguments, predicate, and prepositions (15).

(15) You take it to the train.
Syntactic pattern: ARG1 VERB ARG2 *to* ARG3

Case-marking features are added for nouns and adjectives in Russian and German, with actual morphological cases being encoded (e.g., ACC). For many words, the case-marking is ambiguous. To give an example, the noun form *yabloko* "apple" can function either as a nominative (16) or as an accusative (17). Because the form is ambiguous, in both sentences, the case-marking for the target noun is annotated as {ACC, NOM}.

(16) Yabloko upalo.
apple.NOM fell.N
'An apple has fallen!'

(17) Hochu yabloko.
want.1SG apple.ACC
'I want an apple.'

Using these guidelines, all the instances are annotated. The resulting corpus of argument usages is described in the next section.

2.2.4 Resulting data set

The resulting data sets are comparable in size, although the L2 English data set is smaller because of the original corpus size. Each verb in the corpus is associated with its average frequency provided in Table 2.1, to preserve the original distribution of the annotated verbs in each language. The distribution of the most frequent values of each feature is shown in Figure 2.2. Note that these distributions reflect the frequencies of individual values in our annotations, not weighted by the frequencies of the individual verbs.

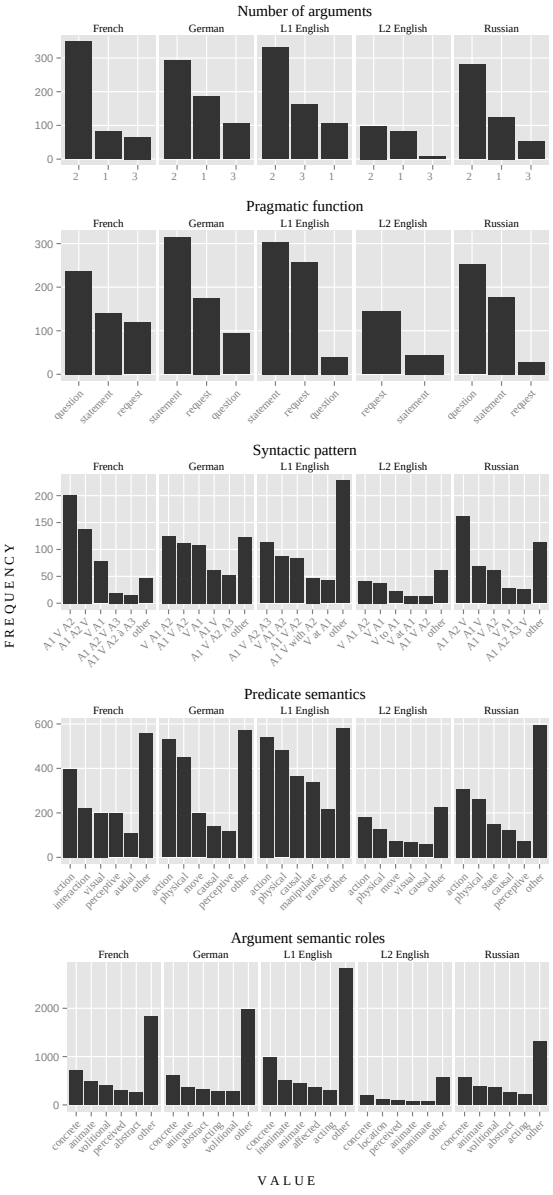


Figure 2.2: Distribution of feature values across languages.

We can see both similarities and differences across the language samples:

- Most instances in each language contain two arguments, while instances with only one argument are the least frequent in most languages, except for L1 English. This is not surprising, considering that the L1 English subset includes two verbs which often occur with three arguments: *take* and *give*.
- Statements are the most frequent pragmatic type in L1 English and German data, while Russian and French have more questions, and L2 English have more requests. The disproportionately low number of questions in the English data may be related to the use of auxiliary verbs *do, have, will* in most types of questions: recall that verb usages with auxiliary verbs were excluded. At the same time, Russian questions generally do not employ auxiliaries, and most French questions in our data occur with words *qu'est-ce que* and *qu'est-ce qui*.
- The most frequent syntactic patterns in all languages in general do not contain prepositions, with some exceptions in English and French. The verbs tend to occupy either the second position (as in English or German statements), or the first position (imperatives and questions). One exception is the pattern A1 A2 V, which is most frequent in Russian. This reflects the relatively free Russian word order.
- In terms of predicate semantics, the instances most frequently include verbs of PHYSICAL ACTION, although not in French: recall that four out of five annotated French verbs were not the verbs of physical action.
- In terms of argument semantic roles, CONCRETE arguments are the most frequent ones in each data set, while ANIMATE tend to occupy the second place (or third, in both English data sets). This reflects the fact that adults tend to talk to children about concrete, rather than abstract concepts: in particular, most events described in the instances involve at least one animate participant.

Figure 2.3 shows a full annotated example for each language. The next section describes the larger bilingual subcorpus compiled automatically from existing resources.

2.3 English–German automatically compiled corpus

2.3.1 Resources

The Penn Treebank (Marcus et al., 1994) is an English constituency-parsed corpus consisting of three subcorpora: the *Wall Street Journal* (WSJ), the Brown Corpus, and the Automatic Terminal Information Service data. We focus on the WSJ part consisting of one million words, because this is the only part that is semantically annotated in the Proposition Bank.

The Proposition Bank (PropBank; Palmer et al., 2005) provides an additional layer of semantic annotation to the WSJ part of the Penn Treebank. More specifically, the verbs are marked as predicates. The verb arguments are marked with their semantic role labels: ARG0 stands for agents, causers, or experiencers, ARG1 (patients), etc., and ARG2–ARG5 are rather diverse and verb-specific. Adjuncts, or verb modifiers, are also annotated: they can express location, time, etc. (18). Note that the roles are not

English
no we're not playing at that game today
predicate: play
index: 91
pragmatics: statement
semantics: action, physical, playful
argument1: we
argument2: game
proto-roles1: acting, animate, concrete, playing
proto-roles2: abstract, activity
case1: n/a
case2: n/a
pattern: arg1 verb at arg2

Russian
ona govorit nogi a ne nozhki "she says feet and not little feet"
predicate: govorit
index: 23
pragmatics: statement
semantics: action, communicative, audial, causal, interaction
argument1: ona
argument2: noga
proto-roles1: animate, concrete, producing, speaking, volitional
proto-roles2: abstract, message, perceived, produced
case1: nom
case2: acc, nom
pattern: arg1 verb arg2

French
fais un peu bravo! "do some cheering"
predicate: faire
index: 54
pragmatics: request
semantics: action, communicative, interaction, physical
argument1: bravo
proto-roles1: abstract, activity, produced
case1: n/a
pattern: verb arg1

German
dann kommt der Frühling "then comes the spring"
predicate: kommen
index: 42
pragmatics: statement
semantics: state, appear
argument1: fruehling
proto-roles1: abstract, coming, period
case1: acc, nom
pattern: verb arg1

Figure 2.3: Four annotated examples, one per language. Note that argument semantics is stored in a separate file, and is not shown in this figure.

necessarily assigned to a particular word: any constituent (phrase) may bear a semantic role.

(18) [$_{Arg0}$ The ministers] meet [$_{ArgM\text{-}LOC}$ in Australia] [$_{ArgM\text{-}TMP}$ next week].

The 2008 CoNLL Shared Task Data (CoNLL-08: Surdeanu, Johansson, Màrquez, Meyers, & Nivre, 2009) additionally provides syntactic dependency annotations for the PropBank data.

FrameNet (Ruppenhofer et al., 2006) is a database consisting of semantic frames – structured representations of events, relations, entities and their participants. For example, a frame MEET WITH describes an event which consists of PARTY1 meeting PARTY2 at a prearranged TIME and PLACE, possibly with a specific PURPOSE. The two parties are the core frame elements (FEs), while the other FEs are non-core. This frame can be evoked by a lexical unit *meet (with)*.

WordNet (G. A. Miller, 1995) is a rich lexical database, containing nouns, verbs, adjectives, and adverbs grouped into sets of synonyms (synsets). Importantly, the synsets are organized into a network based on semantic relations between them, such as hyponymy and hyperonymy.

VerbNet (Kipper Schuler, 2006) is a verb lexicon, in which verbs are organized into verb classes. Importantly for this study, verb classes are annotated with their semantic primitives: e.g., {TRANSFER, CAUSE, HAS-POSSESSION} for the verb *give*.

SemLink (Palmer, 2009) is a resource mapping existing lexical resources: PropBank, VerbNet, FrameNet, and OntoNotes. In this study, we only employ the existing mappings between PropBank and FrameNet. As we mentioned above, PropBank semantic labels ARG2–ARG5 are used inconsistently for different verbs, and we are

interested in replacing such labels with the respective more specific FrameNet labels (PURPOSE, PLACE, etc.).

The FrameNet–WordNet mapping (Bryl, Tonelli, Giuliano, & Serafini, 2012) provides a repository of WordNet synsets for each frame element in a given FrameNet frame. For example, for the frame element CAUSE in the frame CAUSATION the repository provides the following distribution of synsets: ABSTRACTION-NOUN$_6$: 76 occurrences, PHYSICAL ENTITY-NOUN$_1$: 43 occurrences, ENTITY-NOUN$_1$: 5 occurrences.

The TIGER corpus (Brants et al., 2004) is a German corpus of newspaper text from *Frankfurter Rundschau*, containing approximately 900,000 word tokens. Importantly for us, it is annotated with morphological information and with syntactic constituency structure.

The 2009 CoNLL Shared Task Data (Hajič et al., 2012) provide a version of TIGER annotated with syntactic dependency annotations.

The SALSA corpus (Burchardt et al., 2006) enriches TIGER with semantic role annotations. It mostly employs the existing frames and FEs from FrameNet, but some of the existing frames are adapted to German, and a number of predicate-specific proto-frames are used.

2.3.2 General approach

The general steps that we took to prepare both the English and the German subcorpus include the following:

1. Extract all the frame instances from an annotated corpus, containing lexical verb predicates and frame elements (role-bearing constituents) with their labels.
2. Filter out frames which are unsuitable for reasons explained below.
3. Align branches of the dependency parse trees and constituency parse trees (if needed).
4. Filter out adjuncts, or non-core frame elements.
5. In each role-bearing constituent of each frame, identify a preposition (if present), and the lexical head.
6. For all verbs and lexical arguments, extract distributed semantic features from WordNet and/or VerbNet.
7. Ensure that the role labels used in German and English data are consistent.
8. Expand each role label into a set of semantic primitives.
9. Combine the above-mentioned features in a single data set.

The next two sections explain how each of these steps was performed for the two data sets.

2.3.3 English data

We use all the sentences from the Wall Street Journal corpus in the Penn Treebank. Each sentence becomes associated with the respective semantic role annotation extracted from the PropBank/SemLink, and with its dependency parse available from the CoNLL-08 data (see Figure 2.4 below). The sentences which are absent from the SemLink data

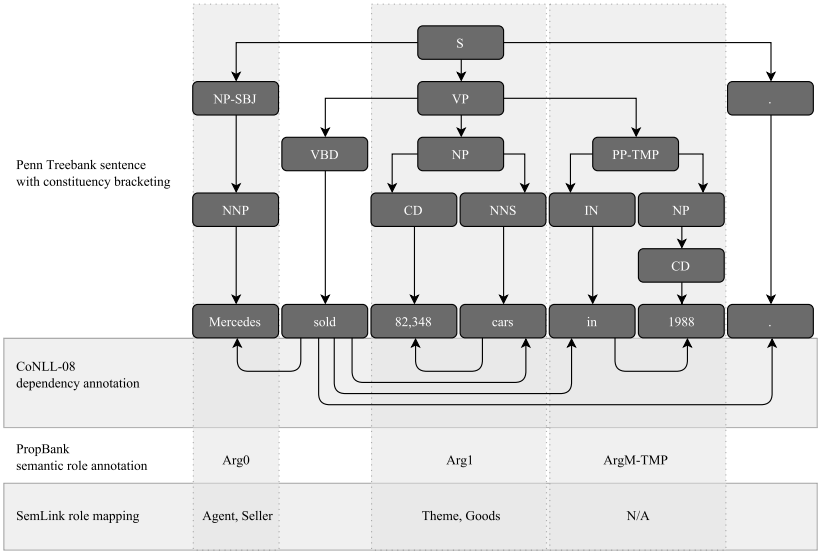


Figure 2.4: An English sentence with various types of annotations.

are immediately filtered out, as well as the sentences which contain arguments with no SemLink mappings between PropBank-style (e.g., ARG0) and FrameNet-style (e.g., AGENT) semantic labels.

Following the consolidation of resources, we start building the argument structure (AS) instances. We iterate over the sentences, considering each frame present in the data. The frame-evoking predicate can be a single personal verb form (as in the example in Figure 2.4), but also other forms or POSs (19–21). Only the frames evoked by a personal verb form are preserved: for each of such forms, a new AS instance is created and assigned the lexical predicate as one of the features.

(19) So would someone recently **divorced** or widowed.

(20) International Paper and Weyerhaeuser declined **to comment**.

(21) By late 1988, they **were banning** Soviet publications.

Predicates in SemLink have FrameNet frames associated with them (e.g., COMMERCE-SELL for the example in Figure 2.4) – the new AS instance is assigned the respective frame name as a feature.

Next, we focus on the semantic annotations of role-bearing constituents in SemLink, which refer to particular branches of constituency trees in the Penn Treebank. Note that we eliminate all the adjuncts, or non-core frame elements (*in 1988* in Figure 2.4), by considering only the numbered arguments; although for the locative verbs *follow, go, lead, live*, and *sit* we also keep the arguments with the semantic roles DIR (direction) and LOC (location) (22). For each core argument in SemLink, we align its constituency branch from the Penn Treebank with the respective dependency branch in the CoNLL

data, as shown in Figure 2.4. Sometimes such an alignment fails, in which case we perform another attempt, using the original PropBank argument annotation instead of the SemLink annotation. If this fails as well, the created AS instance is dismissed.

(22) The Dalai Lama lives in exile **in India**.

The dependency parse is used to identify the head of each role-bearing constituent: the semantic roles in SemLink can be assigned to constituents of any length. For each constituent, we traverse its dependency tree starting from the root, with the goal of finding the first proper lexical argument (e.g., a noun or a pronoun) that will carry the respective semantic role. This step is needed, because the heads of the role-bearing constituent in a dependency tree are often assigned to an adverb (23) or to a preposition (24). In these cases, the following nouns – *stake* in (23) and *treatment* in 24 – will be taken as the respective lexical arguments, while the prepositions (*about*, in this case) are stored independently. If at any time during traversing the tree the currently top branch appears to initiate a clause (25), the AS instance is, again, dismissed. At the end of this step, each lexical argument and preposition are added to the current AS instance as independent features.

(23) The holding acquired **most** of its stake.

(24) The Palestinians complain **about** their treatment.

(25) We want **to be a lot more liquid**.

At this point, we can also determine the relative order of the verb, its arguments, and prepositions, and build a representation of the syntactic pattern (26), which is also added to the AS instance as a feature.

(26) He proposed the movie to his producer.
Syntactic pattern: ARG1 VERB ARG2 *to* ARG3

During the next step, we create a distributional representation of the semantic meaning of each verb and each lexical argument. Just as in the smaller multilingual corpus, we recursively extract from WordNet the hypernyms of each noun (although this time we always use the most common first sense), and add the first element of each synset to the distributional semantic representation of the target word meaning. Since the verb network in WordNet is not as rich as the noun network, we additionally extract all the available features for each verb from VerbNet (27). Besides, adjectives and adverbs are not hierarchically structured in WordNet, for which reason we include all their synonyms into their distributional representations instead. We also manually compile distributional representations for frequent pronouns. All the other words with no features are associated with empty semantic representations. The respective semantic features for the verb and the lexical arguments are added to the AS instance.

(27) sell: {$\underbrace{\text{EXCHANGE, TRANSFER}}_{\text{WordNet features}}$, $\underbrace{\text{HAS-POSSESSION, CAUSE}}_{\text{VerbNet features}}$}

WordNet is also used to expand FrameNet-style semantic role labels into distributional representations, with the help of the existing mapping between FrameNet

Table 2.2: Expanding the role label for FE EVENT in the frame PARTICIPATION: the frequency and the probability of occurrence of each synset for this frame–FE pair in the repository are shown. The two synsets in gray are excluded from the resulting semantic representation.

Synset	Frequency	Probability
COUNTRY-NOUN$_2$	147	.9018
DRILL-NOUN$_1$	4	.0245
SUBSIDIARY-COMPANY-NOUN$_1$	4	.0245
BATTLE-NOUN$_1$	2	.0123
POLITICS-NOUN$_5$	2	.0123
VENTURE-NOUN$_1$	2	.0123
COMPLEX-NOUN$_2$	1	.0061
YELLOWCAKE-NOUN$_1$	1	.0061

and WordNet (Bryl et al., 2012). Recall that this resource provides a repository of synsets for each FE–frame pair. Both frames and FEs are available from SemLink in our data, and for each pair of these we combine all the synsets in the repository to build a distributed representation of the FrameNet semantic label. For some FE–frame pairs the distribution of synsets has a rather long tail, resulting in a very long list of synsets. To avoid this, for FE–frame pairs occurring at least ten times in the repository, we only consider the synsets whose frequency of occurrence for the given pair is higher than 2, and whose probability of occurrence is higher than .01 (see Table 2.2 for an example). Adding the resulting features to the AS instance concludes its assembly.

2.3.4 German data

A similar procedure is carried out for the German data. We start from considering all the sentences present in the SALSA corpus, and associate them with the dependency annotations from the CoNLL-09 data. Just as in English data, this is followed by filtering out the frames evoked by any lexical item other than a personal verb form, but also the predicate-specific proto-frames, which are only present in SALSA, but not in FrameNet (e.g., KENNEN1-SALSA). A new German AS instance is created and assigned a frame name.

Again, we dismiss all the non-core FEs and then identify a lexical argument for each core argument by aligning its constituency tree branch from SALSA with the corresponding dependency tree branch from CoNLL-09. In case of a successful alignment, the dependency tree is traversed in search for a proper lexical argument. If the lexical instantiations of all the core arguments are successfully identified, the lexical features (predicate and arguments) are added to the AS instance, as well as syntactic pattern. Also, the case-marking is available from the original TIGER annotations, and we add the case-marking for each argument as an independent feature to the AS instance.

For the semantic annotation, one option would be to use the GermaNet – a German equivalent of WordNet (Kunze, 2000). However, the difference between the two resources would lead to having inconsistent semantic representations across our English and German data sets, damaging the ecological validity of our computational simulations. Therefore, we translated the verbs and their lexical arguments from English into German. Since our data sets are organized around verb predicates, the verb meanings are rather important, and we translate all the verbs manually, making sure that no two German words obtain the same English translation. The manual translation of all the lexical arguments, in contrast, is not feasible – instead, we translate them by a simple look-up in a German–English dictionary.[3]

Importantly, German words are often translated into English not as single words, but as phrases (e.g., *Wahlverhalten – electoral behavior*). This is troublesome, because the English translations are used afterwards for extracting semantic features from WordNet. Combining semantic features of multiple words into a single set would create very complex semantic representations, incomparable with those in the English data. This is why a single word is required for the semantic feature extraction. Our general approach to this problem is to make sure that the German word and its English translation belong to the same part of speech (POS). For this, the English phrase is POS-tagged using the Stanford tagger (Toutanova, Klein, Manning, & Singer, 2003), while the POS of the original German word is available from the SALSA annotations. In the phrase described above, comparing the POS-tags helps us to select the key word: *behavior*. However, German compounds are often translated as a combination of two nouns (e.g., *Produktinnovation – product innovation*). In this case, only the last English noun is taken (*innovation*). When multiple synonyms are available for translation (e.g., *Antwort – reaction, reply, response*), we use the one whose lemma occurs in the WSJ most frequently. In case the German compound noun is not present in the word list, we try splitting it into subparts using jWordSplitter,[4] and look up the last German noun in the word list. For each word translated into English, its semantic representation is built using WordNet, as described in the previous section. This semantic representation is then mapped back to the original German word. The semantic features are added to the AS instance.

Since the frames and FEs in SALSA are consistent with FrameNet, we extend the semantic role labels into distributional sets in the exact same way as for the English data. The resulting features are added to the AS instance, which concludes the process.

2.3.5 Resulting data set

The resulting data sets are comparable both in type of language (newspaper texts) and in size (3,624 AS instances in the English subcorpus, and 3,370 in the German subcorpus). The number of verb types is also comparable: 319 English vs. 301 German verbs. The distribution of the most frequent values for some of the features are shown in Figure 2.5.

[3] Available from http://www.dict.cc/

[4] http://www.danielnaber.de/jwordsplitter/

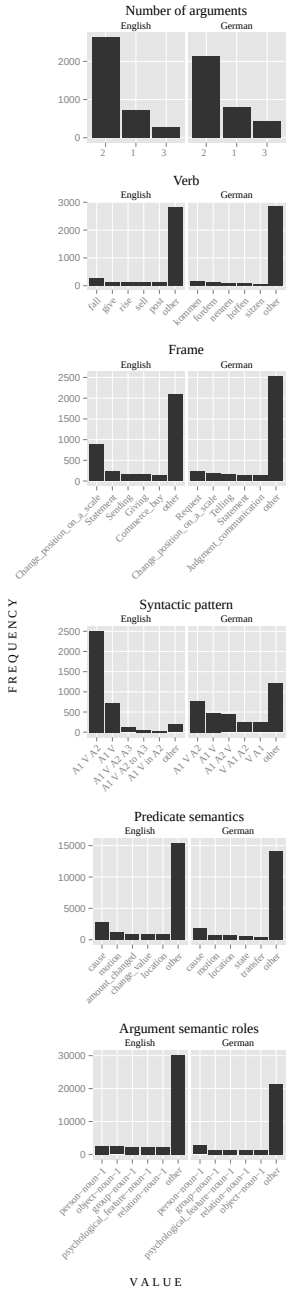


Figure 2.5: Distribution of feature values across languages.

Just as in the small corpus, instances with two arguments prevail. The most frequent verbs as well as the most frequent frames reflect the type of language used in the WSJ and in *Frankfurter Rundschau*. Note that the WSJ is more focused on business and financial news, hence the frequent occurrence of verbs such as *fall, rise* (about the price) and *sell*, as well as frames such as CHANGE POSITION ON A SCALE and COMMERCE BUY in our English subcorpus. In contrast, *Frankfurter Rundschau* is a general-interest newspaper, and more common verbs (*kommen* "to come", *fordern*, "to demand", etc.) are the most frequent ones, while the most frequent frames include REQUEST, TELLING, STATEMENT, etc. The more specialized frame CHANGE POSITION ON A SCALE is also frequent, but is not as prevailing as in the English subcorpus.

The distribution of syntactic patterns differs across the two languages: the neutral word order A1 V A2 clearly dominates in English, while in German the alternative word order A1 A2 V is frequent as well. This pattern mostly originates from sentences with the verb preceded by a clause or an adjunct eliminated from our analysis (28).

(28) **Heute** weiss man es besser!
today know.3SG one it better
'Today we know it better.'

In terms of predicate semantics, the most frequent verbs in both the English and the German subcorpus are those of *caused motion*. Finally, the distribution of argument semantic primitives is rather uniform in both subcorpora, the top five most frequent synsets refer to the first WordNet senses of nouns *person, object, group, psychological feature*, and *relation*.

2.4 Conclusion

We have presented two corpora of verb usages annotated with the argument structure information. The two corpora complement each other: the manually annotated corpus is based on child-directed speech and is free of noise, but is relatively small, while the automatically compiled corpus is extracted from newspaper texts and is larger, but contains certain noise because of the compilation procedure. In particular, word sense disambiguation has not been performed on verbs and their arguments.

One advantage of the presented corpora is their feature-based structure: the necessary features related to the verb argument structure can be easily extracted and used in computational modeling. Additionally, features can be combined or dismantled. To give an example, the syntactic pattern (29a) can be dismantled into multiple features, such as a more abstract pattern and a preposition (29b) or a number of positional features (29c).

(29) He gave a toy to his friend yesterday.
   a. Syntactic pattern: ARG1 VERB ARG2 TO ARG3
   b. Syntactic pattern: ARG1 VERB ARG2 PREP ARG3
      Prepositions: *to*

c. Verb position: 2
Arg1 position: 1
Arg2 position: 3
Arg3 position: 4
Arg1 preposition: N/A
Arg2 preposition: N/A
Arg3 preposition: *to*

Importantly, each corpus includes data from more than one language, and provides a rare combination of syntactic and fine-grained semantic features. This makes the corpora suitable primarily for the development of computational models of multilingual learning, as well as of automatic systems for cross-lingual semantic (SRL) and syntactic parsing, which is timely given the recent interest in this latter domain (Akbik et al., 2015; van der Plas, Apidianaki, & Chen, 2014; Kozhevnikov & Titov, 2013; M. Lewis & Steedman, 2013, etc.). Other possible applications include various natural language understanding tasks based on SRL: information extraction (Surdeanu, Harabagiu, Williams, & Aarseth, 2003), question answering (Shen & Lapata, 2007; Sun et al., 2005), etc.

Speaking about computational models of multilingual learning, the corpora have already been used to sample the input data to the model of learning argument structure constructions from bilingual input (see chapters 3–5). To give a brief example, in chapter 4 I train and test the model using the automatically compiled corpus, in order to study how second language (L2) proficiency depends on the amount of linguistic input, and on the moment of L2 onset. Making use of the existing feature structure in the corpus, I employ a battery of test tasks: in each of these tasks the value of one of the features (e.g., verb or syntactic pattern) was masked, and the model has to predict the most probable value based on the acquired linguistic representations. In chapter 5, I make use of the four languages (excluding L2 English) from the smaller corpus to study cross-linguistic influence in argument structure constructions. I train the model on different pairs of languages (English–German, French–German, French–Russian, etc.), focusing on the phenomenon of case-marking comprehension in German and Russian. The results demonstrate that positive and negative cross-linguistic influence contribute to the model's performance in a case-marking comprehension task.

The small manually annotated subcorpus is published online, while the larger automatically compiled corpus can be generated from the existing resources using the available code.[5]

[5] http://ilk.uvt.nl/~yevgen_m/#data

CHAPTER 3

Modeling verb selection within argument structure constructions[1]

3.1 Introduction

Speakers' language use is conditional on the linguistic means they possess. In a way, an individual's language use provides us with a "window to the mind" (Gilquin, 2010): linguistic representations are studied through language use (see a review by Clahsen, 2007). At the same time, one of the tenets of cognitive linguistics is that linguistic knowledge is directly grounded in previous usage events (e.g., Kemmer & Barlow, 2000). Such events include both language production and comprehension, thus an individual's language use depends to a certain extent on the properties of the input (s)he has been exposed to. Indeed, it is known that input-related (e.g., distributional) properties of a linguistic unit affect how this unit is used or processed (e.g., Gor & Long, 2009; N. C. Ellis, 2002; Hoff & Naigles, 2002). But to determine the importance of various input-related factors, we need formal models predicting language use from multiple factors at once.

In this chapter, we study the processing of argument structure constructions through a verb production task. In the traditional view of argument structure, the term describes how the arguments of a predicate (typically a verb) are realized: the verb *eat* involves two participants, hence two arguments; importantly, the verb is believed to predict its structure (Haegeman, 1994). In constructionist accounts, in particular Goldberg's construction grammar (Goldberg, 2006; Goldberg, Casenhiser, & Sethuraman, 2004;

[1] This chapter is derived in part from an article published in *Language, Cognition and Neuroscience* 30 June 2016 © Taylor & Francis, available online: http://dx.doi.org/10.1080/23273798.2016.1200732

Goldberg, 1995), argument structures obtain properties independent of particular verbs through the emergence of abstract argument structure *constructions*, a particular type of linguistic constructions (or form–meaning pairings) that "provide the means of clausal expression" (Goldberg, 1995, p. 3): for example, the verb *eat* often participates in a transitive construction, which has the form SUBJ VERB OBJ and the meaning X *acts on* Y. Such constructions slowly emerge in a learner's mind as (s)he categorizes individual verb instances. Although this is a simplistic description, argument structures can be seen as verb-centered mental categories (Goldberg, Casenhiser, & Sethuraman, 2005, 2004), where a variety of verbs may occupy the central slot in each construction.

The studies mentioned above investigate, among other things, the role of individual verbs and their properties in formation of argument structure constructions, considering their abstract nature. Within a given construction, speakers prefer some verbs over others. In particular, some verbs within a construction are produced more frequently than others, they come to mind first, and they are learned earlier (e.g., N. C. Ellis & Ferreira-Junior, 2009; Goldberg et al., 2004; Theakston, Lieven, Pine, & Rowland, 2004; Ninio, 1999b; Naigles & Hoff-Ginsberg, 1998): e.g., the SUBJECT VERB LOCATION construction attracts such verbs as *go*, *come*, and *get*, while *sleep* and *telephone* are rather rare (data from N. C. Ellis & Ferreira-Junior, 2009). Two groups of factors have been considered to predict verb preference: distributional and semantic factors, yet there is no conclusive evidence on the exact contribution of each factor. At the same time, it is important to reveal their exact contributions, in order to better understand the underlying nature of links between verbs and constructions in speakers' minds. Understanding which input properties enable individual verbs to group into constructions would contribute to our knowledge about the mental grammar, or "constructicon".

Our goal in this study is to evaluate the role of specific distributional and semantic factors. As a methodological tool, we use a computational model of construction learning. Computational models enable us to overcome some of the methodological limitations imposed by studying human subjects and, as a result, make informed predictions about the role of some of the proposed factors. Ultimately, our study endeavors to propose a refined prediction model explaining verb selection in argument structure constructions. This will help us to understand which factors are responsible for the emergence of links between verbs and constructions in the minds of language users.

The chapter is organized as follows. In the next section (3.2.1) we review some existing studies on the issue, motivate our focus on particular studies (N. C. Ellis et al., 2014a, 2014b), and expose two methodological issues that we plan to address. We also introduce distributional and semantic factors considered in the study, and explain why these factors may be important (3.2.2). This is followed in section 3.3 by the description of the setup of our study: computational model, input data, test stimuli, and the exact predictor variables representing the distributional and semantic factors under consideration. Section 3.4 consists of three studies: the first one is intended to simulate the original experiments: we demonstrate a reasonable performance of our model in the target task, and fit a regression explaining this performance as a function of the predictor variables. The second study addresses two methodological issues: we show how the regression coefficients change when each of the issues is resolved. In the final

study (section 3.4.4) we consider alternative combinations of predictor variables that may better explain the model's performance in the target task. Section 3.5 summarizes the chapter, and is followed by a short conclusion 3.6.

3.2 Theoretical overview

3.2.1 Predicting verb selection

N. C. Ellis et al. (2014a, 2014b), henceforth EOR, provided native and non-native English speakers[2] with a set of stimuli, which schematically represented argument structure constructions with a verb missing: *it ___ about the...*, *s/he ___ across the...*, *it ___ as the...*, etc. Each stimulus was presented both with an animate (*he* or *she*) and with an inanimate (*it*) pronoun. Participants had to spend a minute to produce verbs fitting the slot. Note that EOR's stimuli have a very weak semantic component: they are, in fact, form-based patterns, and participants are free in their interpretations of the arguments' thematic roles. Römer, O'Donnell, and Ellis (2015) motivate such an approach by the fact that they analyze semantic associations between verbs and constructions, and therefore it is "important to initially define the forms that will be analyzed in a semantics-free, bottom-up manner" (p. 45). Although this is a controversial point (and we return to it in the discussion), in this study we follow their approach.

Importantly, this task is used to investigate the acquired associations between verbs and constructions, and it is not suitable for studying language production as such. In production speakers start from the intended meaning, and then encode this meaning using some of the suitable forms (words, grammatical patterns, etc.). In contrast, EOR's participants are cued with a pattern with little semantic information and have to select a verb (that is, a form and a meaning at the same time) that fits the pattern. In this capacity, the task is similar to other psycholinguistic tasks often used for studying human memory, implicit knowledge of words, and mental grammar: the fill-in-the-blank (cloze) task, the free word association task, and the cued recall task (see Shaoul, Baayen, & Westbury, 2014, for a review).

Following the task, the cumulative frequency of production of each verb in each construction was calculated. Statistical analyses revealed that the cumulative production frequency could be predicted from three input variables – verb frequency in the construction, contingency of verb–construction mapping, and prototypicality of verb meaning – with an independent contribution of each variable. Here we only briefly define the variables, more information on each of them is given below (section 3.2.2).

- Verb frequency in the construction: how frequently a verb appears within a specific construction in the linguistic input.

[2] Note that in EOR's setup virtually no distinction is made between first (L1) and second language (L2) speakers. This is in line with the theories of incidental (statistical) language learning, and with the proposal in cognitive linguistics that much of the L2 learning relies on the same cognitive mechanisms used in L1 learning (MacWhinney, 2012; N. C. Ellis & Larsen-Freeman, 2006; Ervin-Tripp, 1974).

- Contingency of verb–construction mapping: to what extent the use of a specific construction is indicative of a particular verb, compared to other constructions/verbs.
- Prototypicality of verb meaning: how representative the verb meaning is for the general semantics of a construction.

Some of these findings are in line with some existing studies in language acquisition, which look at verb production by children. In particular, the verb frequency effect has been also found by Naigles and Hoff-Ginsberg (1998), Ninio (1999a), and Theakston et al. (2004). However, Ninio (1999a) suggests that the effects of frequency and prototypicality are not independent, and Theakston et al. (2004) find no effect of prototypicality after the frequency is accounted for.

Additionally, there is a number of studies carried out by Ambridge and colleagues, who investigate whether distributional and semantic factors help children and L2 learners to learn restrictions for the verb use in various argument structure constructions (Ambridge, Bidgood, Twomey, et al., 2015; Ambridge, Pine, Rowland, Freudenthal, & Chang, 2014; Ambridge & Brandt, 2013; Ambridge, Pine, & Rowland, 2012, etc.). Although these studies mostly use grammaticality judgments, a production experiment has been reported as well (Blything, Ambridge, & Lieven, 2014). This line of research demonstrates the role of both distributional and semantic factors in construction learning. Their results in terms of the role of distributional factors are consistent with other studies mentioned above. As for the role of semantics, Ambridge and colleagues in their studies use a very different interpretation of verb semantics, focusing on fine-grained discriminative features of the verb meaning, which are based on Pinker's (2013) verb classes (we return to this issue in the final discussion). This makes it difficult to compare their findings in terms of verb semantics to what other studies report.

In short, there is no conclusive evidence about the exact contribution of each specific factor to explaining the verb use within argument structure constructions. We focus on the studies of EOR, because they investigate both groups of factors on a large set of constructional patterns.

Methodological issues

There are two potential methodological issues in EOR's analyses, which may have some implications for the ecological validity of their studies. The first issue relates to how the values of the predictor variables (in particular, frequency and contingency) are obtained. All input estimates are based on the British National Corpus (BNC). Although the use of large corpora for approximating language input to learners is rather common and well justified overall, the method has certain shortcomings when it comes to accounting for the individual variation between speakers (e.g., Blumenthal-Dramé, 2012). The variation in individual experiences with a language may lead to the formation of different linguistic representations in learners (Dąbrowska, 2012; Misyak & Christiansen, 2012). The variation is even higher among L2 learners, whose learning trajectories may vary greatly (e.g., Grosjean, 2010). In EOR's case, verb production data obtained from multiple individuals are predicted by input-related measures computed

from a corpus, which is, again, generated by a language community. This way, EOR demonstrate that their model predicts verb selection on the population level. But cognition is individual, and for making informed claims about cognitive representations we need to test the selection model on the input to individual speakers and individuals' production data. This is a challenging task for studies with human subjects, because it is nearly impossible to account for the whole learning history of an individual.

Another issue we focus on relates to the use of cumulative frequency of verb production. Calculating the total number of times each verb has been produced by all the speakers in a specific construction results in losing the information about the order of production. Yet, the order of verb listing must also be taken into account. For example, the verb position in a produced list has been shown to correlate with the frequency of production of this verb in a category-listing task (Plant, Webster, & Whitworth, 2011). Similarly, studies on sentence production show that, all things being equal, the more accessible (prototypical, frequent) word in a word pair tends to be placed earlier in a sentence than the less accessible one (e.g., Onishi, Murphy, & Bock, 2008; Bock, 1982). These findings suggest it is important to account for the order of verb production in the experimental task described above. In fact, EOR briefly mention this issue among the limitations of their study.

One objective of the current study is to simulate EOR's experiments using the computational model of argument structure construction learning (Alishahi & Stevenson, 2008). The second objective is to test whether the findings of EOR still hold after addressing the two methodological issues described above; the computational model is particularly helpful in this respect. First, it provides us with control over the input to each simulated learner, and eliminates other possible sources of individual variation, related to learners' cognitive abilities, propensities, etc. (R. Ellis, 2004). Second, the model generates the probability of production of each verb, which makes it easy to account for the order of verb preference (see section 3.3.4 below).

Our final objective relates to the original prediction model, which uses frequency, contingency and prototypicality to explain verb selection. Based on some theoretical premises presented in the next section, we propose a refined prediction model in the current study, and show that it may have a higher explanatory power than EOR's original model. We proceed with a critical overview of the three variables used in the original experiments.

3.2.2 Factors affecting verb selection

Input frequency

Language learners are sensitive to frequencies of occurrence of linguistic units in the input. Frequency effects have been demonstrated in many domains of language processing and language use (see overviews by Ambridge, Kidd, et al., 2015; Divjak & Caldwell-Harris, 2015; Lieven, 2010; Diessel, 2007). Frequencies also relate to the concept of entrenchment in cognitive linguistics: more frequent words (in this case, verbs) get entrenched stronger in learners' minds, which makes them more accessible (Schmid, in press; Bybee, 2006; Langacker, 1987). Although the existence

Table 3.1: A verb–construction contingency table.

	Target construction	**Other constructions**	**Total**
Target verb	a	b	$a+b$
Other verbs	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

of frequency effects is commonly recognized in cognitive linguistics, it is unclear yet which frequencies count (N. C. Ellis, 2012): of a particular word form (*goes*), of a lemma (all occurrences of *go*, *went*, etc.), of a form used in a specific function (*go* as an imperative), of an abstract meaning alone, etc. The frequency effect may also depend on the level of granularity of the examined units (Lieven, 2010). The complexity of the issue is reflected in the number of different kinds of frequencies discussed in the literature:

- Token vs. type frequency (Bybee & Thompson, 1997): the number of occurrences (tokens) of a specific lexical unit in a corpus vs. the number of various specific units (types) in a corpus matching a given abstract pattern.

- Absolute vs. relative frequency (Schmid, 2010; Divjak, 2008): the absolute measure denotes the independent frequency of a unit (e.g., the verb *go* has been produced 25 times in the construction *he/she/it* VERB *across* NOUN), while the relative measure relates the frequency of the target unit to the frequencies of competitor units, capturing this way paradigmatic relations of the units (e.g., the verb *go* takes a 10 percent share of all the verb tokens produced in the construction *he/she/it* VERB *across* NOUN). This difference between the measures has to do with the notion of contingency (association strength), discussed in more detail in the next section. It is useful to visualize it using a verb–construction frequency (or contingency) table (see Table 3.1): the absolute verb frequency is expressed as $a+b$, while the relative frequency must relate this value to the frequency of competing verbs, $c+d$.

- Marginal vs. joint frequency: unlike the previous pair, this distinction concerns the syntagmatic relations of two units. A unit's marginal frequency is its overall frequency in a corpus (e.g., the verb *go* occurs in the BNC approximately 86,000 times); also sometimes referred to as "raw frequency". In Table 3.1, the marginal frequency of the target verb is denoted as $a+b$, and the marginal frequency of the target construction is $a+c$. The joint frequency a, on the other hand, denotes how frequently the target verb occurs in the target construction (e.g., the verb *go* in the construction SUBJ VERB *across* LOC occurs in the BNC approximately 120 times).

This last distinction requires further attention here. EOR in their analysis always employ the joint verb–construction frequency as one of the predictors. This measure

has been considered in studies of some linguistic behaviors, such as acceptability judgments (e.g., Divjak, 2008), as well as in language acquisition (e.g., Theakston et al., 2004). However, these studies also take into account the marginal verb frequency. In particular, Ambridge, Kidd, et al. (2015) argue that both types of frequencies affect child language learning. Talking about production in particular, Blything et al. (2014) carried out a production experiment with children, and used, among others, measures called "entrenchment" and "preemption" to predict the probability of verb production. Both measures were based on the overall frequency of a verb (or verbs) in the BNC, and their observed effects also support the idea that the marginal verb frequency is important. This idea is also in line with the theoretical account of units' entrenchment in the cognitive system, proposed by Schmid and Küchenhoff (2013), Schmid (2010). They distinguish between cotext-free and cotextual entrenchment: while cotext-free entrenchment is related to the marginal item frequency, cotextual entrenchment captures syntagmatic associations between items, just as the joint frequency of two items does.[3] For measuring the syntagmatic association strength, various association measures have been proposed, which we discuss in the next section.

At this point it is important to note that the verb selection model of EOR does not take into account the marginal verb frequency, and we believe that including this variable in the model could improve it. EOR motivate their exclusion of the marginal verb frequency ("raw", in their terminology) by the fact that verb selection in their test correlates better with the joint verb–construction frequency than with the marginal verb frequency. But assuming the potentially independent effects of the two kinds of frequencies, the inclusion of the marginal verb frequency into the model may be justified.

Contingency of mapping

The second factor in EOR's model is contingency, or the reliability of verb–construction mapping. Although EOR use a particular measure explained below, contingency is an umbrella term for multiple measures of the association strength between a particular verb and a particular construction. The notion of contingency comes from the paradigm of human contingency learning, focusing on learning associations between stimuli, which are often described in terms of cues and outcomes. The term is rarely used in linguistic studies, which prefer talking about association strength, or about "contextualized" frequency measures (Divjak & Caldwell-Harris, 2015). Joint verb–construction frequency is the simplest example of such a measure, while other measures represent more sophisticated ways to quantify how well a verb and a construction go together. Therefore, we argue that the simultaneous use of two contingency measures within the same model may be redundant.

In various disciplines, the impact of contingency has been shown to be independent from that of frequency. In particular, some classical models of memory recall implement the effects of frequency and association strength independently of one another (Gillund

[3] We follow the existing literature in assuming that the entrenchment of a unit is a mere product of its frequency, although the impact of each individual use may, in fact, be strongly modulated by pragmatics (Schmid, in press).

& Shiffrin, 1984; J. R. Anderson, 1983). Studies on item- versus association-memory in word retrieval also indicate that these two types of memories are independent of each other (e.g., Madan, Glaholt, & Caplan, 2010; Hockley & Cristi, 1996). However, these studies talk about the marginal item frequency, which, as we have mentioned, deals with an item in isolation. Therefore, the mentioned studies can hardly be used as an argument in favor of the independent effects of *joint* frequency and contingency within the same model.

The second issue related to contingency has to do with the ongoing discussion in cognitive linguistics about which contextualized measure has a higher predictive power (Gries, 2015; Küchenhoff & Schmid, 2015; Gries, 2013; Schmid & Küchenhoff, 2013; Bybee, 2010; Divjak, 2008; Stefanowitsch & Gries, 2003). Just as in the previous section, these measures are commonly presented using a contingency table (see Table 3.1). Despite a great number of proposed association measures (see overviews by Pecina, 2010; Wiechmann, 2008; Evert, 2005), we can make a simple distinction between three types, based on how many of the table cells *a*–*d* the measure takes into account e.g., Divjak and Caldwell-Harris, 2015; Divjak, 2008:

1. Raw joint frequency (cell *a*) is the most intuitive way to measure how well a verb and a construction go together: the verb *go* in the construction SUBJ VERB *across* LOC occurs in the BNC approximately 120 times.

2. Conditional probabilities relate the joint frequency to the marginal token frequency of either a construction ($Attraction = \frac{a}{a+c}$) or a verb ($Reliance = \frac{a}{a+b}$). Such normalization of the raw joint frequency is useful when, for example, multiple constructions with different frequencies are studied: the same number of 120 occurrences of a particular verb may account for 90 percent of all verb usages in one construction, but only for 10 percent in another one.

3. Complex associative measures take into account all the four cells *a*–*d*. An example of such a measure is $\Delta P_{Attraction}$, or $\Delta P(construction \rightarrow word) = \frac{a}{a+c} - \frac{b}{b+d}$, which is used in the original studies of EOR. Other popular measures include, e.g., Minimum Sensitivity (Wiechmann, 2008) and the *p*-value of Fisher–Yates exact test (Stefanowitsch & Gries, 2003). The use of such measures can be motivated by the need to capture the competition between the verbs and the constructions at the same time, in particular to address the problem of hapax legomena. For example, in a study of *as*-predicative (Gries, Hampe, & Schönefeld, 2005) the unrepresentative verb *catapult* scored highest in Reliance among many other verbs, only because it never occurred in other constructions in the corpus. The use of a complex measure solved the problem in their case. At the same time, other researchers (e.g., Schmid & Küchenhoff, 2013; Blumenthal-Dramé, 2012; Divjak, 2008) suggest that complex measures may have little advantage over the conditional probabilities (type 2 above).

To summarize, we think that including both joint frequency and ΔP (or any other contingency measure) into the model, as in EOR's studies, may not be well justified. We suggest that only one such measure should be considered in the analysis, while the

other is redundant. In the current study we consider one measure of each type specified above, as well as their combinations, to test which of them predicts verb selection better.

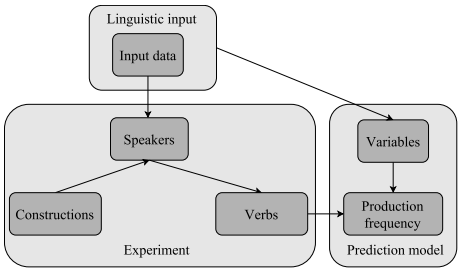
Semantic prototypicality

Semantic prototypicality is a concept borrowed from studies on category structure; it is also known under alternative names, such as "family resemblance" (Rosch & Mervis, 1975), "goodness-of-example" (Mervis, Catlin, & Rosch, 1976), "typicality", "goodness of membership" (Onishi et al., 2008), etc. It is common in cognitive science to estimate the typicality of concepts within a semantic category using so-called category norms – ranked lists of items based on human production data e.g., Plant et al., 2011; Kelly, Bock, and Keil, 1986. EOR, however, do not use this approach, as it would lead to circular reasoning: prototypicality is used to predict the production data, and thus cannot be computed based on other production data. Instead, for each considered construction (e.g., *he/she/it* VERB *across* NOUN) they build a semantic network of verbs participating in this construction (*go*, *move*, *face*, *put*, etc.). This network is organized according to the similarity of verb meanings, as informed by WordNet (G. A. Miller, 1995). Using a network for a particular construction, they compute a measure called betweenness centrality, which indicates the centrality of each verb's meaning in this construction. This way, the most general verbs in the construction (in this case, *go* and *move*) tend to obtain higher prototypicality values (see Gries & N. C. Ellis, 2015; Römer et al., 2015, for more detail). In this sense, "semantic generality" would be a more suitable term, however we follow EOR and other studies mentioned next in using the word "prototypicality". An additional advantage of EOR's method to compute prototypicality is that the resulting values are independent of the corpus-based frequency and contingency measures.

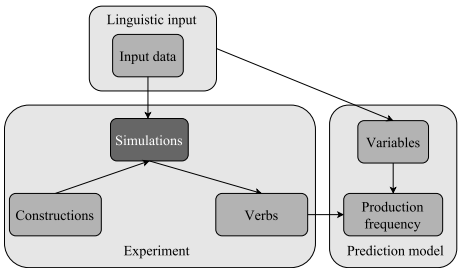
Semantic prototypicality has also been studied in language acquisition research: semantically general verbs have been suggested to be "pathbreaking" in child language use (e.g., Ninio, 1999a, 1999b). However, semantic generality is often confounded with input frequency: general verbs tend to be used most frequently (Goldberg et al., 2004; Ninio, 1999a), and the independent effect of semantic generality is not always found (Theakston et al., 2004). At the same time, EOR argue that the effect of semantic prototypicality is independent of frequency: while frequency relates to entrenchment, prototypicality has to do with the spreading activation in semantic memory (J. R. Anderson, 1983): if verbs within a construction form an interconnected network, then more central (general, prototypical) verbs in this network are more likely to be activated, and thus to be produced. To summarize, there is no conclusive evidence on whether the semantic prototypicality of a verb is a good predictor of its use.

Summary

This theoretical overview shows that the role of both the distributional (frequency, contingency) and the semantic factors (prototypicality) requires further research. In particular, it is unclear yet whether marginal verb frequency plays an independent



(a) Original study



(b) Our initial study: computational simulations replace human speakers.

Figure 3.1: Design of EOR's study and its simulation; updated components are marked with a darker color.

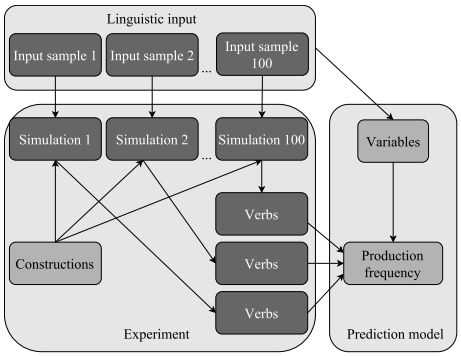
role in predicting verb selection; which measures of contextual frequency should be included into a prediction model, and how many of such measures; finally, the role of semantic prototypicality is under discussion. We will address these issues in our study, but first we proceed with its methodological description.

3.3 Material and methods

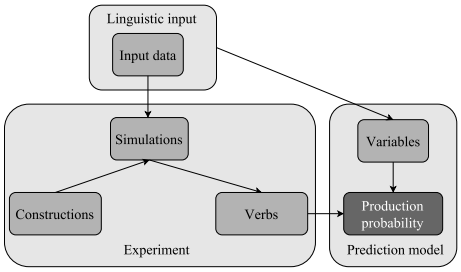
3.3.1 Study overview

Figures 3.1–3.3 present a schematic overview of the design employed in the original studies and in the present study, the latter being divided into three main steps. Only a brief summary for each step is given here, while more detail can be found in the respective sections below.

There are three main blocks of the original study: (1) experiment, (2) linguistic input, and (3) prediction model (Figure 3.1(a)). During the experiment, L1 or L2 speakers are exposed to a set of constructions with the main verb missing, and produce a set of verbs. Three predictor variables are extracted from the BNC, under the assumption that this corpus provides an approximation of the linguistic input that participants have been exposed to in their lifetime. These variables are then used in the prediction model to explain the frequency of production of verbs within constructions.



(a) Accounting for individual differences: specific input samples and individuals' production lists are used.



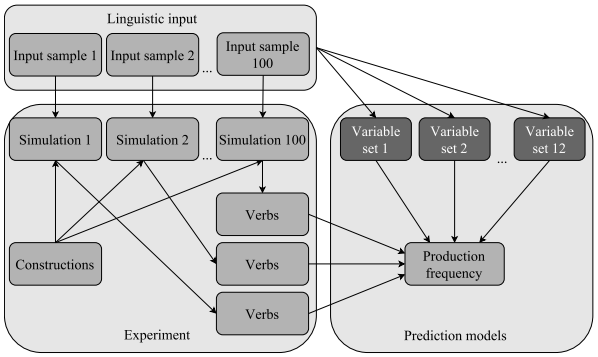
(b) Accounting for order of preference: production probability replaces production frequency.

Figure 3.2: Analyses addressing methodological issues; updated components are marked with a darker color.

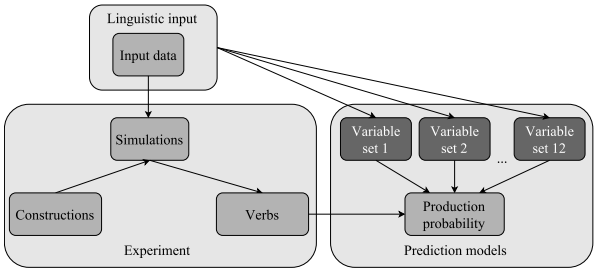
The overall design of our first step (Figure 3.1(b)) is almost identical, except we use computational simulations instead of human speakers, and different data sets. The goal of this step is to check the validity of our computational model; that is, to see whether it selects verbs that fit the target constructions, and whether such selection can be explained by the same input-related features as in EOR's experiments.

At step two we address the methodological issues described earlier (Figure 3.2). First, we distinguish between individual input samples instead of generalizing over the whole population (section 3.4.2 below, also Figure 3.2(a)). Second, in a parallel analysis we employ the production probability instead of production frequency, to account for the order of verbs produced by speakers (more detail below, in section 3.4.3, also Figure 3.2(b)).

At the final step three we test various prediction models to select the one that explains the simulated data sets best, using the two types of design from step two (see Figure 3.3). The following sections describe the essential components of the study: computational model, input data, experimental setup, and predictor variables.



(a) Models accounting for individual differences: alternative sets of predictors are considered, cf. Figure 3.2(a).



(b) Models accounting for order of preference: alternative sets of predictors are considered, cf. Figure 3.2(b).

Figure 3.3: Refining the prediction model; updated components are marked with a darker color.

3.3.2 Computational model

The model used in the current study is based on a model of human category learning, which was shown to replicate multiple experimental findings in this area (J. R. Anderson, 1991). Alishahi and Stevenson (2008) employed the same learning algorithm for simulating early learning of argument structure constructions (which is sometimes seen as a categorization task: Goldberg et al., 2004). The model of construction learning demonstrated similarity to human data in terms of U-shaped learning patterns, use of syntactic bootstrapping (both in production and comprehension), phenomena of over-generalization and recovery (Alishahi & Stevenson, 2010, 2008).

The model relies on some theories of cognitive linguistics and construction grammar, in particular those of Tomasello (2003), Goldberg (1995); for more details, see Alishahi and Stevenson (2008). Most importantly, the input is processed iteratively, so that constructions gradually emerge from categorizing individual instances item by item (similar to the theory described by Goldberg et al., 2004). At the end of the learning process, the model uses its knowledge of argument structure constructions in the elicited verb production task. While the learning model has been used before, the implementation of the test task for this model is novel. We describe these steps in more detail.

Input representations

The model is exposed to a number of instances, each of which represents a single verb usage in a specific construction. Each instance comprises several information cues characterizing the respective verb usage. Table 3.2 shows such a usage, with the full set of features listed in the left column.

We make a simplifying assumption that the model can infer the values of all the provided features from the utterance and the respective perceptual context. This means, in particular, that the model can recognize the words in the utterance and infer their meanings and linguistic cases (where appropriate),[4] as well as to identify the role of each participant in the described event.

Each feature F_k is assigned a value within an instance I, so that I is a unique combination of specific feature values (F_k^I). Following some linguistic theories e.g., McRae, Ferretti, and Amyote, 1997; Dowty, 1991, features expressing semantic and thematic role properties are represented as a set of elements each, and these sets were semi-automatically obtained from the existing resources (see section 3.3.3 below). Regarding the thematic roles, it has been shown that the model used in this study can learn representations of "traditional" thematic roles (e.g., AGENT, THEME) from distributed sets of properties (Alishahi & Stevenson, 2010). A distributed representation of the thematic roles in the current study provides at least two advantages over representing each role as a single symbol. First, set representations enable the model to estimate

[4] Note that we do not assign case-marking to personal pronouns (e.g., *me* = *I*-ACC), but use the actual forms used in the corpus instead. Given the exceptionally high token frequencies of these forms, it is sometimes argued that forms such as *I* and *me* co-exist in the speaker's lexicon, without *me* being derived from *I* (e.g., Diessel, 2007; Hudson, 1995).

Table 3.2: An instance for the verb usage *We sold the house.*

Feature	Value
Head predicate	*sell*
Predicate semantics	{EXCHANGE, TRANSFER, POSSESSION, CAUSE}
Number of arguments	2
Argument 1	*we*
Argument 2	*house*
Argument 1 semantics	{REFERENCE, PERSON ..., ENTITY}
Argument 2 semantics	{DWELLING, HOUSING ..., BUILDING}
Argument 1 thematic role	{COMPANY (N_1), PERSON (N_1) ..., CIVILIZATION (N_1)}
Argument 2 thematic role	{RELATION (N_1), MATTER (N_3) ..., OBJECT (N_1)}
Argument 1 case	N/A
Argument 2 case	N/A
Syntactic pattern	ARG1 VERB ARG2

how similar lexical meanings or thematic roles are to each other. Second, computing the semantic prototypicality of a verb is rather straightforward for set representations of verb meanings (see section 3.3.5). As can be seen in Table 3.2, each verb meaning is represented as a set of semantic primitives describing this meaning: e.g., {EXCHANGE, TRANSFER, POSSESSION, CAUSE} for the verb *sell*. These elements are automatically extracted from available sources (see section 3.3.3). An argument structure construction (henceforth ASC) emerges as a generalization over individual instances, where each feature contributes to forming the generalization. An ASC combines the feature values from all the participating instances, but it is impossible to recover individual instances from an ASC (unless it only contains a single instance). An individual instance is a set F^I of feature values F^I_k ($F^I_k \in F^I$), and an ASC S is a set F^S of feature values F^S_k ($F^S_k \in F^S$), but in an ASC each feature value ($e \in F^S_k$) may occur more than once, depending on the number of participating instances with the value $F_k = e$.

Learning mechanism

The learning is performed using an unsupervised naive Bayes clustering algorithm. As we mentioned, the model receives instances one by one, and its task is to group the incoming instances into ASCs by finding the "best" ASC (S_{best}) for each given instance I:

$$S_{best}(I) = \underset{S}{\operatorname{argmax}}\, P(S|I) \tag{3.1}$$

In other words, the model considers each ASC it has learned so far, seeking the most suitable category for the encountered instance. It makes little sense to talk about the probability of an ASC (prior knowledge) given an instance (new evidence), therefore, the Bayes rule is used to estimate the conditional probability in equation (3.1):

$$P(S|I) = \frac{P(S)P(I|S)}{P(I)} \tag{3.2}$$

The denominator $P(I)$ is constant for each ASC, and therefore plays no role in making the choice. The choice of ASC for the new instance is affected by the two factors in the numerator:

1. The prior probability $P(S)$, which is proportional to the frequency of the ASC in the previously encountered input (or the number of instances that the ASC contains so far, $|S|$):

$$P(S) = \frac{|S|}{N+1}, \tag{3.3}$$

   where N is the total number of instances encountered by that moment. The learner always has an option to form a new ASC from a given instance. Although initially such a potential ASC contains no instances, its value $|S|$ is assigned to 1, to avoid 0s in the multiplicative equation (3.2). The determining role of frequency is grounded in usage-based linguistics: a frequent ASC is highly entrenched and is easier to retrieve from memory, so that new instances are more likely to be added to it.

2. The conditional probability $P(I|S)$, which takes into account how similar an instance I is to S. The higher the similarity between I and S, the more likely I to be added to S: this is based on studies pointing to the importance of similarity in categorization tasks (e.g., Sloutsky, 2003; Hahn & Ramscar, 2001). The model compares each instance to each ASC by looking at the independent features listed in Table 3.2, such as the head predicate, argument roles, etc. For example, all being equal, two usages of the same verb are more likely to be grouped together than two usages of different verbs, yet this can be compensated by other features. Technically speaking, the overall similarity is a product of similarities for individual features:

$$P(I|S) = \prod_{k=1}^{|F^I|} P\left(F_k^I \middle| S\right) \tag{3.4}$$

   The probability $P\left(F_k^I \middle| S\right)$ in this equation is estimated differently depending on the feature type. A smoothed maximum likelihood estimator is used for features with values represented as a single symbol, such as head predicate, number of arguments, lexical arguments, cases and syntactic pattern:

$$P\left(F_k^I \middle| S\right) = \frac{\left|\left\{F_k^I \middle| F_k^I \in F_k^S\right\}\right| + \lambda}{\left|F_k^S\right| + \lambda |F_k|} \tag{3.5}$$

where $\left|\left\{F_k^I \middle| F_k^I \in F_k^S\right\}\right|$ shows how many times F_k^I occurs in F_k^S, and the smoothing parameter λ determines the default probability of F_k^I in S when $\left|\left\{F_k^I \middle| F_k^I \in F_k^S\right\}\right| = 0$. The lower bound of λ (when a new ASC is created for each encountered instance) can be computed based on how many values each feature F_k in the data set can take. More specifically, $\lambda_{min} = \prod_k \frac{1}{F_k}$. For the data sets in the present study (when they are used jointly), $\lambda_{min} = 10^{-17}$. We chose a moderate value 10^{-9}.

Equation (3.5) cannot be used for features with set values, because there is rarely a full overlap between any two sets of properties (e.g., semantic properties). In other words, $\left|\left\{F_k^I \middle| F_k^I \in F_k^S\right\}\right|$ is almost always 0. Alishahi and Pyykkönen (2011) propose the following way to compute the probability for such features, which we employ in this study:

$$P\left(F_k^I \middle| S\right) = \left(\prod_{e \in F_k^I} P(e|S) \times \prod_{e \in F_k \setminus F_k^I} P(\neg e|S)\right)^{\frac{1}{|F_k|}} \tag{3.6}$$

where F_k denotes the set of all values of this feature in the data, and $F_k \setminus F_k^I$ subtracts from this set all elements occurring in F_k^I. The probabilities $P(e|S)$ and $P(\neg e|S)$ can be computed using equation (3.5), replacing F_k^I with an individual element e.

Based on the computed values of the prior and the conditional probability, the model either places I into an existing ASC or creates a new ASC containing only one instance I. Note that when the model receives instances from two languages during a simulation, L1 and L2 instances are not explicitly marked as such. The only relevant information is implicitly present in the values of such features as head predicate, arguments, and syntactic pattern (in case it has prepositions). This ensures the model treats all instances equally, irrespective of their language.

3.3.3 Input data and learning scenarios

Following the original experiments, we simulate L1 English (as in N. C. Ellis et al., 2014b) and L2 English learning (as in N. C. Ellis et al., 2014a). Although the latter study was carried out with native speakers of German, Spanish, and Czech, we only use L1 German due to poor data availability. Manual annotation of argument structures proved to be rather time-consuming, therefore we used available annotated resources for English and German to automatically extract the data we needed.

We use the data sets described in chapter 2; here we briefly outline how they were obtained.

1. The Penn Treebank for English (WSJ part, Marcus et al., 1994) and the TIGER corpus for German (Brants et al., 2004) were used to obtain syntactically annotated simple sentences.

2. Argument structures were extracted from these sentences, using the annotations in English PropBank (Palmer et al., 2005) and the German SALSA corpus (Burchardt et al., 2006).

3. We further used only the sentences containing FrameNet-style annotations (Ruppenhofer et al., 2006), either via the PropBank–FrameNet mappings in SemLink for English (Palmer, 2009), or in the SALSA corpus for German.

4. Word semantic properties were obtained from WordNet (G. A. Miller, 1995) and VerbNet (Kipper Schuler, 2006).

5. Symbolic thematic roles were semi-automatically replaced by sets of elements through the WordNet–FrameNet mappings (Bryl et al., 2012).

The resulting German and English data sets contain 3,370 and 3,624 ASC instances, respectively, which are distributed across 301 (German) and 319 (English) verb types. The corpora mentioned above were the only large sources of English and German data for which the annotations of argument structure were available. We acknowledge that the kind of language in these corpora (mostly newspaper texts) differs from what L1 and L2 learners are normally exposed to. Moreover, the distributions of verbs and constructions in the corpora may be genre- or domain-specific and differ from English and German in general, and the data sets are limited in size: many constructions occur with only a few verb types (we look at this in more detail below, see section 3.4.1). This prevents us from making statements about specific English verbs or constructions, yet the extracted data sets do suit our goal of studying the impact of individual input-related factors on the production of verbs in constructions.

Input to the computational model is sampled randomly from the distribution of instances in the presented data sets. This way, the exact input to the model varies between simulations, to simulate a population of learners with individual linguistic experiences. In the L1 learning setup, 100 simulated learners receive a cumulative number $N = 6,000$ English instances. Clearly, human adult speakers are exposed to much more input than $6,000$ utterances, but given the size of our data sets, this value is large enough: the model achieved a stable level of ASC knowledge on the target input data set after receiving $6,000$ instances. In the L2 setup, 100 learners are exposed to $N = 12,000$ instances: 6,000 L1 German instances, followed by 6,000 instances of "bilingual" input, in which English and German are mixed in equal proportions. This way, L2 learners only encounter $\frac{1}{2} \times 6,000 = 3,000$ English instances, to simulate non-native speakers whose L2 proficiency is lower than L1 proficiency.

3.3.4 Test data and elicited production

Learning was followed by the elicited production task. The model was provided with a number of test items, each of which was intended to elicit the production of verbs in a single construction. Following the original experiments, we looked at the representation of verbs within form-based constructions, without the semantic component: just as EOR's participants, the model is free in its interpretation of the arguments' thematic

roles. We further refer to these units as "constructions", to distinguish them from the emergent ASC representations in the computational model. We did not limit our analysis to prepositional constructions with only two arguments (as did EOR), because this would substantially reduce the amount of the available data in our case. Instead, we used all the available constructions. In terms of ASC representations used by the model, each construction was defined as a syntactic pattern, e.g. ARG1 VERB *about* ARG2 (for a full list of patterns, see Table 3.4 below). To follow the design of the original experiments, we constructed the test stimuli as follows. Following EOR's approach, two stimuli were generated for each construction: the first one had either a pronoun *he* or a pronoun *she* (randomly selected) as the first argument head, and the second one had a pronoun *it* as the first argument head. This way, each stimulus occurred once with an animate (*s/he*) and once with an inanimate pronoun (*it*). The other argument heads were masked, together with the verb. Therefore, during the testing the model was provided with a number of test ASC instances I_{test}, which only contained the values of a few features: number of arguments, syntactic pattern, the first argument (the selected pronoun) and its semantics (e.g., {REFERENCE, PERSON …, ENTITY} for *he*). As a result, test stimuli were similar to those used in the original experiments (in this case, *he ___ about the...*). Given a test instance, the model's task was to produce a list of verbs fitting the empty slot. Such elicited production is implemented as a generation of a set of verbs enumerated with their respective probabilities of production ($V_{produced}$). There is no upper boundary for the number of verbs produced, but verbs with low probabilities of production are excluded from the analysis. The probability of each $V_j \in V_{produced}$ given a test instance I_{test} is calculated as follows:

$$P(V_j|I_{test}) = \sum_S P(V_j|S)P(S|I_{test}) \tag{3.7}$$

The right side of equation (3.7) is a sum of the products of two probabilities, computed for each acquired ASC. $P(V_j|S)$ is estimated as provided in equation (3.5), and $P(S|I_{test})$ is transformed and computed in exactly the same way as during the learning (see equations 3.2–3.4). In other words, to select verbs to fill in a test stimulus, the model first computes how similar the stimulus is to each ASC, and assigns the similarity weights to ASCs. Next, the model considers each verb associated with an ASC, and takes into account both the frequency of the verb in this ASC and the similarity weight of the ASC, to obtain the evidence from this ASC in favor of selecting particular verbs. Finally, such evidence values from all the existing ASCs add up, determining the final selection probability of each verb.

Note that our model is not equipped with explicit language control mechanisms, which human speakers can use for inhibiting activated representations from a non-target language (Kroll, Bobb, Misra, & Guo, 2008; Green, 1998). Therefore, the model may produce L1 verbs in the L2 elicited production task, which is taken into account in our analysis of production data.

3.3.5 Predictor variables

The predictor variables proposed in the original experiments are the joint verb–construction frequency $F(v,c)$, the ΔP-contingency $\Delta P_A(v,c)$, and the prototypicality of verb meaning $Prt(v,c)$. These measures are used for predicting the selection of verbs within each construction. Therefore, the measures are obtained based on the input data which the input to the model is sampled from. Two different methods are used for computing the values.

Our first goal is to simulate the original experiments of EOR closely following their analysis, therefore we adopt their approach of calculating the values of $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ from the whole English data set, without accounting for the individual variation in the input. The value of joint frequency $F(v,c)$ is extracted from the input data set directly, together with additional measures such as the marginal verb frequency $F(v)$, and the marginal construction frequency $F(c)$: these were needed for computing the value of contingency $\Delta P_A(v,c)$:

$$\Delta P_A(v,c) = P(v|c) - P(v|\neg c) = \frac{F(v,c)}{F(c)} - \frac{F(v) - F(v,c)}{N - F(c)}, \tag{3.8}$$

where N denotes the total size of the input data, in this case 3,624 instances. In simple terms, ΔP-contingency is the probability of a verb given a construction minus the probability of the verb's occurrence in all the other constructions. ΔP can take values as high as 1 (when the verb mostly occurs with the target construction) and as low as -1 (when the verb is proportionally much more frequent in other constructions).

As for prototypicality, recall that each verb meaning in ASC instances is represented as a set of elements (e.g., {EXCHANGE, TRANSFER, POSSESSION, CAUSE}), and we consider a verb v to have a higher prototypicality in a construction c when its meaning M_v shares more elements with the meanings M_i of all the other verbs i (excluding v) occurring in c ($i \in c \setminus v$):

$$Prt(v,c) = \frac{\sum_{i \in c \setminus v} \frac{|M_i \cap M_v|}{|M_v|}}{|c \setminus v|}, \tag{3.9}$$

where $|c \setminus v|$ is the number of verb types participating in c, excluding v. We did not use EOR's betweenness centrality values, because they were based on a so-called path similarity between verbs in WordNet, but the hierarchy of verbs in WordNet did not reflect the true hierarchy of verb meanings in our data sets.[5] At the same time, $Prt(v,c)$, as defined here, operates on the actual sets used in ASC instances, and suits our setup. The two measures, however, are conceptually similar: more general verbs with fewer semantic components (*give*: {POSSESSION, TRANSFER, CAUSE}) tend to score higher

[5] As an alternative, we tried to calculate the similarity between verb meanings using the actual sets of semantic elements used in our data sets, build a resulting network based on these similarity values for each construction, and then calculate betweenness centrality on this network. Recall, however, that many constructions in our data sets occurred with only a few verb types: computing betweenness centrality on such a small network yielded an abundant number of 0s, which was damaging for our analysis.

than more specific ones (*purchase*: {BUY, GET, POSSESSION, TRANSFER, CAUSE, COST}).

Our second goal is to address the methodological issues, in particular individual variation, therefore in the respective analysis the values of the three measures are calculated for each simulated learner individually, based on the actual input sample it receives. To do this, during each simulation we record the information about the occurrence of individual verb usages in the actual input: $F(v)$, $F(c)$, and $F(v,c)$. Thus, the value of joint frequency $F(v,c)$ is directly available from the recorded information, and the values of contingency $\Delta P_A(v,c)$ and prototypicality $Prt(v,c)$ are calculated as given above (equations 3.8–3.9), but based on a particular input sample instead of the whole data. N in this case is equal to the actual amount of input: $6,000$ for L1 or $12,000$ for L2 simulations.

The goal of our final study is to identify the best set of variables predicting verb selection. In particular, when presenting the three types of contingency measures, we have mentioned that we plan to test one measure of each type. A raw frequency measure $F(v,c)$ is available directly, and a complex measure $\Delta P_A(v,c)$ is calculated according to equation (3.8). Therefore, we only need a measure of the second type, a conditional probability. We use $Attraction(v,c)$, henceforth $A(v,c)$, which normalizes the joint verb–construction frequency by the marginal construction frequency:

$$A(v,c) = P(v|c) = \frac{F(v,c)}{F(c)} \tag{3.10}$$

The next section describes our simulations and the obtained results. First, we simulate the original experiment for L1 (N. C. Ellis et al., 2014b, experiment 2) and for L2 (N. C. Ellis et al., 2014a), keeping our setup and analysis as close as possible to the original experiments, to see whether our model produces results similar to those of the original experiments. Next, we address the two methodological issues by reanalyzing the data obtained from the same simulated learners, to examine whether the original results still hold in the new analysis. Finally, we use a number of regression models which include different combinations of predictions, to determine which factors predict the production data best.

3.4 Simulations and results

3.4.1 Simulating the original experiments

In this section we employ the elicited production task described in section 3.3.4 above to obtain a list of produced verbs. Using this list, we look at the verbs produced within some individual constructions, run correlation tests for individual constructions, and perform a combined analysis on the whole data set as described next.

Methodological details

Each simulated learner has produced a list of verbs fitting every given construction. EOR in their experiments limited the number of produced verbs by allocating a minute

for each stimulus. To adopt a similar approach, we had to filter out verbs whose probability of production was lower than a certain threshold. The value of .005 was established empirically, by testing values between .05 and .001. Using this threshold value, for each verb in a certain construction we calculate the total production frequency of this verb by all learners, henceforth $PF(v,c)$. If a verb has not been produced by any learner in a certain construction, the verb–construction pair is excluded from the analysis, to obtain data similar to EOR's. For analyzing L2 production data, we exclude all L1 verbs produced by the model, because these are irrelevant for our analysis.

First we look at the verbs produced within a sample of ten individual constructions: four most frequent constructions in our data set, and six constructions present in both EOR's and our data set.

Next, to compare our model to EOR's human subjects, we look at whether each of the three factors – $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ – correlates with $PF(v,c)$ within each construction in our data set, using Pearson correlation coefficient.[6]

Finally, we proceed with a combined regression analysis on the whole data set. Again, to make the results comparable with EOR's findings, we first consider only the six constructions present in both their and our data set. However, this is a rather small sample, therefore we run an additional regression analysis on our whole data set of 44 constructions. Before fitting the models, we standardize all the variables, to make the β coefficients directly comparable and to reduce the collinearity of predictors. We run multiple regression analyzes to predict $PF(v,c)$ by the three factors: $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$. Note that the values of the mentioned variables in this simulation set are computed using the first method described in section 3.3.5 – that is, for the whole input data set, following the original experiments.

L1 simulations

First we look at the verbs produced by the model within ten individual constructions selected as described above: the produced lists are provided in Table 3.3. We can see substantial differences between the frequencies of occurrence of individual constructions in the input data. Some of them are rather frequent: e.g., A1 V A2 occurs 2,508 times with 224 verb types, and A1 V occurs 724 times with 119 verb types. In contrast, most prepositional constructions are infrequent: in particular, the six constructions from EOR's data set occur only 1 to 11 times with 1 to 6 verb types. Respectively, the number of verb types generated by the model per construction also varies between 2.4 and 84.2 in this subset of ten constructions. It is also clear from the table (see bold font) that the model sometimes produces verbs which are unattested in the target construction in the input. We discuss this in the interim discussion below.

To see whether the frequencies of verb production correlate with each of the three target factors, as in EOR's study, we run a series of correlation tests reported in

[6] As in the original study, we add 0.01 to all the predictors as well as to the outcome variable. We additionally increment $\Delta P_A(v,c)$ by 1, to avoid having negative values in the data. The last step is necessary, because we log-transform all the variables as in EOR's studies. The log-transformation is justified by the fact that practice (which in our case is reflected in production frequency) is believed to be a power function of experience (Newell & Rosenbloom, 1981), and therefore a power transformation can linearize the relationship between $PF(v,c)$ and at least one of the predictors, namely $F(v,c)$.

Table 3.4. We can see that both the joint frequency $F(v,c)$ and ΔP-contingency are correlated with the production frequency $PF(v,c)$ for almost all constructions: verbs which appear more frequently in a construction or which are associated more strongly with a construction are also produced more frequently by the model. This is not always the case for the third predictor, prototypicality $Prt(v,c)$: significant correlations of this variable with production frequency are only observed for 23 out of 44 constructions. In particular, there is no such correlation for any of the six constructions present in EOR's data (marked with an asterisk in Table 3.4). We address this issue below in the interim discussion. The next step, as we mentioned above, is to provide combined regression analyses of the data set.

The summary of the three models is provided in Table 3.5(a–b). Overall, the results are similar to what EOR report: all the three variables contribute to predicting the verb production frequency. However, the difference is that $Prt(v,c)$ in our experiment appears to be a less important predictor, which is reflected in the β values (from 0.05 to 0.06 in our study, depending on the set of constructions, vs. 0.29 in the original study). We have run an additional analysis, in which we kept the verbs that appeared in a construction in the input, but were not produced in this construction by the model: $PF(v,c)$ for such verbs was assigned to 0. Besides, we have run mixed-effects models (e.g., Baayen, 2008), as implemented in *R* (D. Bates, Mächler, Bolker, & Walker, 2015), for the same two sets of constructions, with a random intercept and random slopes for all the three factors over individual constructions. The results appeared to be very similar to what is reported here, therefore we leave them out for brevity.

Table 3.3: Ten constructions with their frequencies and produced verbs. Verbs in bold are unattested with target construction in input.

	Construction				
Property	A1 V A2	A1 V	A1 V A2 A3	A1 V A2 *to* A3	A1 V *about* A2
Verb tokens in input	2,508	724	112	52	11
Verb types in input	224	119	8	12	4
Verb types produced	228	115	66	47	146
Avg. verb types produced	84.2	35.5	9.7	11.6	10.6
Verb types with	*want*: 185	*want*: 169	*give*: 143	*send*: 139	*complain*: 175
their production	*buy*: 184	*begin*: 135	*send*: 117	*give*: 137	*inquire*: 154
frequencies	*sell*: 182	*die*: 108	*pull*: 90	*elect*: 99	*brag*: 131
	announce: 170	*exist*: 104	*tell*: 58	*propose*: 87	*shout*: 96
	receive: 169	*happen*: 103	*place*: 37	***disclose***: 77	***listen***: 41
	hold: 167	*expire*: 102	*disclose*: 36	*donate*: 71	***sit***: 19
	see: 162	*rise*: 99	*drag*: 33	*pass*: 70	***groan***: 17
	start: 159	*sell*: 96	***elect***: 32	*pressure*: 51	***scoff***: 14
	post: 154	*decline*: 90	*hang*: 31	*explain*: 41	***live***: 11
	lead: 153	*drop*: 89	***pressure***: 24	*peg*: 39	***send***: 11
	...	...	...	...	...
	unnerve: 1	*exhale*: 1	***wear***: 1	***want***: 1	***withdraw***: 1

Table 3.3 (continued from previous page).

Property	Construction				
	A1 V *into* A2	A1 V *with* a2	A1 V *for* A2	A1 V *against* A2	A1 V *of* A2
Verb tokens in input	9	7	3	1	1
Verb types in input	6	5	2	1	1
Verb types produced	206	106	77	20	21
Avg. verb types produced	24.0	10.4	6.0	2.6	2.4
Verb types with	*buy*: 107	*join*: 174	*search*: 164	*lean*: 174	*disapprove*: 154
their production	*run*: 88	*cooperate*: 141	*scream*: 135	***groan***: 17	***scoff***: 14
frequencies	***sell***: 78	*merge*: 138	***sit***: 43	***scoff***: 16	***sit***: 14
	eat: 69	*respond*: 134	***scoff***: 20	***sit***: 13	***groan***: 11
	erupt: 68	*sit*: 118	***obtain***: 19	***gaze***: 7	***gaze***: 7
	pack: 64	***scoff***: 23	***glance***: 17	***live***: 6	***live***: 7
	turn: 63	***glance***: 21	***groan***: 17	***rely***: 5	***squint***: 7
	acquire: 62	***groan***: 19	***gaze***: 8	***listen***: 4	***rely***: 5
	hold: 51	***scream***: 18	***live***: 8	***squint***: 4	***glance***: 4
	want: 50	***gaze***: 16	***rely***: 6	***glance***: 3	***listen***: 4
	...	...	...	...	...
	thrill: 1	***write***: 1	***steal***: 1	***shout***: 1	***spout***: 1

Table 3.4: Summary of correlation tests between $PF(v,c)$ and each of the three factors for individual constructions in L1 replication data.

	$F(v,c)$		$\Delta P_A(v,c)$		$Prt(v,c)$	
Construction	r	p	r	p	r	p
A1 V	.96	<.001	.17	.002	.05	.372
A1 V A2	.94	<.001	.13	.020	.08	.162
A1 V A2 A3	.44	<.001	.22	<.001	.11	.044
A1 V A2 *about* A3	.18	.001	.18	.001	.21	<.001
A1 V A2 *above* A3	.21	<.001	.21	<.001	.14	.011
A1 V A2 *across* A3	.33	<.001	.33	<.001	.22	<.001
A1 V A2 *among* A3	.19	.001	.19	.001	.03	.622
A1 V A2 *as* A3	.43	<.001	.42	<.001	.13	.020
A1 V A2 *at* A3	.43	<.001	.28	<.001	.17	.003
A1 V A2 *by* A3	.28	<.001	.28	<.001	.13	.023
A1 V A2 *for* A3	.43	<.001	.43	<.001	.12	.029
A1 V A2 *from* A3	.36	<.001	.31	<.001	.18	.001
A1 V A2 *in* A3	.35	<.001	.34	<.001	.21	<.001
A1 V A2 *into* A3	.30	<.001	.30	<.001	.09	.125
A1 V A2 *of* A3	.22	<.001	.21	<.001	.12	.034
A1 V A2 *on* A3	.46	<.001	.33	<.001	.15	.006
A1 V A2 *over* A3	.42	<.001	.42	<.001	.15	.008
A1 V A2 *through* A3	.27	<.001	.27	<.001	.20	<.001
A1 V A2 *to* A3	.61	<.001	.39	<.001	.19	.001
A1 V A2 *under* A3	.16	.003	.16	.003	.21	<.001
A1 V A2 *until* A3	.32	<.001	.32	<.001	.14	.011
A1 V A2 *with* A3	.49	<.001	.46	<.001	.10	.062
A1 V *about* A2*	.27	<.001	.22	<.001	.02	.663
A1 V *against* A2*	.36	<.001	.36	<.001	−.01	.875
A1 V *at* A2	.31	<.001	.29	<.001	.02	.692
A1 V *below* A2	.13	.020	.13	.020	.13	.021
A1 V *by* A2	.29	<.001	.23	<.001	.02	.779
A1 V *for* A2*	.65	<.001	.63	<.001	−.12	.294
A1 V *from* A2	.13	.022	.13	.022	−.09	.095
A1 V *from* A2 A3	.56	<.001	.56	<.001	.10	.086
A1 V *in* A2	.25	<.001	.17	.002	−.04	.468
A1 V *into* A2*	.21	<.001	.17	.002	−.07	.220
A1 V *of* A2*	.35	<.001	.35	<.001	.08	.133
A1 V *on* A2	.40	<.001	.31	<.001	.12	.037
A1 V *on* A2 A3	.23	<.001	.23	<.001	.20	<.001
A1 V *to* A2	.15	.009	.13	.020	.01	.828
A1 V *to* A2 A3	.23	<.001	.23	<.001	.16	.003
A1 V *to* A2 *about* A3	.48	<.001	.48	<.001	.09	.101
A1 V *to* A2 *of* A3	.49	<.001	.49	<.001	.09	.094
A1 V *up* A2	.09	.107	.09	.107	.19	.001
A1 V *upon* A2	.23	<.001	.23	<.001	.06	.255
A1 V *with* A2*	.36	<.001	.32	<.001	−.06	.285
A1 V *with* A2 *in* A3	.26	<.001	.26	<.001	.07	.216
A1 V *with* A2 *on* A3	.44	<.001	.44	<.001	.17	.002

* Constructions present in EOR's data.

Table 3.5: Summary of the multiple regression models fitted to the L1 replication data.

a. L1 simulations: constructions present in EOR's data set

$PF \sim F + \Delta P + Prt$

Variable	β	SE	p	LMG[a]	VIF
$F(v,c)$	0.69	0.03	$< .001$	.59	2.75
$\Delta P_A(v,c)$	0.25	0.03	$< .001$	.40	2.74
$Prt(v,c)$	0.05	0.02	.008	.01	1.02
Multiple $R^2 = .83$, adjusted $R^2 = .82$					

b. L1 simulations: all constructions

$PF \sim F + \Delta P + Prt$

Variable	β	SE	p	LMG	VIF
$F(v,c)$	0.57	0.01	$< .001$	.73	1.13
$\Delta P_A(v,c)$	0.25	0.01	$< .001$	.25	1.14
$Prt(v,c)$	0.06	0.01	$< .001$	.02	1.02
Multiple $R^2 = .50$, adjusted $R^2 = .50$					

c. L2 simulations: constructions present in EOR's data set

$PF \sim F + \Delta P + Prt$

Variable	β	SE	p	LMG	VIF
$F(v,c)$	0.70	0.02	$< .001$	.57	2.73
$\Delta P_A(v,c)$	0.29	0.02	$< .001$	.41	2.73
$Prt(v,c)$	0.05	0.01	.002	.02	1.02
Multiple $R^2 = .90$, adjusted $R^2 = .90$					

d. L2 simulations: all constructions

$PF \sim F + \Delta P + Prt$

Variable	β	SE	p	LMG	VIF
$F(v,c)$	0.59	0.01	$< .001$	.75	1.12
$\Delta P_A(v,c)$	0.24	0.01	$< .001$	.23	1.14
$Prt(v,c)$	0.06	0.01	$< .001$	.02	1.03
Multiple $R^2 = .51$, adjusted $R^2 = .51$					

[a] This measure is used in EOR's studies: it computes the importance of each predictor relative to the other predictors by analyzing how the regression coefficients change when various combinations of predictors are excluded from the model. The measure was proposed by Lindeman, Merenda, and Gold (1980) and implemented in *R* by Grömping (2006).

L2 simulations

For the sake of space we omit the lists of verbs produced in the L2 simulations, as well as the correlational results per construction. There were some differences between the actual sets of verbs produced in L1 and L2 simulations, but these would not be immediately obvious from verb lists or correlation tables. Although comparing L1 to L2 simulations was not our goal in this study, to further demonstrate that our model performed as expected on the simulated task, we quantified the differences between the verbs produced in L1 and L2 simulations, to compare these differences to what Römer, O'Donnell, and Ellis (2014) report. We adopted an approach similar to theirs and ran a mixed-effects regression analysis predicting the frequencies of verbs produced in L2 simulations from those in L1 simulations, with the random slope over individual constructions. The model fit was reasonable (marginal $R^2 = .57$, conditional $R^2 = .65$)[7], and the β-coefficient reflecting the correlation between the produced verb frequencies in L1 and L2 simulations was equal to 0.71, which is rather close to the average value of 0.75 reported by Römer et al. (2014) for native English vs. native German speakers.

Next, we proceed with reporting on the combined regression analysis of the L2 simulation data set. Table 3.5(c–d) summarizes the regression results for the simulated L2 production data. Overall, the results are similar to those for L1, and to those of EOR. Note that the values of the three target variables, following EOR's study, were computed for English constructions only. For the same reason, although the model produced some German verbs in the test task, these verbs were excluded from our analysis. However, the input to the model consisted of both English and German constructions, many of which are shared by the two languages. Since our model treated L1 German and L2 English instances in exactly the same way, it could be fairer to compute the values of $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ for the whole data set, assuming that each construction may be associated with both English and German verbs. This is why we ran an additional analysis, in which all the produced German verbs were kept during the analysis, and the values of the three variables were computed for the whole bilingual data set. Again, the results were very similar to the ones reported above.

Interim discussion

To summarize, the model performs as expected on the target task: verbs which appear in a construction in the input tend to populate the top of the respective list of produced verbs for this construction. Since there are six constructions present both in this study and in EOR's study, we would ideally compare the verbs produced by the model and by human participants. Yet, in our input data set these constructions occur with only 1 to 6 verb types, and the model tends to produce these verbs first. In contrast, naturalistic language input to human participants is more varied: each construction occurs with a greater variety of verb types, and EOR's participants are not as limited in their verb choice as the model is. Besides, the distribution in the input per construction differs across the two studies: human participants are mostly exposed to colloquial language,

[7] These coefficients indicate the amount of variance explained by the fixed factors and by the full model, respectively (Johnson, 2014), and are computed using an existing R implementation (Bartoń, 2016).

while our input data set is based on business newspaper texts from the Penn Treebank (WSJ part). This is reflected in verb selection: human participants tend to produce colloquial verbs (e.g., *go*, *be*, *dance with ...*), while the model often prefers specialized verbs (*join*, *cooperate*, *merge with ...*), although in both cases verbs produced first tend to be the most frequent ones in the respective input data set.

Given the low number of verb types in some prepositional constructions, the model generalizes and produces verbs unattested in these constructions, marked with bold in Table 3.3. These verbs mostly appear at the bottom of the list for each construction, with a few exceptions, such as A1 *elect* A2 A3, A1 *disclose* A2 *to* A3, and A1 *sell into* A2. Although these usages may not be the most common ones, they are not ungrammatical either, and could easily appear in a larger language sample: e.g., *they elected him president*; *he ... discloses it to others*; *rivals ... sell into that market* (examples taken from the BNC). This suggests that our model is able to find reasonable generalizations using the input. At the same time, some occasionally produced verbs are ungrammatical, such as A1 *send about* A2, A1 *listen of* A2, etc. This happens because the model's exposure to the target construction is limited in terms of participating verb types, and there may not be enough support for making correct generalizations. Besides, as we argue below in this section, verb semantic representations in the input data are not rich enough. This is why the model overgeneralizes and produces such ungrammatical usages. However, as we mentioned, the ungrammatical usages tend to appear at the bottom of the list, and do not compromise the model's performance on the verb production task. Besides, the difference between the frequencies of verb production in L1 and L2 simulations is very close to the value reported by Römer et al. (2014), which further defends the performance of our model on this task. Nevertheless, the fact that we could not compare the model's performance to human data in terms of specific verbs leaves the possibility that the model does not perform exactly like humans in the target task.

As for the correlations and the combined regression analysis, the frequency of production of verbs in our simulations can be predicted by joint verb–construction frequency, ΔP-contingency, and to some extent by verb semantic prototypicality. However, prototypicality does not correlate with the production frequency in all constructions, and its contribution to predicting production frequency is smaller than in EOR's studies. We propose three possible explanations of this result.

The first explanation is that our computational model does not rely on this factor to the extent human speakers do when generating verbs in constructions. This, indeed, may be the case, because the predicate semantics is only one out of many features in our representation of verb usages (recall Table 3.2). In other words, our model may underestimate the importance of the verb meaning in learning argument structure constructions. Note, however, that EOR in one of their studies (N. C. Ellis et al., 2014a) also did not observe significant correlations between the production frequency and semantic prototypicality for 5 out of 17 constructions in the data obtained from L1 English as well as L1 German speakers. In our simulations prototypicality was correlated with the production frequency in 23 out of 44 constructions, and it had an independent contribution in all the regression models reported above.

The second explanation relates to the type of semantic representations that the

model operates on. Human speakers are often believed to possess fine-grained semantic representations of verbs: for example, Pinker (2013) proposes such narrow semantic rules as "transfer of possession mediated by separation in time and space" (p. 129). In contrast, semantic representations in our data set are extracted from WordNet and VerbNet and are more simplistic than that (e.g., give: {POSSESSION, TRANSFER, CAUSE}). This is not critical for the simulated learning process, because the discrimination between different verbs is supported by other features in the data, such as arguments' thematic proto-roles. However, in our analysis the prototypicality values are computed based on the verb semantics only, and the impoverished semantic representations may lead to the lower impact of semantic prototypicality in our study.

Our final explanation relates to how the prototypicality measure operates on a large and dense (as in EOR's study) vs. a small and sparse data set (as in our study). EOR computed semantic prototypicality of a verb in a construction based on a rich semantic network of all verbs that appear in this construction in the BNC. BNC is a rather large source, and it is unlikely that EOR's participants, given a construction, would produce a verb which is unattested in this construction in the BNC. In contrast, some constructions in our data set appeared with only a few verb types, in which case the prototypicality values were computed based on a rather small set of these few verbs. Yet the model often produced verbs which were unattested in this construction (non-members), but were semantically similar to other verbs that did appear in the target construction (members). To give an example, a construction ARG1 VERB ARG2 *for* ARG3 appeared in our data set with only five verbs: *substitute*, *elect*, *hail*, *criticize*, and *remove*. In the production task, the model generated these five verbs rather frequently, but there were other frequent verbs, in particular *praise*, *chastise*, and *indict*. Clearly, these verbs are allowed in the target construction, partly because they are somewhat synonymic to the construction members, at least when used in the target context (*to* VERB *someone for a reason*): *chastise* and *indict* are similar to *criticize*, while *praise* is similar to *hail*. In fact, the non-members must have been included into the target set of verbs, and the semantic prototypicality of all the verbs must have been calculated on this extended set. Since we had no way to predict beforehand which verbs would be produced by the model (and thus, should be included into the set), we computed all prototypicality values on the smaller set of verbs. This was particularly the case for the six constructions shared between our data set and EOR's data set: recall that these constructions appeared in the input with only a few verb types. As a result, prototypicality values for such constructions might not be very objective, hence the rather low contribution of this variable to predicting the frequency of verb production. At the same time, the correlation between prototypicality and production frequency is also very small for some frequent constructions, such as ARG1 VERB ARG2 and ARG1 VERB (at the top of Table 3.4), which cannot be explained by the account outlined above. We believe this has to do with the incoherence of semantic networks for such constructions, and we leave this issue for the final discussion.

The small effect of semantic prototypicality in data simulated by our model should be addressed in the future; for now it is important to keep in mind that the reported impact of semantic prototypicality in the current study may be underestimated. Apart from the described limitation, our model was able to replicate the main effects reported

in the original studies, both for L1 and L2. In the next section we address the two methodological issues of the original study discussed earlier (see section 3.2.1 above).

3.4.2 Addressing the methodological issues: Individual variation

In this second analysis we take into account the individual variation in the linguistic input, while trying to keep the rest of the design as close as possible to the previous analysis. We use the same set of simulated learners described in the previous section to predict verb production by the three target variables. The differences from the previous analysis are described next.

Methodological details

This time we do not calculate the cumulative frequency of production of each verb in a specific construction, $PF(v,c)$, as we did earlier. Instead, for each verb produced by each simulated learner we define a binary outcome variable, which is set to 1 if the probability of production of this verb equals at least .005 (the threshold value from the previous analysis), and to 0 otherwise. This way, we now do not combine the data from all learners into a single $PF(v,c)$ value, but instead have data from individual simulated learners, while keeping the rest of the design very close to what was reported in the previous section. Besides, we compute the values of the three target variables – $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ – for each simulation individually, based on a specific input sample. To keep up with the previous analysis, we apply the same data transformations as described before. To account for potential individual variation between constructions and learners, we use logistic mixed-effects models with the binary outcome variable described above, with $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ as fixed factors, and with constructions and learners as random factors. All the mixed-effects models for both L1 and L2 simulated data were fit to the two data sets: EOR's constructions only, and the whole data set, just as in the previous section. We started from maximal random effect structure with the random intercept and three random slopes (for each predictor), however the maximal model only converged for EOR's subset of L2 simulated data, therefore we removed some random slopes.

Results

The results are provided in Table 3.6. We did not use the *LMG* relative importance measure from the previous analysis, because it could not be applied to mixed-effects models. In this set of models the β-coefficients for $\Delta P_A(v,c)$ are generally small (0.02 to 0.08), with the exception of the model fitted to EOR's constructions in L1 simulations ($\Delta P_A(v,c) = 0.26$). However, even in the latter case the respective *SE* value is rather high (0.18), suggesting high variation in the data regarding the effect of $\Delta P_A(v,c)$. Besides, there is substantial variability among the coefficients for $Prt(v,c)$: between -0.08 and 0.38. The coefficients are greater in the models fitted to all constructions (0.38 and 0.37), compared to the models fitted to EOR's constructions only (0.10 and -0.08). Note that, surprisingly, in the latter case this coefficient has a negative value,

Table 3.6: Summary of the mixed-effects models accounting for individual language experience.

a. L1 simulations: constructions present in EOR's data set

$Prod. \sim F + \Delta P + Prt + (1 + \Delta P|learner) + (1 + \Delta P|constr.)$

Variable	β	SE[a]	$95\%CI$[a]	VIF
$F(v,c)$	0.58	0.01	$[0.56, 0.60]$	1.03
$\Delta P_A(v,c)$	0.26	0.18	$[-0.09, 0.61]$	1.00
$Prt(v,c)$	0.10	0.02	$[0.06, 0.13]$	1.03

b. L1 simulations: all constructions

$Prod. \sim F + \Delta P + Prt + (1|learner) + (1|constr.)$

Variable	β	SE	$95\%CI$	VIF
$F(v,c)$	0.89	0.00	$[0.88, 0.90]$	1.95
$\Delta P_A(v,c)$	0.02	0.00	$[0.01, 0.02]$	2.01
$Prt(v,c)$	0.38	0.01	$[0.36, 0.39]$	1.07

c. L2 simulations: constructions present in EOR's data set

$Prod. \sim F + \Delta P + Prt + (1 + F + \Delta P + Prt|learner) + (1 + F + \Delta P + Prt|constr.)$

Variable	β	SE	$95\%CI$	VIF
$F(v,c)$	0.75	0.06	$[0.62, 0.87]$	1.32
$\Delta P_A(v,c)$	0.08	0.06	$[-0.04, 0.20]$	1.41
$Prt(v,c)$	-0.08	0.09	$[-0.26, 0.09]$	1.09

d. L2 simulations: all constructions

$Prod. \sim F + \Delta P + Prt + (1|learner) + (1|constr.)$

Variable	β	SE	$95\%CI$	VIF
$F(v,c)$	0.89	0.01	$[0.88, 0.90]$	1.42
$\Delta P_A(v,c)$	0.05	0.00	$[0.04, 0.06]$	1.50
$Prt(v,c)$	0.37	0.01	$[0.35, 0.39]$	1.07

[a] Due to the large sizes of the data sets, the reported *SE* and *CI* values for all the models are approximate, based on the Wald tests (D. Bates, Mächler, Bolker, & Walker, 2015).

however the respective variation in the data is high again ($SE = 0.09$). Besides, the respective model (fitted to EOR's constructions in L2 simulations) is the only one which includes random slopes for $Prt(v,c)$ over individual constructions and individual learners (see Table 3.6(c)), suggesting that some of this variation may come from accounting for the individual variation in the data.

Interim discussion

The models reported above predict verb production while taking into account differences in individual linguistic experiences of simulated learners. By comparing this kind of analysis to the original one, we can investigate whether taking into account individual variation may potentially lead to different results. Although our goal was to keep the data and the analysis maximally consistent with the previous setup, there are still differences in the type of outcome variable used (numeric production frequency vs. binary outcome) and, as a result, in the type of models fitted to the data (linear vs. logistic regression). This does not allow us to compare coefficients pairwise across the two types of analysis, however the general pattern of difference suggests that the effect of ΔP-contingency may not be as high as predicted earlier, as soon as individual variation is taken into account.

The results on the individual variation in terms of semantic prototypicality are somewhat inconclusive. On the one hand, the positive effect of semantic prototypicality is present in the new models fitted to the full data sets, in both L1 and L2 simulations, and in the new model fitted to EOR's constructions in L1 simulations. On the other hand, there is not enough evidence for such effect in EOR's constructions obtained from L2 simulations. This must relate to whether the respective prediction model accounts for the variation between individual learners regarding this factor: we fitted an additional model to the same data, this time without the random slope for prototypicality over individual learners, and this model did predict a positive effect of semantic prototypicality. In other words, our data suggest that semantic prototypicality may play a role for some learners, but not for others.

3.4.3 Addressing the methodological issues: Order of preference

In the third set of analyses we look into the order of verb production by the same simulated learners, trying again to keep the rest of the design as close as possible to the original procedure.

Methodological details

In this set of analyses we record the actual probability of production of each verb by each simulated learner in each construction and then compute the cumulative probability, $PP(v,c)$, using it as the outcome variable in regression, instead of cumulative frequency. Cumulative frequency of a verb only shows how many times it is produced overall, while cumulative probability preserves the order of verb production by adding up the actual values of verb production probability for each learner. Unlike in the

previous section, we are not interested in the variation between learners' individual experiences, therefore we use the values of $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ computed for the overall data set, to keep this analysis as close as possible to the original one. Again, we use the threshold value of .005 and apply the same data transformations as before. To account for the variation between constructions, we use linear mixed-effects models with $PP(v,c)$ as the outcome variable, with $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ as fixed factors, and with the random intercept and three random slopes (for each predictor) over individual constructions. One random slope has been removed from one final model to ensure its convergence. The rest of the analysis follows the originally outlined procedure.

Results

The summaries of the prediction models are provided in Table 3.7. Just as in the previous set of analyses, we can see that the effect of ΔP-contingency is small (the greatest β-coefficient is 0.03), and even negative (-0.04) for one of the models. Besides, in all cases the respective 95%CI includes 0, suggesting that the contributions of ΔP-contingency are not significant in these models.

In other respects this new set of models is similar to the original analysis. The other two factors, joint frequency $F(v,c)$ and prototypicality $Prt(v,c)$, have their independent contributions, although in one case the 95%CI for prototypicality includes 0. The overall fit of the models to the data is lower than reported in our first analysis: they explain 34 to 66% of the variance in the data (see R^2_c values in the table), and only 28 to 47% of this is explained by the fixed factors (R^2_m values): to compare, the overall fit of the models in the original analysis varies between 50 and 90%. [8]

Interim discussion

The models reported above predict the cumulative probability of verb production by the simulated learners. Unlike the originally reported models (see section 3.4.1), this type of analysis accounts for the order of verb preference by our simulated L1 and L2 learners. Most importantly, none of the four models suggest that $\Delta P_A(v,c)$ is an independent predictor, when the order of verb production is taken into account. Recall that both joint frequency and ΔP-contingency are measures of the contextual frequency: this may explain why we do not observe the independent effects of both measures at the same time. Indeed, the approximate correlation coefficient between βs for joint frequency and ΔP-contingency (this coefficient is not included into the tables) appears to be rather large, between -0.50 and -0.74. In other words, the higher the β for frequency, the lower the β for ΔP-contingency, and vice versa.

The poorer fits of the models support our idea that there is space for refining the original prediction model used so far in the analyses: another set of variables may explain the data better without predicting so much random variation between constructions. We will investigate this issue in the next section.

[8] For a fairer comparison of model fits across the two types of analysis, we also looked at the mixed-effects models mentioned in section 3.4.1, and their fits were still higher than reported here.

Table 3.7: Summary of the mixed-effects models accounting for the order of verb preference.

a. L1 simulations: constructions present in EOR's data set

$PP \sim F + \Delta P + Prt + (1 + F + \Delta P + Prt | constr.)$

Variable	β	SE[a]	$95\%CI$[a]	VIF
$F(v,c)$	0.56	0.08	$[0.41, 0.72]$	1.74
$\Delta P_A(v,c)$	-0.04	0.09	$[-0.21, 0.13]$	1.70
$Prt(v,c)$	0.06	0.05	$[-0.03, 0.15]$	1.39
$R^2_m = .28$, $R^2_c = .34$[b]				

b. L1 simulations: all constructions

$PP \sim F + \Delta P + Prt + (1 + F + \Delta P + Prt | constr.)$

Variable	β	SE	$95\%CI$	VIF
$F(v,c)$	0.84	0.04	$[0.77, 0.93]$	1.86
$\Delta P_A(v,c)$	0.03	0.02	$[-0.01, 0.07]$	1.96
$Prt(v,c)$	0.14	0.02	$[0.10, 0.18]$	1.09
$R^2_m = .47$, $R^2_c = .64$				

c. L2 simulations: constructions present in EOR's data set

$PP \sim F + \Delta P + Prt + (1 + F + \Delta P | constr.)$

Variable	β	SE	$95\%CI$	VIF
$F(v,c)$	0.53	0.08	$[0.37, 0.68]$	1.64
$\Delta P_A(v,c)$	0.02	0.08	$[-0.13, 0.18]$	1.63
$Prt(v,c)$	0.14	0.04	$[0.06, 0.22]$	1.02
$R^2_m = .32$, $R^2_c = .37$				

d. L2 simulations: all constructions

$PP \sim F + \Delta P + Prt + (1 + F + \Delta P + Prt | constr.)$

Variable	β	SE	$95\%CI$	VIF
$F(v,c)$	0.85	0.05	$[0.75, 0.95]$	2.20
$\Delta P_A(v,c)$	0.03	0.02	$[-0.01, 0.08]$	2.25
$Prt(v,c)$	0.14	0.02	$[0.10, 0.17]$	1.05
$R^2_m = .47$, $R^2_c = .66$				

[a] The reported SE and CI values are estimated via parametric bootstrap with $1,000$ resamples (D. Bates, Mächler, Bolker, & Walker, 2015).

[b] R^2_m and R^2_c stand for marginal and conditional R^2 coefficients.

3.4.4 Refining the prediction model

Our next goal is to test whether there is a better set of predictors explaining the production data. Based on our theoretical overview, we have three issues to address. First, theoretical accounts suggest that the marginal verb frequency may play an independent role in verb selection, therefore we believe that including marginal frequency into the prediction model would improve its fit to the data. Second, the presence of two contextual frequency (association) measures in the model may not be well justified, and eliminating one of them might not necessarily damage the model. Finally, there are multiple measures of contextual frequency, three of which we plan to test: joint frequency, ΔP (as in the previous analyses), and Attraction.

Methodological details

We start by fitting a number of mixed-effects models of the type described in the second analysis (logistic models taking into account individual differences) and in the third analysis (linear models taking into account order of preference). To ensure that the models generalize well over different constructions, we use the full set of constructions for fitting each model, and not EOR's subset. The structure of fixed factors in the models is defined as described below.

		I		**II**		**III**
(m1a) *Production*	$\sim$			$joint\,freq. \times \Delta P$	$\times$	$prototyp.$
(m2a) *Production*	$\sim$			$joint\,freq. \times attr.$	$\times$	$prototyp.$
(m3a) *Production*	$\sim$			$attr. \times \Delta P$	$\times$	$prototyp.$
(m4a) *Production*	$\sim$			$joint\,freq.$	$\times$	$prototyp.$
(m5a) *Production*	$\sim$			$attr.$	$\times$	$prototyp.$
(m6a) *Production*	$\sim$			ΔP	$\times$	$prototyp.$
(m1b) *Production*	$\sim$	$verb\,freq.$	$\times$	$joint\,freq. \times \Delta P$	$\times$	$prototyp.$
(m2b) *Production*	$\sim$	$verb\,freq.$	$\times$	$joint\,freq. \times attr.$	$\times$	$prototyp.$
(m3b) *Production*	$\sim$	$verb\,freq.$	$\times$	$attr. \times \Delta P$	$\times$	$prototyp.$
(m4b) *Production*	$\sim$	$verb\,freq.$	$\times$	$joint\,freq.$	$\times$	$prototyp.$
(m5b) *Production*	$\sim$	$verb\,freq.$	$\times$	$attr.$	$\times$	$prototyp.$
(m6b) *Production*	$\sim$	$verb\,freq.$	$\times$	ΔP	$\times$	$prototyp.$

In all the equations above, component I represents the marginal verb frequency, component II comprises contextual frequency measures, and component III is the semantic prototypicality. We start with the original model tested in the previous sections, m1a. Models m2a–m3a resemble m1a, but they test alternative pairs of the three contextual frequency measures. Models m4a–m6a, in contrast, eliminate one of the contextual frequency measures, keeping only one. Finally, the other six models (m1b–m6b) mirror

Table 3.8: Model rankings.

	L1 data				L2 data			
	Individual differences		Order of preference		Individual differences		Order of preference	
Rank	Model	$\Delta AICc$	Model	$\Delta AICc$	Model	$\Delta AICc$	Model	$\Delta AICc$
1	m2b	0	m2b*	0	m2b	0	m2b*	0
2	m1b	2,601	m1b	20	m1b	1,893	m1b	21
3	m4b	5,842	m4b	39	m4b	5,236	m4b	41
4	m2a	9,867	m3b*	57	m2a	5,810	m3b*	66
5	m1a	13,892	m5b	131	m1a	8,282	m5b	135
6	m4a	16,539	m6b	614	m4a	11,475	m6b	579
7	m3b*	23,403	m2a*	846	m3b*	19,627	m2a*	749
8	m5b	34,996	m1a	855	m5b	29,330	m1a	760
9	m3a*	36,100	m4a	858	m3a*	30,887	m4a	762
10	m5a	55,234	m3a*	921	m5a	43,980	m3a*	833
11	m6b	64,844	m5a	1,023	m6b	53,950	m5a	932
12	m6a	93,828	m6a	1,496	m6a	72,918	m6a	1,363

* Models which showed multicollinearity problems ($VIF > 3$ for some predictors).

models m1a–m6a, respectively, but add the marginal frequency measure to their counterparts. Note that the models are multiplicative due to the log-transformation of all the variables: $log(y) = log(a) + log(b) + log(c) \Rightarrow y = abc$. Studying and interpreting interactions between variables in such models are not straightforward, and for simplicity we do not include any interaction terms in the prediction models.

We compare the fit of all the 12 models using their corrected Akaike information criterion (AICc), as implemented in R (Bolker & R Development Core Team, 2016). This is a common method to compare models in a multimodel inference paradigm (Burnham & D. R. Anderson, 2002).[9]

Results: model comparison

The ranked list of the models with their respective AICc values is provided in Table 3.8, which is also visualized in Figure 3.4.

[9] It has been argued (Greven & Kneib, 2010) that using AICc to compare models with different structures of random factors leads to a bias in favor of a more complex random factor structure. For this reason, to ensure the model comparison is fair, in linear models we only use random intercepts over individual constructions. In logistic models (accounting for individual differences) we would ideally use random intercepts over individual learners and constructions, but some of the models with random intercepts did not converge, therefore we used simple logistic regression without random effects.

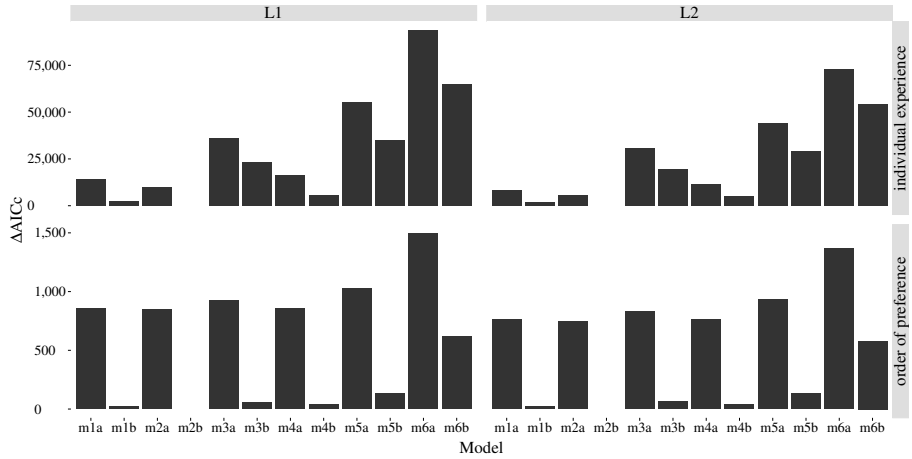


Figure 3.4: Model rankings visualized. Δ*AICc* for a model *M* in each subplot shows the difference between the *AICc* of the best model in that subplot and the *AICc* of the model *M*. Δ*AICc* of the best model in each subplot is 0, and higher *AICc* values correspond to worse model fits.

First we have to note that models m2a–m3a and m2b–m3b in some cases yielded multicollinearity problems. This was caused by the presence of two contextual frequency measures in these models, which sometimes were highly correlated even after applying the data transformations. The models which show this problem, even if ranked rather high, may not be very informative in terms of their coefficients.

Furthermore, we notice that the order of the models in the four lists is not identical, although there are clear similarities. The original model m1a is far from being the best one in any list. A pairwise comparison of the models demonstrates that m1b–m6b, which include the marginal verb frequency $F(v)$, always fit the data better than their respective counterparts without $F(v)$: m1a–m6a. In other words, adding $F(v)$ to any model improves its fit. If we further look only at the ranks of the "better" models m1b–m6b, we can see that the models with two contextual frequency measures (m1b–m3b) generally outperform the models with only one such measure (m4b–m6b). The only exception from this pattern is the single-measure model m4b, which is ranked third in each list, always higher than m3b. In all the four lists, the best model is m2b, therefore we look at this model in more detail in the following section.

Predictive power of each factor

To look at the impact of individual predictors in the refined model, in Table 3.9 we provide the summary of the model m2b ranked highest in each list. To account for the random variance, we refit m2b to each data set, this time including random slopes for each predictor (linear models accounting for order of preference), or random intercepts (logistic models accounting for individual variation).

Table 3.9: Summary of the best models of m2b type.

a. L1 simulations: model accounting for individual differences

$Prod. \sim F(v)+F(v,c)+A+Prt+(1|learner)+(1|constr.)$

Variable	β	SE[a]	$95\%CI$[a]	VIF
$F(v)$	0.64	0.01	$[0.63, 0.66]$	1.03
$F(v,c)$	0.65	0.01	$[0.64, 0.67]$	6.48
$A(v,c)$	0.11	0.00	$[0.10, 0.11]$	6.54
$Prt(v,c)$	0.33	0.01	$[0.31, 0.35]$	1.07

b. L1 simulations: model accounting for order of preference

$PP \sim F(v)+F(v,c)+A+Prt+(1+F(v)+F(v,c)+A+Prt|constr.)$

Variable	β	SE[b]	$95\%CI$[b]	VIF
$F(v)$	0.29	0.04	$[0.21, 0.37]$	1.36
$F(v,c)$	0.73	0.10	$[0.53, 0.92]$	5.14
$A(v,c)$	0.07	0.05	$[-0.02, 0.17]$	5.78
$Prt(v,c)$	0.10	0.02	$[0.06, 0.14]$	1.14
$R^2_m = .52$, $R^2_c = .72$				

c. L2 simulations: model accounting for individual differences

$Prod. \sim F(v)+F(v,c)+A+Prt+(1|learner)+(1|constr.)$

Variable	β	SE[a]	$95\%CI$[a]	VIF
$F(v)$	0.63	0.01	$[0.62, 0.66]$	1.06
$F(v,c)$	0.60	0.01	$[0.58, 0.62]$	6.29
$A(v,c)$	0.16	0.01	$[0.15, 0.17]$	6.31
$Prt(v,c)$	0.33	0.01	$[0.32, 0.35]$	1.07

d. L2 simulations: model accounting for order of preference

$PP \sim F(v)+F(v,c)+A+Prt+(1+F(v)+F(v,c)+A+Prt|constr.)$

Variable	β	SE[b]	$95\%CI$[b]	VIF
$F(v)$	0.30	0.04	$[0.22, 0.38]$	1.45
$F(v,c)$	0.77	0.11	$[0.55, 1.00]$	6.18
$A(v,c)$	0.06	0.05	$[-0.05, 0.16]$	6.79
$Prt(v,c)$	0.08	0.02	$[0.05, 0.12]$	1.13
$R^2_m = .53$, $R^2_c = .75$				

[a] Values are based on the Wald tests.

[b] Values are estimated via parametric bootstrap with $1,000$ resamples.

Looking at the summary in Table 3.9, we first observe that the four fixed factors in the linear models explain 52% of the variance for L1 data, and 53% for L2 data (see R^2_m coefficients). This is higher compared to the original prediction models for the same data sets (47%, see section 3.4.3).

Next, we can see that all the models yield collinearity problems: the variance inflation factor for $A(v,c)$ and $F(v,c)$ varies between 5.14 and 6.79. This suggests high collinearity between the two predictors. This is supported by the high correlation between βs for $A(v,c)$ and $F(v,c)$ in all models, varying between -0.89 and -0.91. Considering that the random slopes for $A(v,c)$ and $F(v,c)$ could not be included into the logistic models, the random variation in these models may be underestimated. In sum, even though $A(v,c)$ and $F(v,c)$ demonstrate their independent effects in the two logistic models, the respective β-coefficients may not be very informative.

The coefficients for $Prt(v,c)$ in linear models are also rather small, 0.10 and 0.08. Most importantly, the effect of $F(v)$ is high in all the models.

Interim discussion

The comparison of prediction models supports our proposal that the marginal verb frequency plays an independent role in predicting verb production in our simulated data. The parallel use of two contextual frequency measures appears to improve the model fit overall, contrary to our expectations. Yet, including two contextual frequency measures leads to collinearity issues: there is often a trade-off between the overall fit of the model to the data and the informativeness of its β coefficients. The use of a single measure is supported by our analysis of individual predictors, which suggests that the contextual frequency can be considered as a single component: joint frequency and Attraction capture the same type of syntagmatic relation between verbs and constructions. In other words, it is the combined effect of contextual frequency which is important, but not the individual effect sizes of joint frequency and Attraction. If one needs to chose a single contextual frequency measure between joint frequency, Attraction, and ΔP-contingency, our analysis suggests that joint frequency is the best measure: recall the high ranks of model m4b.

Considering contextual frequency as a single component, its individual impact in all the models is the highest, compared to the other predictors. The impact of prototypicality appears to be rather small in some refined models, but so it is in the original models as well: again, recall that our computational model may underestimate the importance of this factor.

3.5 General discussion

In this study we examined whether the selection of verbs within constructions could be explained by the distributional and semantic properties of these verbs and constructions, to see which factors may be responsible for establishing links between verbs and constructions in speakers' minds. We started from adopting the proposal by EOR that the frequency of production of a verb in a construction can be predicted by the joint

verb–construction frequency, the contingency of verb–construction mapping, and the prototypicality of the verb meaning. In what follows, we first briefly recapitulate how our simulations are similar and dissimilar to the human data. Since semantic prototypicality is the main issue in this respect, we discuss it next. The discussion is continued with a comparison of the results across three types of analysis provided above. Next we explain how the prediction model can be improved by avoiding multiple measures of contextual frequency, and by including marginal frequency instead. Additionally, we discuss how the use of form-based representations of constructions may have affected the findings, and address other theoretical challenges. Finally, we briefly talk about the computational model used in this study, and provide a short conclusion.

3.5.1 Simulations vs. human data

We used a computational model of construction learning to simulate the verb production experiments from EOR's studies. The analysis of verbs produced in the computational simulations demonstrated the model's reasonable performance on the target task: given a construction, the model mostly produced verbs that had been attested in this construction in the input. There were some exceptions, which suggest that the model was able to perform sensible generalizations over individual verb usages. At the same time, the type of the input data used in this study made it impossible to directly compare the verbs produced by the model to those produced by human participants, suggesting that we cannot claim that the model exactly replicated human linguistic behavior in the target task.

Our initial correlational and regression analyses showed main effects similar to those in the original experiments of EOR. In particular, we observed independent contributions of all the three predictors to explaining the frequency of verb production. Additionally, a preliminary comparison of the verb lists produced by the model in L1 vs. L2 simulations demonstrated that the degree of difference between the two lists was similar to that reported by Römer et al. (2014) for native German vs. native English speakers. However, a qualitative comparison between the simulated L1 and L2 verb lists is still needed. The main difference between the results obtained in our simulations and those reported by EOR related to the effect of semantic prototypicality, which appeared to be lower in our simulated data. We discuss this issue next.

3.5.2 Meaning prototypicality, data sparsity, and semantic coherence

We proposed three possible explanations for the low impact of semantic prototypicality: (1) the role of verb semantics is underestimated in the learning algorithm used by our model; (2) verb semantic representations in our data sets are impoverished compared to those in human speakers; (3) our semantic prototypicality measure performs poorly on infrequent constructions due to the data sparsity. Regarding the last explanation, we also found that the correlations between semantic prototypicality and verb production frequency were also low within some frequent constructions in our data set, for which dense information on verb use was available: ARG1 VERB and ARG1 VERB ARG2.

We suggest this has to do with the degree of semantic coherence of a construction. Following the setup of the original studies, we have used highly abstract constructions defined by their shallow form, which may not be semantically coherent. In particular, if we look at the verbs produced within the most frequent construction ARG1 VERB ARG2, these comprise several semantic groups: verbs of mental state (e.g., *want*), verbs of transfer (e.g., *buy*, *sell*), verbs of communication (e.g., *announce*), and many others. Given this variety, the construction is unlikely to have a single semantic core surrounded by multiple peripheral verbs. Instead, there are multiple semantic centers, and a single measure of semantic prototypicality may not capture such organization well, in particular when some semantic verb classes within a construction are much richer than others. This might be why we do not observe an effect of prototypicality in such constructions. In contrast, the effect is larger in constructions whose semantics is more coherent, because they actually have a single "prototypical" core. To give an example, the ARG1 VERB ARG2 ARG3 construction in our data (which comprises ditransitive verb usages, but also allows for adverbial arguments) is represented by eight verbs: *drag*, *give*, *hang*, *lead*, *place*, *pull*, *send*, and *tell*. Most of these are physical action verbs, the only exception being *tell*, hence high semantic coherence and a high effect of semantic prototypicality.

To compare, Theakston et al. (2004) in their study of early verb use did not find enough support that semantic prototypicality of a verb could predict the age when this verb first appeared in the child's speech, and the constructions they used – SVO, VO, and the intransitive – were highly abstract, and thus unlikely to be semantically coherent. This may also explain why the prototypicality effect was observed in the studies of EOR: they only focused on various constructions with locative semantics in their analyses, which may be more semantically coherent.

The question whether the effect of prototypicality is related to the degree of semantic coherence of a construction requires further investigation. As a counter-argument to this claim, Ambridge, Bidgood, Pine, Rowland, and Freudenthal (2015) find the effect of semantics in the passive, a semantically general construction. Note, however, that the interpretation of semantics in their study (and in other related studies: e.g., Ambridge et al., 2014, 2012) differs from semantic prototypicality as defined in this study. The reasoning behind this study (following EOR) is that more prototypical verbs are produced more frequently (because of how the activation spreads within a semantic network). This is why semantic verb features used in our study must capture the essential properties of the respective events. In contrast to this, the idea in the series of studies mentioned above is that particular nuances of verb meanings help in acquiring restrictions on the verb use. Therefore, these studies focus on very specific fine-grained features of a verb meaning, which do not necessarily provide much information about the general semantics of the event, but do help in discriminating between different verbs and verb classes. This account is largely based on Pinker's (2013) theory, in which "it's not what possibly or typically goes on in an event that matters; it's what the verb's semantic representation is choosy about in that event that matters" (p. 127). For this reason, the effect of semantics in this study and in EOR's study is not immediately comparable to the findings of Ambridge and colleagues. Building more comprehensive verb meaning representations based on both general event features and fine-grained

discriminatory features could open new prospects in this area: such representations could be used for training both our computational model and the model of Ambridge and Blything (2015).

3.5.3 Comparing the results across three types of analysis

We further carried out two additional analyses, to account for the potential between-learner variation in the linguistic input, and for the order of verb production by each (simulated) learner. These additional analyses of our simulated data suggest that the type of analysis may affect the main findings, in particular in terms of the observed effect of ΔP-contingency, which we address below. This is consistent across the two additional analyses, suggesting that both individual variation and order of learners' preference is important, which is in line with studies suggesting that individual differences play a role in language learning e.g., R. Ellis, 2004, and that speakers do not arrive at the same mental grammar e.g., Dąbrowska, 2012; Misyak and Christiansen, 2012. To verify the predictions made by our model in this respect, we would need to compare the results to human empirical data on individual variation and order of preference, which are missing yet.

3.5.4 Multiple measures of contextual frequency

Contingency may sometimes fail to demonstrate its independent effect because of the other variable included into the prediction model: joint verb–construction frequency. Both variables capture how well a verb and a construction go together (i.e., contextual frequency). If the hypothesized cognitive effect of the verb–construction association is loaded on both variables, one of them may show no independent impact. This issue was addressed by testing a number of alternative prediction models. One of our questions was whether models with one or with two contextual frequency measures would predict the data better. Our findings in this respect were somewhat inconclusive. On the one hand, prediction models which included two such measures were in general ranked higher than models which included only one measure. On the other hand, the independent effects of both joint frequency and contingency were not always present within the same prediction model. In fact, it was the combined impact of the two measures that was consistent across prediction models, but not the independent effect of each contextual frequency measure. This is why we suggest that it is a single effect of the contextual frequency that is cognitively plausible, while each measure (i.e., joint verb–construction frequency, Attraction, or ΔP-contingency) provides a particular quantitative representation of this effect. The correlation between the measures may be lower or higher in a specific data set, and this is why sometimes, but not always, it is justified to include two contextual frequency measures into a prediction model.

The relation between association strength and joint verb–construction frequency may also resemble the relation between the effects of entrenchment and preemption on learning argument structure restrictions, described by Ambridge, Bidgood, Twomey, et al. (2015). Both the entrenchment and preemption hypotheses predict that the distribution of verbs over argument structure constructions affects the learning of the related

usage restrictions, because of the verb's occurrence in either competing constructions (preemption hypothesis), or in all constructions (entrenchment hypothesis). In fact, independent contributions of these two factors within the same prediction model have been sometimes found (e.g., Blything et al., 2014; Ambridge, 2013). Yet, Ambridge, Bidgood, Twomey, et al. (2015) suggest that entrenchment and preemption are not independent mechanisms, but only effects that may or may not be observed, depending on the exact set of constructions in a study. Similarly, the effects of both association strength and joint frequency in our study capture the same mechanism of competition between verbs in the speaker's mind.

Whenever a single measure of contextual frequency must be considered in the analysis, our study supports joint verb–construction frequency as the best measure, although more research is needed in this respect. In particular, a more advanced factor analysis (e.g., of the type employed by Maki & Buchanan, 2008) may clarify the relationship between different measures of contextual frequency.

3.5.5 Marginal verb frequency

The results in terms of marginal (overall) verb frequency are more straightforward. We found a consistent effect of the marginal verb frequency, in line with some data in language acquisition research (Blything et al., 2014; Theakston et al., 2004). Besides, this effect was independent from that of joint verb–construction frequency, in accordance with the proposed distinction between cotextual and cotext-free entrenchment (Schmid & Küchenhoff, 2013; Schmid, 2010). Based on this result, the effect of marginal verb frequency is worth investigating in human production data. In particular, this is theoretically supported by some existing memory research (Madan et al., 2010; Hockley & Cristi, 1996), where item memory (reflected in our case in marginal verb frequency) is believed to be independent of associative memory (in our case: contextual frequency measures).

At the same time, the marginal frequency of a verb may relate to the diversity of syntactic environments in which this verb is used. Although some frequent verbs may be used in only a few types of constructions, in general a verb's frequency is likely to be higher when the verb is used in a great variety of construction types. In this capacity, the observed effect of the marginal verb frequency in our study may be similar to what Naigles and Hoff-Ginsberg (1998) report in their child language study: verbs which appear in diverse syntactic frames are used more frequently.

Speaking about the effect of marginal verb frequency compared to that of contextual frequency, our data suggests that contextual frequency has a higher impact on verb selection than marginal frequency. This is a rather reasonable conclusion: when cued by a construction, speakers are more likely to produce frequent verbs related to the cue, rather than verbs which are frequent overall. However, if there are two verbs fitting the construction equally well, the one which is more frequent overall will be preferred. This is consistent with the fact that constructions attract only some verbs and reject other verbs (e.g., Stefanowitsch & Gries, 2003; Goldberg, 1995).

3.5.6 Alternative construction representations

In this study, constructions were defined solely by their shallow form. This is a common approach in corpus linguistics, because it is easy to automatically look for syntactic forms in a corpus. An efficient search for constructional meanings, on the other hand, would only be possible in a corpus that is semantically annotated, which is most often not the case. At the same time, constructions are commonly defined as pairings of form and meaning e.g., Croft, 2001; Goldberg, 1995; Langacker, 1987. Assuming a priori that a shallow pattern has a meaning does not guarantee that this meaning is unified and coherent, and that the hypothesized construction is cognitively real. Defining a construction by explicitly describing both its form and its meaning may be a better practice.

The described problem is particularly evident in the current study, as well as in EOR's studies. Form-based patterns do not predefine the argument roles, and therefore, could be interpreted by participants in multiple ways. This sometimes resulted in the production of verbs with different argument structures within the same pattern: e.g., *come* and *throw* in *he/she/it ___ across the ...*; or *eat* and *write* in *he/she/it ___ as the ...*; with some usages even looking ungrammatical: *he/she/it knows as the ...*, *he/she/it climbs of the ...*, etc. data from English native speakers in N. C. Ellis et al., 2014b. Similarly, in our study multiple semantic interpretations were possible, for example, for ARG1 VERB ARG2 ARG3. Besides, the problem in both studies is reinforced by the use of both animate (*s/he*) and inanimate (*it*) pronouns as the subject of each test stimulus: it may be argued that the animate pronouns represent an AGENT, while the inanimate pronoun is more likely to be a FORCE, hence two different constructions.

This leads us to the issue of the level of granularity of constructional patterns. It has been suggested that observed frequency effects may depend on the level of granularity of a construction under consideration (Lieven, 2010). The issue has also been touched on by Theakston et al. (2004), who show that different researchers employ different constructions in similar studies: for example, Ninio's (1999) VO and SVO constructions are combined within the same transitive construction by Goldberg (1998). In other words, the results may be also conditional on the chosen level of granularity of constructions. Together, these issues call for a similar analysis of different constructional representations. An earlier study with the same computational model (Matusevych, Alishahi, & Backus, 2015a) suggests that the observed effects of input-related factors on verb selection depend, indeed, on the type of constructional representations. Yet, the issue requires further investigation.

3.5.7 Further theoretical challenges

This study additionally touches on some theoretical questions that need to be addressed in the future. One of them is the relation between naturalistic and experimental verb production data. In this study, just as in EOR, the production of verbs was elicited by constructional stimuli. This is different from related studies of verb production by children (e.g., Theakston et al., 2004; Naigles and Hoff-Ginsberg, 1998; Ninio, 1999a, 1999b), which work with naturalistic samples of child language. It is unclear whether

such "field" data are directly comparable to the experimental data from elicited production experiments: for example, in the natural data some verbs within a construction may be used more often simply because of the higher referential frequency of the actions, states, etc. they refer to.

This leads us to the problem of defining the true nature of such phenomena as a unit's frequency, semantic prototypicality, and entrenchment. In this study we have simplistically assumed that a unit's frequency reflects its entrenchment, and that the frequency is independent of prototypicality, but these relations are not so trivial (Schmid, in press; Geeraerts, Grondelaers, & Bakema, 1994). To mention only some complications, when a unit is perceptually salient in speech (e.g., a word which is very unusual in a given genre or context), it may contribute more to memory consolidation (and entrenchment) than when it is less salient. Besides, it has been argued that the frequency (e.g., the referential frequency) does play a role in determining prototypicality (see an overview in Gilquin, 2006). Highly controlled studies of these phenomena could clarify the theory, and computational modeling can be helpful in this respect.

3.5.8 Computational model of construction learning

The final issue to address is the computational model employed in this study. On the one hand, simulation results always depend to a certain extent on the chosen model. To give an example from this study, semantics in our model is only one out of many features that guide construction learning, and the role of semantics may be underestimated compared to human learners. If that is indeed the case, then the differences in the size of effects reported in this study and in EOR's study may be attributed to the model's inability to replicate the exact linguistic behavior of human speakers.

On the other hand, when the model, as in our case, produced results similar to some existing experimental findings, this supports the plausibility of the model. The similarity of our results based on L1 and L2 simulations to those of EOR supports the assumption that incidental learning takes place in both L1 and L2 learning. Besides, the fact that the model is able to produce verbs relevant for a given construction, suggests that the emergent constructional representations in the model may approximate well what humans learn. Unfortunately, the type of the input data used in the present study does not allow us to compare the production data to the original study in terms of specific verbs and constructions, and this issue should be addressed in the future to better evaluate the potential of this computational model. One fruitful direction may be to investigate the role of frequency vs. verb semantics in the process of learning verb–construction associations (as in Ambridge & Blything, 2015), as opposed to looking at the static knowledge of such associations in simulated speakers.

3.6 Conclusion

In this chapter we presented a computational simulation of the verb production experiments of N. C. Ellis et al. (2014a, 2014b) using a usage-based, probabilistic model of argument structure construction learning. Our experiments showed that the model's

performance in the verb production task could be predicted by the same variables as the performance of human participants in EOR's experiments. Our follow-up analyses addressed some methodological limitations of these experimental studies, and suggested a refined version of the verb production model proposed by EOR. In particular, the frequency of production of verbs within argument structure constructions in our simulated data could be predicted by joint verb–construction frequency, contingency of verb–construction mapping, and prototypicality of verb meaning, although the effect of prototypicality was lower than in the human data. We then carried out two additional analyses on the same simulated data sets, to account for individual variation between speakers and for order of their verb preference. The results suggest that the type of analysis may affect the main findings. In particular, the effects of both joint verb–construction frequency and contingency measure within the same prediction model are not always observed. Finally, we compared a number of prediction models with different variables. The best prediction model included overall verb frequency in the input data, semantic prototypicality, and two contextual frequency measures: joint verb–construction frequency and Attraction. However, the high correlation between the contextual frequency measures suggests that their effects are combined rather than independent. We believe this refined prediction model should be tested on experimental data with human subjects.

CHAPTER 4

The impact of first and second language exposure on learning second language constructions[1]

4.1 Introduction

How is the learning of argument structure constructions in a second language (L2) affected by basic input properties such as the amount of input and the moment of L2 onset? This question touches on an important claim in usage-based theories of learning, namely that our knowledge of language is directly based on our experience with it, in particular the linguistic input we are exposed to. The amount of input and the moment of L2 onset are variables which are widely discussed in the field of second language acquisition (SLA). Yet, the exact question posed above has not received much attention either in usage-based linguistics or in SLA, although many closely related issues have been studied.

The impact of the moment of onset and the amount of exposure has been investigated in the domain of first language (L1) word learning, resulting in a number of competing hypotheses (see overviews by Hernandez & Li, 2007; Juhasz, 2005). Most researchers agree that word learning is affected both by the time of the word onset and the amount of exposure to that word. These findings might be applicable to the development of abstract constructions as well, especially since cognitive linguistics rejects a strict dichotomy between language domains such as lexis and grammar. However, some argue that there is a functional distinction between lexical items and abstract

[1] This chapter is derived in part from an article published in *Bilingualism: Language and Cognition* 16 September 2015 © Cambridge University Press, available online: http://dx.doi.org/10.1017/S1366728915000607

constructions (Boas, 2010). Learning abstract constructions is different from word learning in that it is based on pattern-finding skills such as analogy and categorization (Abbot-Smith & Tomasello, 2006; Tomasello, 2003). There is also some neurological evidence that abstract constructions and lexical items are characterized by different representation in the human brain, and might be subject to different learning mechanisms (Pulvermüller, Cappelle, & Shtyrov, 2013; Pulvermüller & Knoblauch, 2009). The difference in how words and abstract patterns are stored in memory is also one of the central points in the declarative/procedural model (e.g., Ullman, 2015; Pinker & Ullman, 2002). These differences suggest that the findings on word learning are not immediately generalizable to construction learning, and vice versa.

Interest in L2 construction learning has been growing recently (Ambridge & Brandt, 2013; Tyler, 2012; Gries & Wulff, 2009, 2005, etc.). In particular, it has been investigated how L2 construction learning depends on distributional properties of the linguistic input, such as the frequency of using verbs in constructions, or the generality of verb meanings (N. C. Ellis et al., 2014a; Römer et al., 2014; McDonough & Nekrasova-Becker, 2012; Boyd & Goldberg, 2009; Year & Gordon, 2009), but not on the amount of input and the moment of onset – factors commonly discussed in SLA literature.

The biggest challenge of studying input-related factors and their impact on language development is that their effects are often hard to disentangle. Studies on both L1 and L2 learning have shown that the amount of exposure and the time of onset are often confounded (Muñoz & Singleton, 2011; Flege, 2008; Ghyselinck, Lewis, & Brysbaert, 2004), and observational and experimental studies cannot easily solve this problem. In contrast, computational modeling allows researchers to manipulate input properties one at a time and to examine their individual impact on language development (Monner et al., 2013; Zhao & Li, 2010; Monaghan & A. W. Ellis, 2002; A. W. Ellis & Lambon Ralph, 2000).

In this study, we use a computational tool for investigating how the learning of L2 argument structure constructions depends on the moment of L2 onset and the amount of L2 input. Our goal is not to develop a cognitive model of how humans learn a second language, but to simulate L2 construction learning from bilingual input in a purely data-driven fashion and without incorporating any unrelated (e.g., biological or social) factors. This approach allows us to analyze how the development of L2 constructions changes as a result of systematic manipulations of the amount of exposure and the time of onset. Although the use of computational modeling prevents us from making conclusive claims about human L2 construction learning, our simulations can provide useful intuitions on this process, which may then be tested with human subjects.

4.1.1 Variable definitions and the problem of confounding

SLA literature often talks about the age of onset, or the age of acquisition. However, the appropriateness of the term ‘age’ has been questioned. Talking about age has been suggested to be not informative, because this is not a basic variable, but a macrovariable that aggregates multiple interrelated factors (e.g., Flege, 2008; Montrul, 2008; Jia & Aaronson, 2003), which can be grouped into three broader categories (Larson-Hall, 2008; Moyer, 2004; Jia & Aaronson, 2003):

1. Biological–cognitive factors: state of neurological and cognitive development (Birdsong, 2005), neuroplasticity (Long, 1990), etc.

2. Socio-psychological factors: motivation, the need to be fluent, self-perception of fluency, etc. (Moyer, 2004).

3. Experiential factors: amount and distribution of L1 and L2 input, contexts of use, contacts with L2 native speakers, etc. (Moyer, 2004).

The proposed categorization indicates how important it is to exactly specify which 'components' of age are being studied. This can be especially well illustrated by studies on the age of acquisition in L1 processing. Some of them (e.g., Mermillod, Bonin, Méot, Ferrand, & Paindavoine, 2012; Izura et al., 2011; A. W. Ellis & Lambon Ralph, 2000) use the term 'age of acquisition' interchangeably with 'order of acquisition'. This can be confusing, because conventionally 'order' only reflects a sequential nature of the input presentation during the learning, while 'age' is associated with biological changes that accompany maturation. Speaking in terms of the categorization proposed above, order of acquisition falls into the category of experiential factors, while 'age' is a proxy variable for the three groups. Thus, the relative onset of two languages is better described by such terms as 'moment of onset', or 'time of onset', or simply 'onset', to avoid references to biological–cognitive or socio-psychological factors.

Strict variable definitions, however, do not resolve the problem of their confounding. In the SLA literature, the contributions of the amount of L2 input and the L2 onset have been debated. In particular, Flege (2008) claims that the confounding of the variables has resulted in underestimating the predictive power of L2 input, compared to the L2 onset. Similarly, studies on L1 processing have discussed what affects the word processing: the amount of exposure to a specific word (i.e., its frequency), or the moment of its first encounter. Some theories, such as the cumulative frequency hypothesis (M. B. Lewis, Gerhand, & Ellis, 2001) and the frequency trajectory theory (Mermillod et al., 2012), attribute a determining role to the frequency, rather than to the order of acquisition. Other theories, such as the lexical–semantic competition hypothesis (Brysbaert & Ghyselinck, 2006; Belke, Brysbaert, Meyer, & Ghyselinck, 2005), focus more on the order effect, claiming it can be both frequency-related and frequency-independent. The problem of confounding is difficult to solve with human learners, which justifies the use of computational models in the field.

Another reason to use highly controlled computational models is a lack of accurate measures able to capture, for example, the actual amount of language input that learners are exposed to. Muñoz and Singleton (2011) describe some of the difficulties involved in measuring the actual amount of L2 input, both in immersion and in classroom settings. A systematic investigation of the impact of L2 onset and L2 amount requires addressing these methodological challenges. Computational modeling has been widely used to study related issues, as we show in the next section, although no models have simulated the bilingual learning of abstract constructions.

4.1.2 Existing computational models

Connectionist simulations have been widely used in studying the order of acquisition effects in L1 processing (e.g., Mermillod et al., 2012; Lambon Ralph & Ehsan, 2006; Monaghan & A. W. Ellis, 2002; A. W. Ellis & Lambon Ralph, 2000). In particular, A. W. Ellis and Lambon Ralph (2000) demonstrated that, as a neural network is exposed to more words, its plasticity is reduced, limiting its ability to learn new (late) words. They also showed how order of acquisition might interact with frequency. Although this type of research investigates variables relevant to our study, it deals with data from a single language.

As for bilingual learning, Zhao and Li (2010) simulated English–Chinese lexical acquisition under different onset conditions. In their experiments lexical items were represented as pairings of phonological and semantic features. The manipulated variable was the amount of L1 input that their computational model received prior to the moment of L2 onset. When the onset of the two languages was the same (simulating an early bilingual), the model's proficiency in both languages was comparable. However, when the model received a substantial amount of L1 input prior to the L2 onset (i.e., a late L2 learner), it performed better in L1 than in L2. This outcome supported the hypothesized relationship between the level of L1 neural entrenchment and the L2 attainment. In short, Zhao and Li (2010) demonstrated the negative effect of L1 entrenchment on L2 learning in the lexical domain. In another study on bilingual learning, Monner et al. (2013) used computational modeling to investigate the effect of L1 entrenchment in a different domain, namely the learning of morphological gender from phonological features in Spanish and French. Using a similar experimental design, they demonstrated the negative effect of L1 entrenchment on learning L2 lexical morphology.

These two studies demonstrate the negative effect of L1 entrenchment on L2 learning at the word level. However, there are no comparable studies for language units beyond the word level, in particular abstract linguistic constructions. In the next section, we describe the computational model used in this study to simulate bilingual construction learning.

4.2 Method

4.2.1 The model

The model that we use in the current study is an adaptation of a model of early argument structure acquisition (Alishahi & Stevenson, 2008). This original model was inspired by usage-based theories, in particular Construction Grammar (as informed by Goldberg, 1995), and it has successfully replicated several patterns of construction learning by children. The model employs a domain-specific unsupervised learning mechanism, inherited from a model of human category learning (J. R. Anderson, 1991). Just as in human learning, the model processes input iteratively, so that linguistic knowledge slowly builds based on experience. All this makes the model a good candidate for our study.

We simulate one specific task – learning argument structure constructions from linguistic and conceptual exposure in two languages. According to Goldberg (1995), this is a special class of abstract constructions (or form–meaning mappings) that provides the basic means of clausal expression. Goldberg et al. (2004) consider these constructions as "argument structure generalizations" – high-level associations of form and meaning, which gradually emerge from categorizing individual instances. These views on learning are reflected in our computational model.

Next we provide a conceptual description of the model, while its formal description can be found in Appendix B.

Exposure

The exposure consists of a number of argument structure instances (AS instances) represented as assemblies of different information cues (or features). Each instance corresponds to an individual verb usage: an utterance and the respective perceptual context. A sample verb usage and its corresponding AS instance are presented in Table 4.1. The features include the head predicate (verb) and its semantic properties (lexical meaning), the number of arguments that the verb takes, argument heads, their cases, their semantic and event-based (thematic role) properties, prepositions and the syntactic pattern (which reflects the word order and the presence or absence of prepositions at specific slots). Instead of representing lexical meanings or thematic roles symbolically, we use a set of elements for each of these, following the theories of McRae et al. (1997), Dowty (1991). Composite representations allow the model to estimate the similarity between different meanings or thematic roles. Sets of elements may be rather large, therefore for brevity we only show three elements for each feature in Table 4.1. Unlike semantic and role properties, some other features, for example head predicate and prepositions, take language-specific values. When a feature such as argument case is absent in a language (e.g., English), it is assigned a dummy value (N/A). Note that the cases are the only morphological features in our setup, other morphological elements as well as articles are ignored, as they contribute little to differentiating between argument structure constructions.

Learning process

The learner maintains a set of constructions, which are represented as generalizations over AS instances. More specifically, each construction is an assembly of feature values of all instances that the model has decided to add to this construction. The learner tracks the frequency of each construction (the number of participating instances), together with the frequencies of all feature values, yet the original instances are not recoverable. The learner receives one instance at a time and iterates over all the acquired constructions, to find the one that can best accommodate the new instance. Two factors determine which construction the new instance is added to:

1. The frequency of each construction in the previously encountered input. This follows the idea in usage-based linguistics that linguistic units become entrenched through their use (e.g., MacWhinney, 2012; Schmid, 2007; Langacker, 1987).

Table 4.1: An example AS instance extracted from a verb usage *I ate a tuna sandwich.*

Feature	Value
Head predicate	*eat*
Predicate properties	{consume, take in}
Number of arguments	2
Argument 1	*I*
Argument 2	*sandwich*
Semantic properties of argument 1	{SELF, PERSON, . . . , ENTITY}
Semantic properties of argument 2	{SNACK FOOD, DISH, . . . , ENTITY}
Role properties of argument 1	{LIVING THING, ENTITY, . . . , ORGANISM}
Role properties of argument 2	{SOLID, SUBSTANCE, . . . , ENTITY}
Case of argument 1	N/A
Case of argument 2	N/A
Syntactic pattern	ARG1 VERB ARG2
Prepositions	N/A

A construction which already contains a large number of instances is more entrenched, or more readily accessible, therefore the learner is more likely to add the new instance to this construction. Note that this is to a certain extent similar to processing limitations that arise in connectionist models at later stages of learning (e.g., A. W. Ellis & Lambon Ralph, 2000). However, the maximal processing capacity of our model (the number of categories) is not predefined as is the number of units in connectionist models, and we make no claims regarding how similar the two approaches are.

2. The similarity between the new AS instance and each construction, which is measured in terms of each feature independently (see Table 4.1 above). For example, if a construction and the new instance share the number of arguments, the syntactic pattern and the argument role properties, it is likely that this instance belongs to this construction. Vice versa, if a construction and the new instance have little in common in terms of feature values, the new instance is unlikely to be added to this construction. This similarity-based learning mechanism comes from the original model in J. R. Anderson (1991) and reflects the role of similarity in human categorization (e.g., Sloutsky, 2003; Hahn & Ramscar, 2001).

Upon estimating the two values, the learner adds the new AS instance into one of the constructions. However, especially at the beginning of the learning process, the best decision (as informed by the likelihood values) may be to create a new construction and add the new instance to this new construction (which would be identical to the instance). This happens when the new instance is very dissimilar to all the constructions

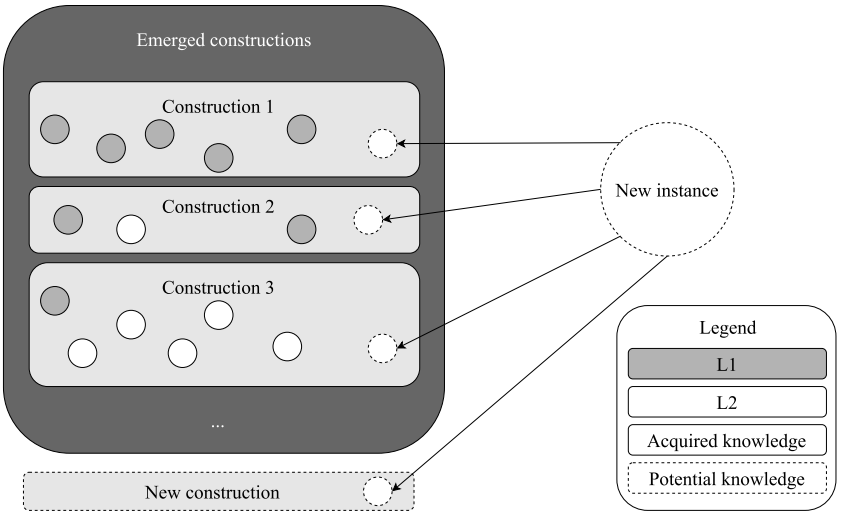


Figure 4.1: Deciding on a construction for a newly encountered L2 AS instance.

the learner has acquired so far.[2]

Exactly the same algorithm applies to L1 and L2 learning, as illustrated by Figure 4.1. Note that constructions may contain instances from only L1 or L2, as well as from both languages. Although such features as head predicate, arguments, and prepositions contain implicit information about the language of each AS instance, the model is not explicitly enforced to distinguish between the two languages. It is the input data and the probabilistic learning mechanism that determine to what extent L1 and L2 share their 'storage resource'.

We further illustrate the learning process in Figure 4.2, where an English speaker learning L2 German encounters an AS instance with the head predicate *gewinnen* "to gain". Note that construction 9 is associated, among other English verbs, with *gain*, and the instance headed by *gain* shares some feature values with the new instance headed by *gewinnen*. Thus, among all the existing constructions, construction 9 may be the most likely candidate for adding the new AS instance. Imagine, however, that the learner, upon receiving a substantial amount of English input, encounters a German instance with a syntactic pattern PREP ARG1 VERB ARG2 ARG3 (e.g., *Über die Nebenwirkungen weiß niemand das geringste.* "No one knows anything about the side effects."). This order of arguments is not typical for English, therefore the learner might not know a suitable construction to accommodate this AS instance, and is likely to create a new construction for the novel instance.

[2] In practice, it is difficult to estimate whether an instance is 'very' dissimilar to a construction. Our model has a parameter determining the cost of creating a new construction, which increases over time: the more constructions the model knows, the less likely a new one to be created (for more detail, see Appendix B.2).

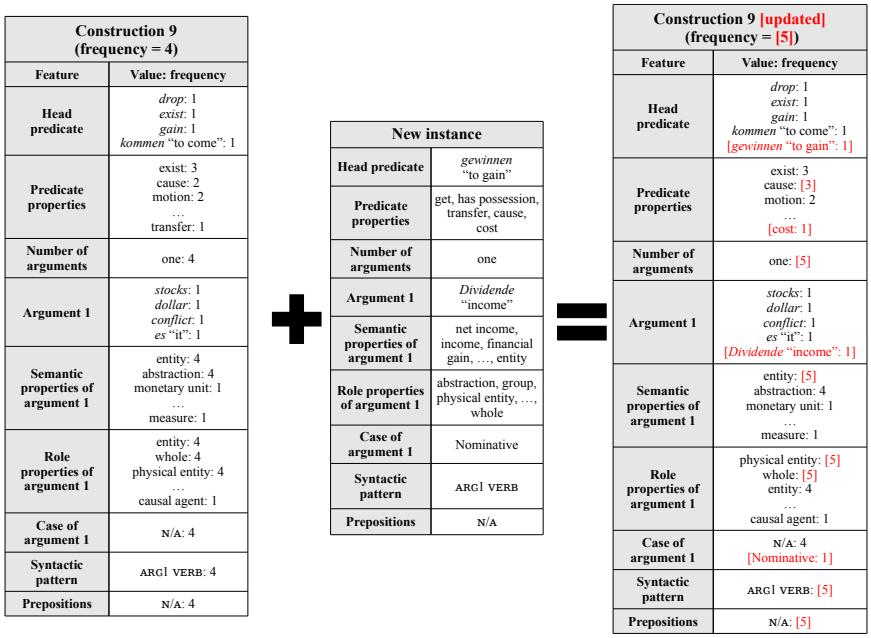


Figure 4.2: Updating a construction with a newly encountered AS instance. The frequency of the construction represents the number of AS instances it is based on. The frequency of each feature value equals to the number of participating AS instances showing this value for the respective feature. Square brackets denote updated elements.

Simplifying assumptions

Like all computational models in the field, our model simulates only certain aspects of learning, and makes a number of simplifying assumptions about the other aspects. Because we focus on the learning of abstract constructions, we assume that our simulated learner is able to segment the utterance and recognize all the words; it knows the meaning of most words in the utterance; it can identify the role of each participant in a given perceptual context; and it is able to infer the information about linguistic cases in the utterance. For the purpose of this study, we assume that the learning mechanism has acquired these types of knowledge and abilities by the moment it starts learning constructions, although we acknowledge that human learners acquire different types of knowledge in parallel (see, e.g., Lieven & Tomasello, 2008, for child learning).

4.2.2 Testing L2 proficiency

The model's knowledge of argument structure constructions is tested in terms of the accuracy of language use, both in production and comprehension. A formal description of the testing method is provided in Appendix B.3, while here we outline the general approach to testing and focus on the actual tasks. We use five tasks for evaluating the model, each of them testing a different aspect (or feature) of the model's construction knowledge. We provide the model with a number of test instances in which the values of some features are masked. Although it is possible to mask the values of multiple features at once, each of the tasks in this study masks only a single feature. Thus, for each test instance, the model has to predict the missing value of a particular feature given the values of the other features. The prediction accuracy in each task is estimated based on the match between the original (masked) value and the value predicted by the model.

Such approach relates to the view in usage-based linguistics that linguistic knowledge is reflected in language use. Although the main motivation for the task choice comes from the model architecture, the tasks we employ map onto some existing methodologies used either in L2 assessment or in experimental studies with children and adults (see Table 4.2). Note, however, that our test tasks are conceptually closer to spontaneous language use rather than to traditional language assessment,[3]; therefore, the examples (30–34) below are provided mostly for illustrative purposes, while the actual testing algorithm can be found in Appendix B.3.

[3] For example, in filling in verbs and prepositions we do not restrain the model from using L1 lexemes that it finds appropriate. In other words, the model has no explicitly implemented control mechanisms, similar to those that humans can use for inhibiting activated representations from a non-target language (e.g., Kroll et al., 2008; Green, 1998). At the same time, mixing L1 and L2 lexemes within the same utterance is not uncommon in bilingual speakers, as the literature on code-switching suggests (e.g., Auer, 2014). Although the lack of inhibitory control negatively affects the model's performance in the mentioned tasks, making it less comparable to human performance, our findings must not be affected, because the inhibitory control is consistently absent in all the experimental conditions.

Table 4.2: Assessment tasks with their descriptions and corresponding features in AS instances.

Masked AS feature	Task name	Description
Head predicate	Filling in verbs	"Fill-in-the-blank" test with removed verb
Prepositions	Filling in prepositions	"Fill-in-the-blank" test with removed prepositions
Syntactic pattern	Word ordering	Placing verb and prepositions in their correct positions
Predicate properties	Verb definition	Verb definition in a sentential context
Arguments' role properties	Role comprehension	Comprehension of argument roles in a given sentence–event pair

Filling in verbs

In this task we elicit the production of verbs that the model finds suitable in a given test instance. This is close to the method used in some experimental studies concerned with the learning of argument structure constructions, as they tend to examine the distribution of verbs in specific constructions (e.g., N. C. Ellis et al., 2014a; Gries & Wulff, 2005):

(30) Fill in a verb: *I ___ a sandwich*.

Filling in prepositions

The same design is used to elicit the production of prepositions. Filling in blank slots with missing prepositions is a classic task in L2 assessment (e.g., Oller & Inal, 1971):

(31) Fill in a preposition: *John gave an apple ___ Mary*.

Word ordering

Given the verb and its arguments, the task is to name a matching syntactic pattern. This is similar to a common L2 assessment task in which learners are asked to unscramble the words into a grammatical sentence (e.g., Wesche & Paribakht, 2000):

(32) Arrange the words to form a grammatical sentence: *ate, (a) sandwich, I*.

Verb definition

The task of deriving lexical meanings from contexts tests learners' ability to comprehend verbs. A similar definition task has been used, for example, for assessing children's vocabulary (Cain, 2007). A schematic example for our setup is given in (4):

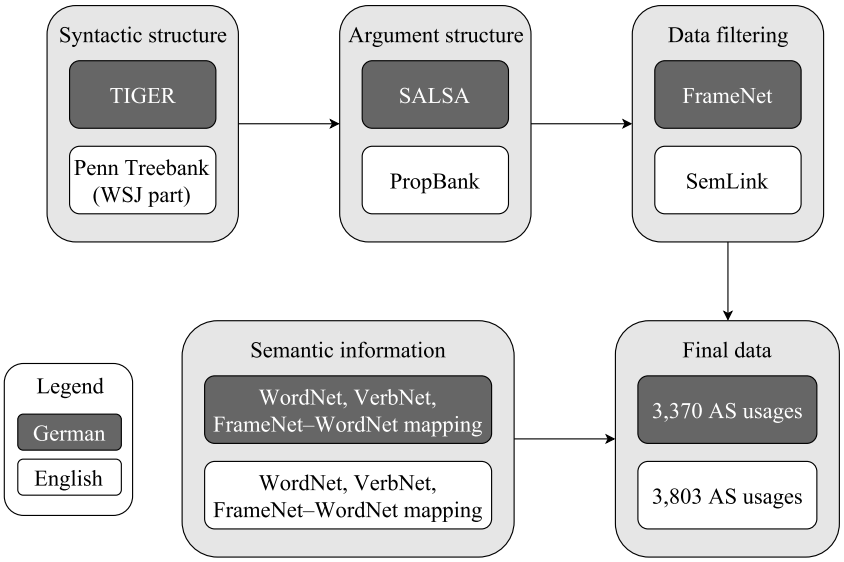


Figure 4.3: Schematic representation of the input data preparation.

(33) Describe the lexical meaning of *ate* in the sentence: *I ate a sandwich*.

Role comprehension

Studies in which humans have to learn new verbs (e.g., Wonnacott, Newport, & Tanenhaus, 2008; Akhtar & Tomasello, 1997) often test the acquisition of verb-general knowledge about the thematic roles of participants in a given event. Similarly, our model is required to describe the role of each participant in a given sentence–event pair:

(34) Describe the thematic roles of *I* and *(a) sandwich* in the sentence: *I ate a sandwich*.

4.2.3 Input and test instances

In preliminary experiments (Matusevych, Alishahi, & Backus, 2013) we tested the model on small data sets of German and English, in which argument structures were annotated manually. However, manual annotation of larger data sets would be very time-consuming. Instead, in the present study we extracted data from available annotated resources for the same languages. Essentially, the data come from German and English newspaper texts. Although these texts do not represent the kind of language that L1 and most L2 learners are exposed to, we used these corpora as the only large sources of English and German that contained all the necessary types of annotations related to argument structure.

Figure 4.3 schematically shows the resources we used and the steps we took for preparing the input data. The syntactically annotated data originate from the TIGER corpus for German (Brants et al., 2004) and the Penn Treebank for English (Marcus et al., 1994). The German SALSA corpus (Burchardt et al., 2006) and English PropBank (Palmer et al., 2005) contained the types of annotations that helped us to extract argument structure from sentences. Further, for consistency between the languages, we filtered the resulting sentences and kept only those that were annotated with FrameNet frames (see Ruppenhofer et al., 2006). While some German data were already annotated so in SALSA, for English we had to use the mappings between PropBank and FrameNet, provided in SemLink (Palmer, 2009). Finally, semantic features for individual lexemes were extracted from (G. A. Miller, 1995) and (Kipper Schuler, 2006). The existing mappings between WordNet and FrameNet (Bryl et al., 2012) also made it possible to automatically expand argument thematic roles into sets of elements. The procedure resulted in German and English data sets containing 3,370 and 3,803 AS instances, respectively. Note that the two data sets have similar, but not identical sizes. Besides, they may differ in the amount of noise originating from either the corpus annotations or from our data extraction procedures. This potentially may result in one of the data sets being more difficult to learn than the other.

Importantly, a substantial part of both German and English AS instances originated from embedded clauses. While in English main and embedded clauses have analogous word order, this is not so for German, where embedded clauses are usually verb-final. Consider the following English sentence (6) translated into German (7):

(35) The group said (that) it sold the shares.

(36) Die Gruppe sagte, dass sie die Aktien verkauften.

The word order in the English embedded clause in (6) is SVO, while the German order (7) is SOV. This is a natural difference if one considers each complex sentence as a whole. However, we represent each AS as an independent language unit, and the unnaturally large number of SOV sentences would make our data set a non-representative sample of German (simple) sentences. Ultimately, this would provide our model with an unrealistic tool to distinguish between English and German syntactic structures. Therefore, we 'recovered' German verb-second word order in embedded clauses by manually assigning the second position to the verb. Note, however, that the order of arguments was never changed, so that the data contained both SVO and OVS sentences.

From the resulting data sets, input to the model was sampled randomly, so each individual simulation represented a learner with a unique history of language exposure. Thereby, in our experiments we sometimes refer to different simulations as individual learners. The exact number of German and English AS instances as well as the temporal pattern of their presentation were determined by the experimental setup, however all the experiments were run twice – using German as L1 and English as L2, and vice versa.

Similarly, test instances are randomly sampled from the data. Learners are tested on different test sets, although every learner is repeatedly offered the same test set at certain intervals. Furthermore, each learner performs most language tasks on a single test set,

except for the task of filling in prepositions, for which an additional test set is prepared. This is because most AS instances in our data (approximately 70% for German and 90% for English) contain no prepositions, and sampling items randomly would result in having no prepositions in the majority of test instances. Therefore, we sample an additional test set for each learner, considering only instances with prepositions. Just as in human language learning, some test items may be identical to input items that the model has encountered. In other words, sampling the input and the test instances from the same data resembles better a natural language learning setting than splitting the data into a train and a test set (a common practice in computational linguistics). It is unlikely that the model can memorize specific instances and then simply reproduce them, because construction learning is implemented as a categorization task, without memorizing actual instances. However, to ensure that the model does not memorize the exact instances, we run an additional set of simulations, in which none of the learning data appear as test instances.[4] The described data is used in all the experiments that we report in the next section.

4.3 Experiments and results

This section describes the design and the results of three experiments. The first two are intended to test whether the general learning principle "the more, the better" holds for statistical learning of argument structure constructions – that is, whether the larger amount of L2 input results in higher L2 performance. We measure L2 amount both in relative (experiment 1) and absolute terms (experiment 2). Experiment 3 is designed to test how learners' L2 performance is affected by the time of L2 onset. In all the experiments, we quantify various amounts of input in terms of the respective number of AS instances. Furthermore, we adopt the following notations (see Figure 4.4):

1. E_T – total language exposure, both L1 and L2. E.g., $E_T = 12,000$ AS instances.

2. TO – the time of onset, expressed as the amount of L1 input prior to the L2 onset. E.g., $TO = 9,000$ L1 instances. $TO = 0$ defines a simultaneous bilingual.

3. E_{L2} – cumulative L2 exposure in absolute terms. E.g., $E_{L2} = 3,000$ L2 instances.

4. R – the ratio of L1 amount to L2 amount at each interval after TO. E.g., $R = 20:1$ means that the learner receives 20 times more L1 input than L2 input.

5. E_B – the amount of bilingual input, in which both L1 and L2 instances are present. E.g., $E_B = 6,000$ indicates that after TO, the learner receives 6,000 instances of bilingual input, where L1 and L2 are mixed in the proportion determined by R.

[4] In all the reported simulations a learning and a test set have been sampled from the same data, therefore the model might have encountered a substantial part of the test instances in the learning data. Yet, the additional simulations yielded very similar results. In other words, the main findings reported in this study are robust and do not depend on the sampling procedure.

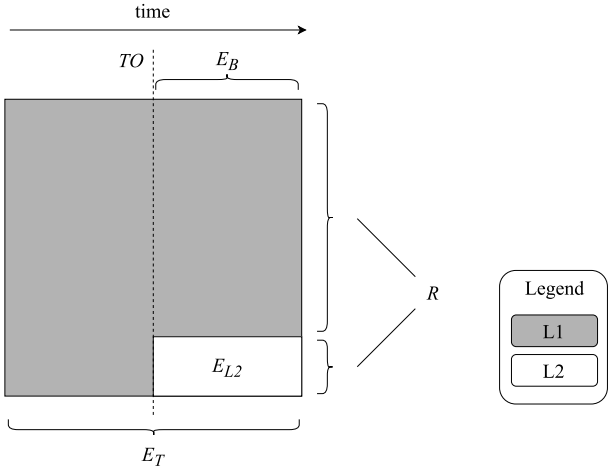


Figure 4.4: Notations used in the experiments.

4.3.1 Amount of L2 input

Experiment 1

In this experiment learners' exposure to L2 was measured in relation to their L1 exposure. To investigate whether the relative amount of L2 input would affect learners' L2 performance, we manipulated the ratio R in four groups of simulated learners, while keeping E_T constant. The first group of learners received equal amounts of L1 and L2 input at each learning interval after L2 onset, $R = 1:1$, while for the other groups R was set to $3:1$, $10:1$, or $20:1$, respectively. Such design simulated a common SLA setting: adult L2 learners are often exposed to the target language in small quantities, while L1 still dominates in their daily use. Each of the four groups consisted of 30 learners, for which both TO and E_B were set to 6,000 instances – to simulate a population of adult L2 learners. Our choice of the TO value 6,000 was justified in our preliminary simulations, which had shown that after encountering approximately 6,000 AS instances learners' L1 performance stabilized (although not completely, and this differed somewhat depending on the task). This way, $E_T = TO + E_B = 12{,}000$. Similarly, we simulated four more groups of early bilinguals ($TO = 0$, $E_T = E_B = 6{,}000$) with different R values (see Figure 4.5).

After every 500 input instances, learners' L2 proficiency was tested using the five tasks described in the previous section. Figure 4.6 shows the average performance curves for each of the four groups of adult learners.

First, we notice that in most tasks the performance curve flattens far below 100%. This is partly because all the tasks underestimate learners' L2 knowledge: while each test item assumes a single 'correct' answer, there may be more than one acceptable answer. When filling in verbs, for example, some empty slots may fit several semantically

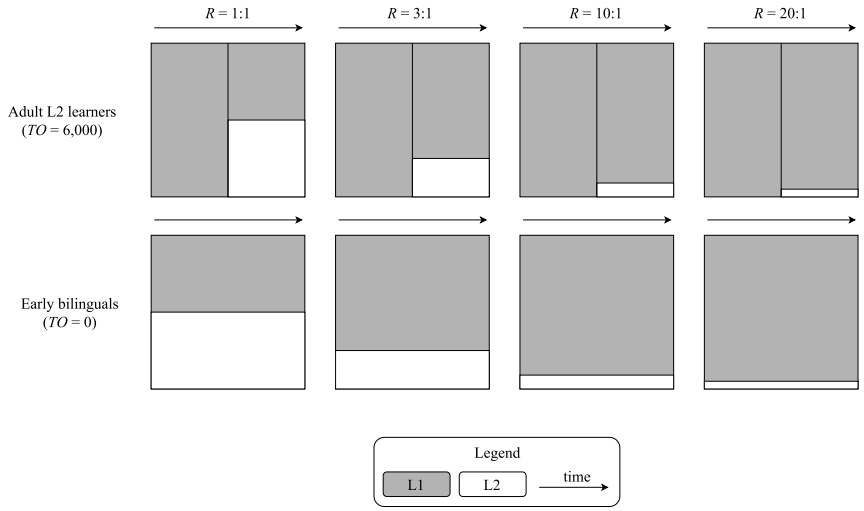


Figure 4.5: The setup of experiment 1. Two rows show the population types, four columns show the learner groups.

related verbs – synonyms (37) or antonyms (38).

(37) He acquired (bought) 300,000 shares of the stock.

(38) Industrial output fell (rose) 0.1% in September.

The size of the described effect is different for each task, which contributes to the different learners' performance across tasks (note that in Figure 4.6 the tasks are plotted on different scales). Additionally, there are certain differences between the model's performance in L2 German and L2 English tasks (compare the plots in Figure 4.6 pairwise). We explain this by possible differences in complexity between the German and English data sets, which we mentioned in subsection 4.2.3 above.

Despite the differences between the tasks, each individual plot in Figure 4.6 reveals the same pattern. Higher relative amount of L2 input corresponds to better L2 performance at each point in time. To statistically test whether the relative amount of L2 input correlated with the L2 performance at the end of learning, we ran Kendall's tau correlation tests[5] (see Table 4.3(a)). The results revealed a highly significant correlation between the amount of L2 input and the performance in each task in late learners, both for L2 English and L2 German. The results for early bilinguals yielded very similar

[5] Alternatively, we could compare the performance in the four groups (e.g., with an ANOVA or the Kruskall–Wallis test). However, this would require a further pairwise comparison of the groups, making the presentation of results less straightforward. Correlation tests are better in this respect, and their use is justified by our TO values being measured on a ratio scale. We use a non-parametric Kendall's tau test, to make no assumptions about the distributions of the performance values. Note that for data with only two groups (experiment 2) this test is equivalent to the Mann–Whitney U test, which is a non-parametric counterpart of the t-test.

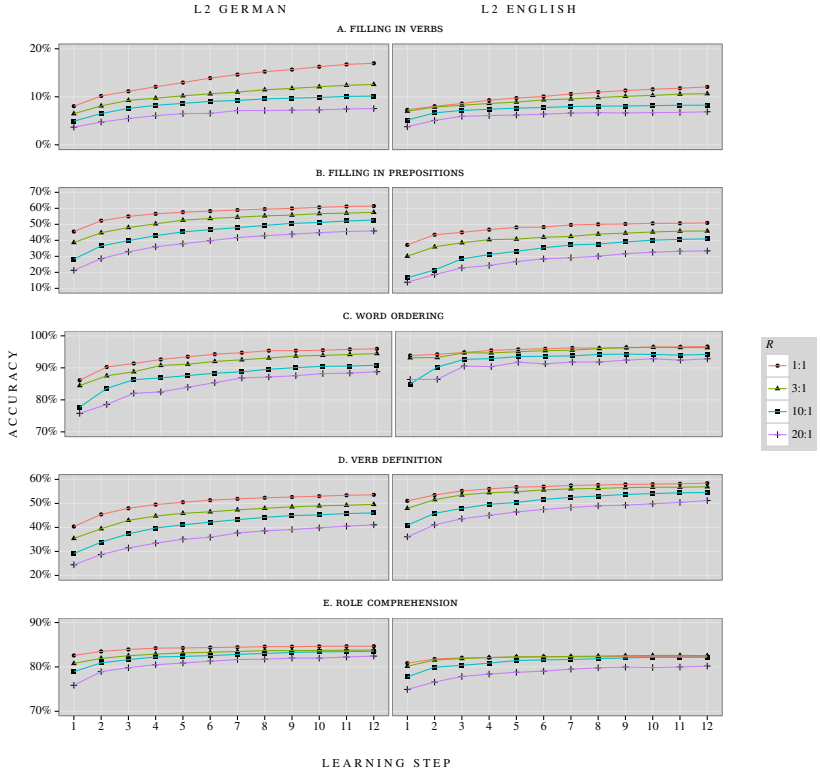


Figure 4.6: Average learning curves for adult learners with different R values, E_T is kept constant.

Table 4.3: Results of correlation tests between R and L2 performance at the end of learning, E_T is kept constant.

a. Simulated population of late L2 learners

	Task									
	Filling in verbs		Filling in prepositions		Word ordering		Verb definition		Role comprehension	
L2	τ	p	τ	p	τ	p	τ	p	τ	p
English	.69	$<$.001	.68	$<$.001	.51	$<$.001	.54	$<$.001	.30	$<$.001
German	.76	$<$.001	.73	$<$.001	.67	$<$.001	.72	$<$.001	.48	$<$.001

b. Simulated population of early bilingual learners

	Task									
	Filling in verbs		Filling in prepositions		Word ordering		Verb definition		Role comprehension	
L2	τ	p	τ	p	τ	p	τ	p	τ	p
English	.65	$<$.001	.65	$<$.001	.49	$<$.001	.59	$<$.001	.22	$<$.001
German	.69	$<$.001	.71	$<$.001	.63	$<$.001	.73	$<$.001	.33	$<$.001

patterns, thus we do not provide the plots of their learning curves, however the results of the correlation tests are shown in Table 4.3(b).

The results in Table 4.3 suggest that receiving more L2 input (in relation to L1 input) by a statistical learner leads to the better knowledge of L2 argument structure constructions. This may be due to the interaction of L2 input with the ongoing exposure to L1 input. However, so far we have assumed that learners' performance achieves its maximum at the end of learning simulations (upon receiving 6,000 mixed AS instances). This may be the case for the easier tasks, but the more difficult ones may take learners more time to achieve the highest possible performance, especially in case their cumulative E_{L2} is low because of a high R value (e.g., 20:1). For example, most learning curves for filling in verbs (see Figure 4.6(a)) do not flatten at step 12. Thus, it may be the case that learners in each group could potentially achieve the same performance, irrespective of the R value, if only they had enough time to learn. In this interpretation the L2 attainment depends not on the relative, but on the absolute amount of L2 input. To test whether this would be true, we ran another experiment.

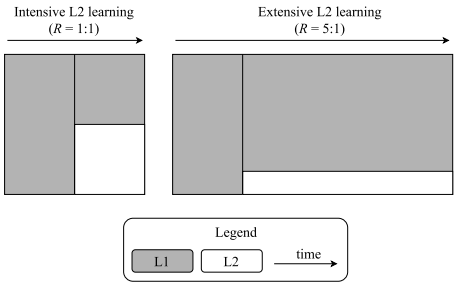


Figure 4.7: The setup of experiment 2.

Table 4.4: Results of correlation tests between R and L2 performance at the end of learning, E_{L2} is kept constant.

	Task									
	Filling in verbs		Filling in prepositions		Word ordering		Verb definition		Role comprehension	
L2	τ	p	τ	p	τ	p	τ	p	τ	p
English	−.25	.021*	−.03	.779	−.03	.750	−.07	.497	.01	.918
German	.15	.156	.12	.268	.01	.918	.08	.442	.18	.099

Experiment 2

The setup of this experiment was similar to that of experiment 1, however this time we kept the absolute amount of L2 input constant (E_{L2} = 1,500), while manipulating R. The latter was set to 1:1 (intensive L2 learning) or 5:1 (extensive L2 learning) – see Figure 4.7 (note that the length of L2 exposure is different in the two conditions, but the total L2 area is identical). Since the results of experiment 1 did not differ substantially for early bilinguals and adult learners, this time we simulated only the latter population by setting TO to 6,000.

If the relative amount of L2, indeed, determines the level of L2 attainment in a statistical learner, then we expect the performance to differ in the two groups. However, if it is only the absolute amount of L2 input that matters, there must be no difference in proficiency between the two conditions. The learning curves are shown in Figure 4.8.

Each individual plot in Figure 4.8 shows that the learner ultimately achieves the same or very similar performance in both conditions. In case of intensive learning, the curve is steep and reaches the highest level fast, while in the extensive condition learning goes much slower. The final performance is comparable, however: see the horizontal lines in Figure 4.8. Again, we ran Kendall's tau correlation tests using the final performance values. Table 4.4 shows the results of these tests.

The results show no significant correlations between R and learners' final performance for most tasks, the correlation reaching significance only for filling in verbs

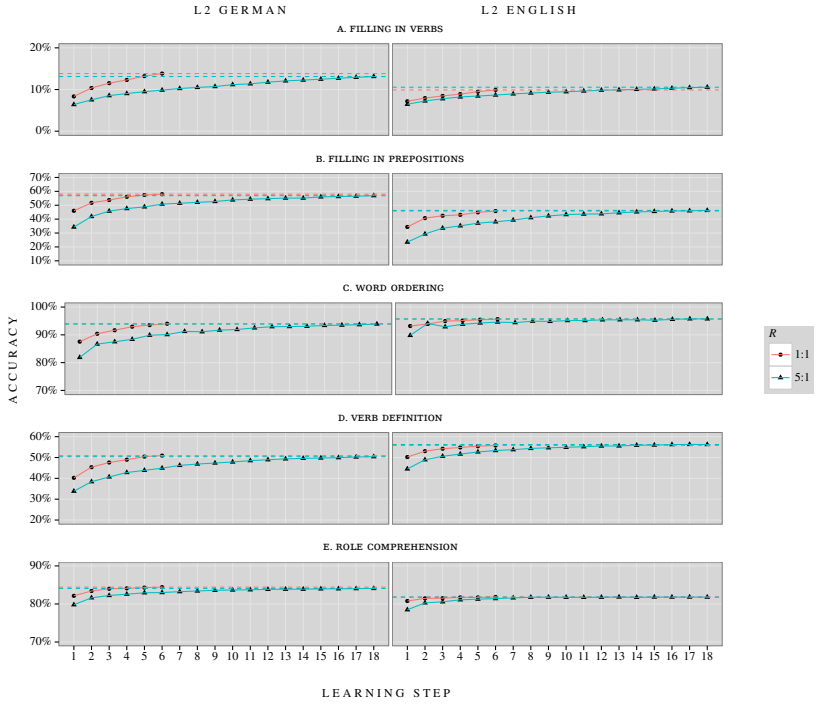


Figure 4.8: Average learning curves for learners with different R values, E_{L2} is kept constant.

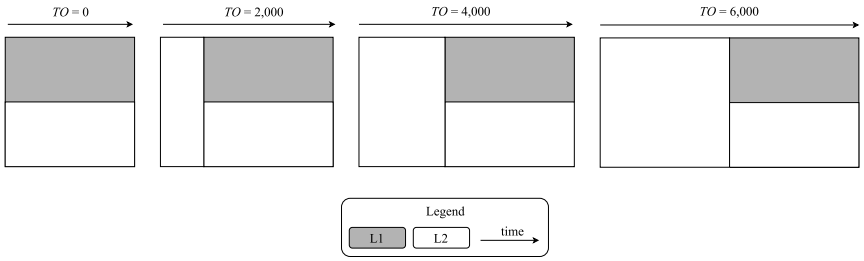


Figure 4.9: The setup of experiment 3.

in L2 English: $\tau = -.25$, $p = .021$. Since the correlation is negative, learners' performance in this task is higher in the extensive condition ($R = 5:1$) than in the intensive condition ($R = 1:1$). We believe this reflects learners' ongoing enhancement of L1 verbs. As we mentioned, filling in verbs is the most difficult task of the five, therefore continuing L1 exposure after TO aids learners in memorizing some contexts in which L1 verbs are used. As the extensive condition exposes learners to more L1 input than the intensive condition, they memorize more of these contexts, which helps them in discriminating between L1 and L2 contexts. As a result, at the end of learning in the extensive condition the model produces fewer L1 instances than in the intensive condition, hence the higher performance.

For the other tasks only the absolute amount of L2 input determines the resulting knowledge of L2 argument structure constructions. This suggests that length of exposure makes no difference, as long as the cumulative amount of L2 input stays the same.

4.3.2 Time of L2 onset

Experiment 3

This experiment was designed to investigate whether learners' L2 performance could be influenced by the time of L2 onset. If constructions and words are learned in a similar manner, then a negative effect of higher L1 entrenchment is to be expected (e.g., MacWhinney, 2012). Later L2 onset would lead to higher L1 entrenchment and, because of the interference this entails, lower L2 proficiency.

We manipulated the prior amount of L1 input by setting TO to 0 (simultaneous bilinguals), 2,000, 4,000 or 6,000 (late L2 learners). As we mentioned, in our preliminary simulations the maximum L1 performance was achieved only after approximately 6,000 AS instances, thereby we chose TO values under 6,000 to ensure that the level of L1 entrenchment is different for each TO. For all the four groups of learners, E_B was set to 6,000, and R was equal for all the groups (1:1), therefore E_{L2} amounted to 3,000 instances for each learner. The only difference between the groups, then, was the TO value. The experimental setup is shown in Figure 4.9, while Figure 4.10 illustrates the average learning curves for each group.

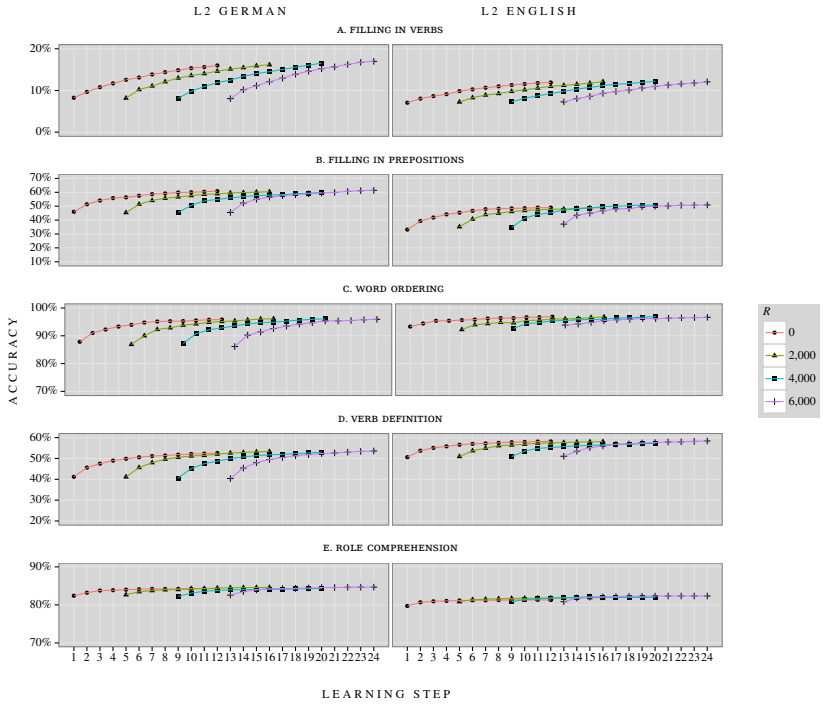


Figure 4.10: Average learning curves for learners with different TO values.

Table 4.5: Results of correlation tests between TO and L2 performance at the end of learning, E_{L2} is kept constant.

	Task									
	Filling in verbs		Filling in prepositions		Word ordering		Verb definition		Role comprehension	
L2	τ	p	τ	p	τ	p	τ	p	τ	p
English	.00	.961	.15	.034*	−.04	.595	.00	.944	.15	.029*
German	.14	.049*	.05	.440	.00	.959	.12	.086	.07	.277

If we look at each individual plot, we can notice no obvious pattern of difference between the four groups – in each case the learning curves seem to reach similar accuracy values. To statistically test whether learners' resulting performance at the end of learning correlated with TO, we ran Kendall's tau correlation tests (see Table 4.5).

The results suggest that the time of L2 onset does not affect the simulated learners' performance at the end of learning, with some exceptions. We do observe significant positive correlations between TO and the ultimate L2 performance for two tasks in L2 English (filling in prepositions and role comprehension), and a marginally significant correlation for filling in verbs in L2 German. Note that the correlations are positive, meaning that later TO leads to better L2 performance. This suggests a positive impact of cross-linguistic transfer from L1 to L2. English and German argument structures have a lot in common, as the two languages are typologically close: they both have SVO order in main clauses, and both are satellite-framed. Thus, the model may use the existing L1 knowledge to perform better in L2 tasks. The higher L1 entrenchment at TO is, therefore, beneficial, and may well give the model a small long-term advantage in L2 performance.

Most correlations in Table 4.5, however, are not significant. To ensure this is not caused by the similar degree of L1 entrenchment at TO in some groups (with $TO = 2,000$, $TO = 4,000$, and $TO = 6,000$), we compared the average L1 performance at TO in the three mentioned groups. Table 4.6 shows that L1 performance in the three groups differs in most tasks. The only deviation from this pattern is observed for role comprehension in L1 English, where the L1 performance of the three groups is approximately equal. This, in fact, makes our correlation result for role comprehension in L2 German non-informative, because the difference in ultimate L2 performance is not to be expected for the three groups with equal degree of L1 entrenchment at TO.

Before drawing any conclusions regarding the effect of the time of onset, we should additionally look at whether such effect is present at the earlier learning stages as well, since the presented correlation results are estimated for the learners' performance at the end of learning only. In addition, the correlation results do not tell us whether the time of onset interacts in any way with learners' cumulative amount of L2 exposure. To test this, we ran a series of regression models that predicted learners' performance at each learning stage.

Table 4.6: Average L1 performance at TO in different learner groups in experiment 3.

		TO			
L1	Task	0	2,000	4,000	6,000
English	Filling in verbs	–	10.20%	12.90%	13.20%
	Filling in prepositions	–	47.20%	48.30%	49.70%
	Word ordering	–	96.10%	96.70%	97.20%
	Verb definition	–	56.70%	58.10%	57.50%
	Role comprehension	–	82.00%	82.00%	81.80%
German	Filling in verbs	–	13.30%	17.50%	19.90%
	Filling in prepositions	–	57.30%	61.60%	62.20%
	Word ordering	–	94.30%	95.90%	97.40%
	Verb definition	–	51.40%	53.30%	54.80%
	Role comprehension	–	84.20%	84.50%	84.90%

4.3.3 L2 performance: contributions of individual factors

Regression models were used to examine the potential effects of TO, E_{L2}, and their interaction. Conceptually speaking, we checked whether at any learning stage learners' L2 performance in a certain task could be predicted by TO and E_{L2}. We ran ten linear mixed-effects models (Baayen, 2008), one for each task in each language, using the *lme4* package for *R* (D. Bates et al., 2015). To account for possible individual variation between learners, we introduced a random factor of learner. Each model had the maximal random effect structure justified by the data sample (Barr, Levy, Scheepers, & Tily, 2013), slightly varying for different tasks and languages due to convergence issues.

All the models were run on the learning results reported on for experiment 3. Recall that in experiment 3 we manipulated TO, but not E_{L2}. Nevertheless, the latter was present in the learning results of our simulations, because we tested the model's performance at different learning stages (that is, after it was exposed to different amounts of L2). Therefore, each performance score had an E_{L2} value associated with it, which we used in the regression. This setup implies that the regression models do not only provide results in terms of ultimate L2 proficiency (as did the correlation tests reported in the previous sections), but at each moment of learning. Importantly, L2 performance is not a linear function of E_{L2} in our experiments (recall the shapes of the learning curves). In general, learning success is believed to be a power function of experience (Newell & Rosenbloom, 1981). To account for this relation between performance and E_{L2}, we log-transformed all the performance values and E_{L2}, but also TO for consistency.[6] To eliminate the problems of multicollinearity and variance

[6] Additionally, we fitted the same models to the data with only two variables log-transformed (perfor-

inflation, and to make the regression coefficients directly comparable, we standardized all the variables. A summary of the models is given in Table 4.7.

L2 amount

The effect of E_{L2} is the only main effect observed for all the tasks in both German and English (see the dark gray cells in Table 4.7. As expected, the effect is always positive: learners' L2 proficiency increases as they are being exposed to more L2 input. This supports the correlation between E_{L2} and learners' L2 performance, found in experiment 1. Note that the standardized regression coefficients (β) for E_{L2} have the largest values, compared to the coefficients of E_{L2} and $TO \times E_{L2}$ in each regression model, which means that the effect of E_{L2} is stronger than that of TO and of the interaction. The only exception is role comprehension in L2 English, for which the coefficient of E_{L2} (0.20) is smaller than that of TO (0.22). Yet, the amount of variance explained by the fixed effects (R^2_m) in the respective regression model is the smallest ($R^2_m = .09$, or 9%), compared to the respective value in all the other models (e.g., $R^2_m = .67$ for verb definition in L2 German). The poor model fit suggests that the β coefficients in the regression model for role comprehension in L2 English might not be informative.

L2 onset

The main effect of TO is present only for L2 English and only for two tasks: filling in prepositions and role comprehension. This is comparable to the results of experiment 3, in which the correlation of TO with learners' final L2 performance was observed for the same two tasks in L2 English. Additionally, in experiment 3 the same positive correlation was observed for a single task in L2 German (filling in verbs), but this was only marginally significant and is not supported by the regression results. As for the other two tasks with a main effect of TO, the analysis for role comprehension, as we mentioned, is not informative due to the poor model fit. This is not the case, however, for filling in prepositions. The impact of TO is positive: late L2 starters perform better than early L2 starters. This could be explained by the positive effect of cross-linguistic transfer. As we mentioned, the model may use the existing L1 knowledge to perform better in L2 tasks, and the higher level of L1 entrenchment is beneficial, especially at the early stages of L2 learning. Indeed, although the effect of transfer can be both positive and negative, the positive effect must prevail here due to the similarity of English and German argument structure constructions. However, the effect can be manifested differently in each of the five tasks used, due to their nature. Since the two languages in our model use shared representations of lexical semantics, participant roles, and word order, in such tasks as verb definition, role comprehension and word ordering, one would expect a positive transfer effect. For example, a simulated learner of L2 English may be able to describe the meaning of a novel English verb *to increase*, because it shares many contexts of use with its German translation *steigen*. This is different for the other two tasks – filling in verbs and prepositions. Since learners are

mance and E_{L2}), and with original non-transformed variables, and they yielded consistent results.

Table 4.7: Summary of mixed-effects models predicting learners' L2 performance.

a. L2 German

Model	Predictor	β	*SE*	95%*CI*
Filling in verbs	TO	0.06	0.05	$[-0.05, 0.16]$
$R^2_m = .66^{\dagger}$	E_{L2}	0.81	0.02	$[0.77, 0.84]$
$R^2_c = .95$	$TO \times E_{L2}$	0.02	0.02	$[-0.01, 0.06]$
Filling in prepositions	TO	0.00	0.05	$[-0.11, 0.10]$
$R^2_m = .58$	E_{L2}	0.76	0.02	$[0.73, 0.79]$
$R^2_c = .89$	$TO \times E_{L2}$	0.00	0.02	$[-0.04, 0.03]$
Word ordering	TO	-0.04	0.05	$[-0.13, 0.05]$
$R^2_m = .61$	E_{L2}	0.78	0.02	$[0.73, 0.83]$
$R^2_c = .84$	$TO \times E_{L2}$	0.05	0.02	$[0.00, 0.09]$
Verb definition	TO	0.04	0.05	$[-0.05, 0.13]$
$R^2_m = .67$	E_{L2}	0.81	0.01	$[0.79, 0.84]$
$R^2_c = .92$	$TO \times E_{L2}$	0.04	0.02	$[0.01, 0.07]$
Role comprehension	TO	0.05	0.08	$[-0.11, 0.21]$
$R^2_m = .21$	E_{L2}	0.46	0.02	$[0.41, 0.50]$
$R^2_c = .91$	$TO \times E_{L2}$	0.02	0.02	$[-0.02, 0.06]$

b. L2 English

Model	Predictor	β	SE	95% CI
Filling in verbs	TO	0.02	0.07	$[-0.11, 0.15]$
$R^2_m = .50$	E_{L2}	0.71	0.02	$[0.67, 0.74]$
$R^2_c = .95$	$TO \times E_{L2}$	-0.01	0.02	$[-0.05, 0.03]$
Filling in prepositions	TO	0.12	0.05	$[0.01, 0.22]$
$R^2_m = .48$	E_{L2}	0.68	0.02	$[0.64, 0.73]$
$R^2_c = .86$	$TO \times E_{L2}$	-0.05	0.02	$[-0.10, -0.01]$
Word ordering	TO	-0.05	0.06	$[-0.16, 0.07]$
$R^2_m = .28$	E_{L2}	0.53	0.03	$[0.47, 0.59]$
$R^2_c = .74$	$TO \times E_{L2}$	0.01	0.03	$[-0.04, 0.08]$
Verb definition	TO	-0.02	0.08	$[-0.17, 0.13]$
$R^2_m = .32$	E_{L2}	0.57	0.02	$[0.53, 0.60]$
$R^2_c = .94$	$TO \times E_{L2}$	-0.02	0.02	$[-0.05, 0.01]$
Role comprehension	TO	0.22	0.09	$[0.04, 0.41]$
$R^2_m = .09$	E_{L2}	0.20	0.02	$[0.16, 0.24]$
$R^2_c = .95$	$TO \times E_{L2}$	-0.02	0.02	$[-0.05, 0.02]$

† R^2_m and $R2_c$ stand for marginal and conditional R^2 coefficients and indicate the amount of variance explained by the fixed factors and by the full model, respectively (Johnson, 2014). The reported SE and confidence interval values are estimated via parametric bootstrap with 1,000 resamples (D. Bates, Mächler, Bolker, & Walker, 2015).

allowed to use their L1 in the two "fill-in-the-blank" tasks, they are likely to produce L1 verbs and prepositions (which are different in German and English), hence the negative effect of transfer.[7] Note, however, that both German and English have a preposition *in*, often used in equal or very similar contexts. In our German data set, *in* is the most frequent preposition, which promotes its use by L1 German speakers during the testing in L2 English. Although the learners, in fact, use the German preposition, it may fit many English test instances that require the use of English *in*, hence the positive effect of lexical transfer from German to English. The same effect from English to German may not be observed, since in our English data set *in* is only the third most frequent preposition. Therefore, learners would more likely use the two more frequent prepositions (*to* and *on*) during the testing.

Interaction term

First we note that the interaction effect of E_{L2} and TO is significant for filling in prepositions in L2 English, with a negative β coefficient. Considering the positive effect of TO in this task we just discussed, this negative interaction can be interpreted as a decrease in the positive TO effect at the later stages of L2 testing. This supports our explanation of the positive TO effect in terms of positive transfer: higher L1 entrenchment is beneficial at the early stages of L2 learning, however at the later stages this benefit diminishes, because learners rely more on their acquired L2 knowledge than on L1 knowledge.

Finally, there is a significant interaction effect in verb definition in L2 German. The respective β coefficient is positive – that is, the positive effect of higher E_{L2} on learners' performance is stronger for learners with later TO. In other words, in this task late L2 starters achieve a certain level of performance faster than early L2 starters. This observation also suggests that transfer has more positive than negative effect in verb definition in L2 German.

4.4 Discussion

In the present study we investigated how the learning of argument structure constructions in L2 was affected by two variables – the amount of L2 input (both relative and absolute) and the time of L2 onset. For this purpose, we computationally simulated the process of statistical construction learning in two languages and ran three experiments to test the performance of simulated learners under different conditions of exposure.

4.4.1 Amount of L2 input

The first variable, the amount of L2 input, affected learners' L2 performance as expected – getting more L2 input resulted in better L2 performance. This is in line with a general learning rule "the more, the better", which has been demonstrated to apply to human

[7] This is a rather broad understanding of cross-linguistic transfer, as it covers not only subconscious cross-linguistic influence, but also the use of L1 instead of L2.

learners for various domains (e.g., Muñoz, 2011; Flege, Yeni-Komshian, & Liu, 1999). In experiment 1, we captured this type of relation using a relative measure of L2 amount, while controlling for the length of L2 exposure. However, when the cumulative amount of L2 was kept constant instead (experiment 2), the model's performance appeared to be the same for varying relative amounts of L2. Intuitively, this is contrary to a well-researched spacing effect: spaced, or distributed, practice leads to higher test performance than massed practice in many domains (Küpper-Tetzel, 2014), including construction learning (Ambridge, Theakston, Lieven, & Tomasello, 2006). However, it has been argued (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006) that the learning depends not only on the length of the interstudy interval (the time between two presentations of an item), but also on that of the retention interval (the time between its last presentation and the test). Thus, simulations with a systematic control of the two intervals (with respect to the presentation of individual L2 instances) are needed to relate our findings to the existing research in this domain.

In the current study we focused only on the quantitative characteristics of L2 input, but the quality of L2 input may be equally important (Moyer, 2005). Obviously, it cannot be the mere amount of input that determines learners' L2 proficiency, as an identical amount of input may be very different for two different learners, in terms of relevance for the learner, grammatical complexity, lexical diversity, native-likeness, discourse style, etc. All these characteristics contribute to learners' level of engagement with the target language and affect the learning process. Therefore, an ideal measure of L2 input should account for much more than its overall amount. Preliminary versions of such measures have already been proposed, but they need further refinement. For example, Ågren, Granfeldt, and Thomas (2014) have developed an individual input profile score, yet they recognize it does not take into account that different input domains may affect the learning to a different degree.

4.4.2 Time of L2 onset

The second variable that we investigated – the time of L2 onset – appeared not to have any impact on performance in most L2 tasks. The only exceptions were two tasks in L2 English – filling in prepositions and role comprehension, where later L2 starters performed better than early starters. The latter exception, as we showed, could be due to the poor fit of the respective regression model. As for filling in prepositions, later L2 starters had a better knowledge of a frequent German preposition *in*, and they could transfer this knowledge into L2 to identify the correct contexts of use of the English preposition *in*. Overall, unlike in other linguistic domains such as lexis and morphology (Monner et al., 2013; Zhao & Li, 2010), a pronounced negative effect of L1 entrenchment (i.e., later L2 onset) on learning L2 argument structure constructions is absent in our experiments. The difference between the domains relates to a discussion in literature on L1 processing or, more broadly, on the age/order effect. It has been shown (Lambon Ralph & Ehsan, 2006) that the negative effect of a later acquisition of a specific item (e.g., word) in cued production is higher for stimuli with more arbitrary cue–outcome mappings (e.g., word phonology and meaning), and lower for stimuli with more consistent mappings (e.g., word phonology and orthography). In case of

arbitrary mappings, the meaning of a novel word can hardly be predicted from its phonological form, despite a potentially large number of earlier acquired mappings. On the contrary, word orthography is often predictable from its phonological form, due to the consistency of the mapping with earlier acquired words. In the context of bilingual learning we look at the consistency of mappings across L1 and L2, rather than across multiple L1 items. In our test tasks each cue (i.e., test AS instance) consisted of multiple features, and the model, in fact, could predict the outcome (i.e., the value of the missing feature) based on the mappings between the features in L1: the languages we used in this study – German and English – were typologically close, and positive transfer was likely to take place. This could be the reason why the negative effect of the late onset was not observed.

In the light of the ongoing discussion about the age/order effect in literature, we can further note that our results do not support the idea proposed by Stewart and A. W. Ellis (2008) that the age/order effect is a property of any learning system. Instead, our findings are consistent with the cumulative frequency hypothesis (Zevin & Seidenberg, 2002; M. B. Lewis et al., 2001), which claims that the accessibility of a word is determined by its cumulative frequency, but not the moment of its first encounter.

Due to the lack of available annotated resources we only used English and German in the current study. We plan to explore new resources and investigate the bilingual learning of argument structure constructions in additional language pairs, to determine the exact contribution of cross-linguistic transfer effects to such learning. The computational tool used for our study focuses on only a subset of (input-related) factors and is not meant to represent the whole picture of how humans learn a second language. Nevertheless, it has provided rather robust and consistent results by allowing for full control of the variable confounding and of the input quantities, which cannot be easily done in human subject studies. These advantages make the presented model a promising tool for future studies.

CHAPTER 5

Quantifying cross-linguistic influence with a computational model: A study of case-marking comprehension[1]

5.1 Introduction

5.1.1 Quantifying cross-linguistic influence

The phenomenon of cross-linguistic influence (CLI) is central to our understanding of bilingual and second language (L2) learning. Languages interact in the bilingual mind, and studies of CLI intend to describe various types of such interaction.[2] One challenging issue that has long interested scholars is measuring the amount of CLI – that is, quantifying the extent to which linguistic representations from one language affect the use of the other language(s). Weinreich (1968) suggested that "no easy way of measuring or characterizing the total impact of one language on another in the speech of bilinguals has been, or probably can be, devised" (p. 63). Measuring the amount of CLI is important to understand to what extent the knowledge of one language is beneficial (in case of positive CLI) or damaging (in case of negative CLI) for the acquisition of other languages.

One common method to measure CLI is through the so-called error analysis: scholars look at the frequency of linguistic errors in a group of learners with a particular first language (L1) background, and estimate the contribution of negative CLI to the non-native L2 use (Born, 1985; Grauberg, 1971; see Palmberg, 1976 for a relevant bibliography). At the same time, CLI is not the only source of non-native language use:

[1] This chapter is based on the article of the same name submitted for publication in a journal.

[2] We adopt a broad cognitive view on CLI (Jarvis & Pavlenko, 2008), which covers manifestations of CLI both in L2 acquisition and in bilingual language use.

other factors such as overgeneralization may play a role, and the non-native use is often caused by a combination of factors (Jordens, 1977). This is why the exact methodology for identifying CLI is not straightforward: it has been argued that one needs to show that the learners within a particular group make similar mistakes, that the mistakes are different across the L1 groups, and that the mistakes have their linguistic equivalents in the learners' L1 (Jarvis, 2000). Given the multitude of interfering variables (e.g., proficiency, learning history, aptitude), it is difficult to identify with confidence all cases of the CLI influence, and to measure the amount of CLI using this method. The same problem persists in more controlled experimental settings, which employ linguistic tasks related to language production or comprehension by bilingual learners (Grosjean, 1998). The number of interfering variables can be reduced in research on multilingual speakers: studying learners' third language use allows for identifying the instances of L1 and L2 influence at individual level (e.g., De Angelis & Selinker, 2001), similar to a within-subject design in experimental studies, but this "individual" approach makes it difficult to generalize over the group of learners.

Another issue related to the described methodologies is that the resulting CLI measures are grounded in language use. This may constitute a methodological challenge whenever such measures are used to *predict* the learner's language use, leading to circular reasoning.

These limitations can be overcome in cognitive computational models of bilingual language learning and use, which allow researchers to look inside the "black box" of linguistic representations. While no computational modeling studies focused on measuring CLI, some of such studies in the field of bilingualism employed quantitative measures that reflected the amount of CLI in the respective models. In particular, Zhao and Li (2010) simulated bilingual acquisition of Chinese and English words using a self-organizing neural network model. The learning process in each simulation yielded a spatial representation (map) of the bilingual lexicon. To explain how their computational model arrived at a particular type of map, the authors computed the average Euclidean distance between lexical translation equivalents in multiple pairs: that is, how far an English word (e.g., *star*) is located from its Chinese equivalent (*Xing1xing*) on the map. A shorter average distance means that many translation equivalents are located next to each other, which is the evidence of high CLI: the location of L1 lexemes has influenced the placement of the corresponding L2 lexemes. Vice versa: a longer average distance corresponds to smaller amount of CLI, because the location of L1 lexemes has not played the determining role in the placement of their L2 equivalents.

In a similar type of model, Shook and Marian (2013) studied bilingual speech comprehension in English and Spanish. They employed an online measure, so-called language activation score. This measure showed how strongly the lexical representations from a particular language (e.g., Spanish) were activated on average, when the model was given a word in either the same or a different language (English). One can argue that the activation score for the non-target language reflects the amount of CLI.

The described measures and the respective models, however, do not go beyond the lexeme level, while there are no computational modeling studies of CLI at the level of abstract constructions. To address this gap in the literature, in this study we use a computational model of learning argument structure constructions from bilingual input.

We choose this model, because it has been used for simulating bilingual learning of argument structure constructions (see chapters 3–4), and it allows for measuring the amount of CLI in this domain. Our goal in the present study is to demonstrate how the amount of CLI can be measured in the learning and use of such constructions, and how a CLI measure can be used to explain the patterns of language use observed in the model. More specifically, we study the acquisition and interpretation of case-marking cues in Russian and German transitive sentences: as we show below, this is one of the aspects discussed in the literature, and the relevant experimental results from human participants are available.

5.1.2 Interpretation of transitive sentences

In some languages, such as English, French, Hebrew, etc., transitive sentences are characterized by a fixed subject-verb-object (SVO) word order (39). In other languages, the word order is more flexible: German transitive sentences can have SVO (40) as well as OVS word order (41).

(39) The dog chases the bear.[3]

(40) Der Hund-∅ jägt den Bär-en.
ART.M.NOM.SG dog-M.NOM chase:3SG ART.M.ACC.SG bear-M.ACC
'The dog chases the bear.'

(41) Den Bär-en jägt der Hund-∅.
ART.M.ACC.SG bear-M.ACC chase:3SG ART.M.NOM.SG dog-M.NOM
'The dog chases the bear.'

To correctly interpret OVS sentences, speakers rely on other cues than the word order: morphological case marking (as in 41), but also animacy, noun–verb agreement, etc. However, learners of a language allowing for OVS sentences may rely on the word order cue and misinterpret participant roles in such sentences; this happens both in adult L2 learners (e.g., Isabelli, 2008; Kempe & MacWhinney, 1998; VanPatten, 1996) and in monolingual children learning various languages (e.g., Smolík, 2015; Kim, O'Grady, & Cho, 1995; Schaner-Wolles, 1989). Speaking of young monolingual German children, it has been suggested that they start by acquiring the more prototypical and more frequent SVO form first (Dittmar, Abbot-Smith, Lieven, & Tomasello, 2008). The situation with bilingual and L2 learners is more complex, because CLI may be at play. There are two general views on the role of CLI in the misinterpretation of transitive sentences.

1. The first view is represented by the First-Noun Principle (e.g., VanPatten, 2012, 1996). According to this principle, learners universally tend to assign the agent role to the first noun or pronoun in a given sentence, while the effect of CLI is negligible. Existing studies have argued that the First-Noun Principle can explain data from L2 learners of various languages: English, French, German, etc. (see an overview by Lee & Malovrh, 2009).

[3] Example from Yoshimura and MacWhinney (2010).

2. The alternative view explains the misinterpretation of OVS sentences by CLI from learners' L1. Under this view, L2 learners adhere to the interpretation strategy which is standard in their L1: if learners do not encounter OVS sentences in their L1, they will misinterpret such L2 sentences as SVO. This general view is compatible with multiple acquisition theories (see an overview by Hanson, Aroline, & Carlson, 2014), but the two accounts mentioned most frequently in this respect are the Unified Competition Model (MacWhinney, 2012) and the L1 Transfer Principle (VanPatten, 2015b).

    2.1. According to the Competition Model (Morett & MacWhinney, 2013; Kempe & MacWhinney, 1998; Mimica, Sullivan, & Smith, 1994; Gass, 1987; Kilborn & Cooreman, 1987; McDonald, 1987, etc.), learners of both L1 and L2 attend to multiple cues in the input, such as word order, case marking, animacy, etc. Importantly, languages differ in the relative importance of various cues (e.g., case marking plays little role in English), and L1 speakers learn to attend to some cues more than to others. These attentional preferences, or cue strengths, are acquired based on the *validity* of the cues. Validity can be calculated using a linguistic corpus, as a product of two other values: cue *availability* and *reliability*. The cue is available whenever it is present as a marker of a particular function: e.g., the nominal case marking of the subject may help discriminating between this subject and the object in the sentence. A cue is reliable whenever its presence ensures the right choice of the function: e.g., the nominal case marking of the object would make the cue unreliable for this sentence. The acquired cue strengths are initially transferred to an L2. As a result, when L1 speakers of a language with fixed SVO word order (e.g., English) start learning an L2 in which OVS sentences are allowed (e.g., German), they fail to attend to case marking and misinterpret OVS sentences as SVO.

    2.2. The L1 Transfer Principle complements the First-Noun Principle mentioned above. Given the combination of the two, learners still tend to interpret the first noun as the agent of a sentence, yet this general strategy is modulated by their L1 knowledge. As an example, Isabelli (2008) demonstrated that L1 Italian students learning L2 Spanish could interpret Spanish OVS sentences better than their L1 English peers. This is because OVS sentences are common in Italian and Spanish, but not in English. Note, however, that the lexical similarity between Italian and Spanish might be a factor in this example (VanPatten, 2015a) – we return to this issue in the general discussion.

To summarize, there is no conclusive evidence about the role of CLI in the interpretation of case-marking cues in transitive sentences. To investigate whether CLI is at play, we simulate an experimental task employed in the two target studies described below, and quantify the impact of CLI in the model's language use with a novel quantitative measure.

The rest of the chapter is organized as follows. First, we briefly introduce two studies on which we focus in our simulations. These studies investigate the interpretation of transitive sentences with case-marking cues by learners whose L1 does not employ such cues. This is followed by the presentation of our computational model, where we also explain how it allows for quantifying CLI. Next, in two sets of simulations we demonstrate that the model's linguistic behavior in the target task is similar to that observed in human learners. The findings are explained in terms of the amount of CLI. Finally, we make two novel predictions on how the model would perform on the same task when trained on different language pairs, and test them using our computational model. Overall, this gives four sets of simulations:

1. Interpretation of German sentences by bilingual learners whose other language has no case marking (Janssen, Meir, Baker, & Armon-Lotem, 2015).

2. Interpretation of German and Russian sentences by L2 learners whose L1 has no case marking (Kempe & MacWhinney, 1998).

3. Interpretation of German sentences by Russian–German bilingual learners (novel).

4. Interpretation of Russian sentences by bilingual learners with various additional languages (novel).

5.2 Target studies on case-marking comprehension

Studies on the interpretation of case-marking cues in transitive sentences have mainly focused on adult L2 acquisition (Morett & MacWhinney, 2013; Kempe & MacWhinney, 1998; Mimica et al., 1994; McDonald, 1987, etc.), while similar studies with early bilinguals have been rare (but see Janssen et al., 2015; O'Shannessy, 2011). We focus on one study from each population: a study with bilingual and monolingual Russian children by Janssen et al. (2015), and a study with adult learners of Russian and German (Kempe & MacWhinney, 1998). In the following sections we explain why we choose these two studies. First, however, we describe a picture-choice task employed in both of them.

5.2.1 Picture-choice task

In this task, participants hear a sentence and see two pictures containing alternative interpretations of the sentence. The participants have to choose the picture which in their opinion corresponds to the correct interpretation of the sentence. In the two target studies, the picture-choice task is employed to study the comprehension of competing cues, in particular case-marking and word order. The target sentences include two nouns (nominative and accusative/dative) and a verb, and the two pictures depict the same event, but the participant roles are swapped in one of the pictures. An example from Janssen et al. (2015):

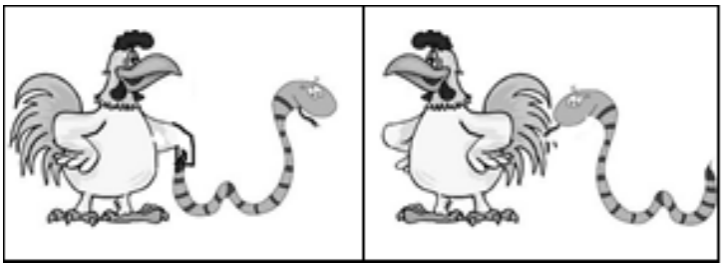


Figure 5.1: Accompanying pictures in the study of Janssen, Meir, Baker, and Armon-Lotem (2015). Reprinted from "On-line comprehension of Russian case cues in monolingual Russian and bilingual Russian-Dutch and Russian-Hebrew children", 2015, by B. Janssen, N. Meir, A. Baker, and S. Armon-Lotem. In E. Grillo & K. Jepson (Eds.), *Proceedings of the 39th Annual Boston University Conference on Language Development*, p. 272. Copyright 2015 by B. Janssen. Reprinted with permission.

(42) Petuh-∅ trogaet zmey-u.
rooster-M.NOM touch:3SG snake-F.ACC
'The rooster touches the snake.'

The sentence (42) is accompanied by two pictures (Figure 5.1), depicting either a rooster touching a snake, or a snake touching a rooster.

5.2.2 Bilingual and monolingual Russian children

Janssen et al. (2015) work with Russian monolingual children, as well as with Russian–Dutch and Russian–Hebrew bilingual children. While Russian is characterized by a free word order and systematic case marking of nouns, the opposite holds for Dutch and Hebrew: these two languages have much stricter word orders and no morphological cases on nouns. The case-marking cue is important in Russian: it marks the thematic roles of the nouns. At the same time, in Dutch and Hebrew the word order is often the only cue that allows to distinguish between SVO and OVS sentences.

In this study, the picture-choice task is employed to investigate whether this difference between Russian and Dutch/Hebrew leads to any differences in sentence interpretation by Russian monolingual and Russian–Dutch or Russian–Hebrew bilingual children. Some of the presented sentences had SVO order, where the word order cue and the case-marking cue supported and complemented each other (the converging cue condition), as in (42) above. Other sentences had OVS word order with the conflicting cues (the conflicting cue condition), such as (43):

(43) Zhiraf-a vidit petuh-∅.
giraffe-M.ACC see:3SG rooster-M.NOM
'The rooster sees the giraffe.'

In addition to SVO and OVS sentences with a subject and a direct object, noun-verb-noun sentences with an indirect dative object were used, such as (44):

(44) Zmey-e ulybayetsa zhiraf-∅.
snake-F.DAT smiles.at:3SG giraffe-M.NOM
'The giraffe smiles at the snake.'

There were 40 stimuli overall: 20 SVO sentences and 20 OVS sentences, the test verbs included *lubit* ('love'), *trogat* ('touch'), *tselovat* ('kiss'), *ulybatsa* ('smile'), *vidyet* ('see'), and *zvonit* ('call').

Both monolingual and bilingual children were expected to perform high in the comprehension of the SVO sentences, but the bilingual children in the conflicting cue condition were predicted to demonstrate a lower accuracy rate and longer reaction time than in the converging cue condition, and than the monolingual children in the conflicting cue condition. This is because the bilingual children may transfer the strength of the word order cue from Dutch or Hebrew into Russian, leading to the misinterpretation of the Russian OVS sentences as SVO. These predictions were met in terms of both accuracy and reaction time. Interestingly, no differences between the two bilingual groups were observed, despite the high variation reported for home language use: 61.1% in the Hebrew group, and 16.7% in the Dutch group.

While the mentioned monolingual and bilingual groups were age-matched, an additional group of younger Russian learners took part in the experiment, and its performance was lower than that of the age-matched Russian group.

For us, this study presents an interesting case: first, the authors mention that their results are compatible with both the First-Noun Principle and the Competition Model. Second, this is one of the only two studies on the interpretation of case-marking cues focusing on early bilingual learning. The other one dealt with rare languages for which it was difficult to obtain the relevant data – Lajamanu Warlpiri and Light Warlpiri (O'Shannessy, 2011).

5.2.3 Adult L2 learners of Russian and German

Kempe and MacWhinney (1998) worked with native English adult learners of L2 Russian and L2 German, who had been exposed to the target languages in classroom for 25–26 months. The picture-choice task with transitive sentences was used. Both in Russian and in German, all the sentences had the verb *look for/find* as the predicate: *iskat* in Russian, and *suchen* in German. The picture-choice task was slightly different in this experiment: the alternative pictures did not depict the full event, but only the two participants instead, and the learners had to decide which participant was the agent, defined as "who or what did the looking or finding" (Kempe & MacWhinney, 1998, p. 557). The 32 Russian and German test sentences were mutual translations of each other: 12 SVO sentences with case-marking, 12 OVS sentences with case-marking, and 8 SVO sentences fully neutralized in terms of their case-marking cues: these contained two nouns whose nominative and accusative cases were marked with the same morpheme, as in (45).

(45) Die Tochter-∅ sucht
ART.F.SG.NOM/ACC daughter-SG.NOM/ACC look.for:3SG
die Mutter-∅.
ART.F.SG.NOM/ACC mother-SG.NOM/ACC
'The daughter looks for the mother.'

Using the methodology commonly adopted in Competition Model studies, Kempe and MacWhinney (1998) compute the availability of a cue as the number of sentences in which the cue is present divided by the total number of transitive sentences. To compute the reliability of a cue, they divide the number of sentences in which the cue correctly indicates the agent by the total number of sentences in which this cue is present. Based on their calculations, Kempe and MacWhinney (1998) show that the case-marking cue in Russian has a higher validity than in German, and this is why Russian L2 learners are more successful in the acquisition of case marking than German L2 learners: they perform the task faster (in terms of decision latencies) and more accurately than German L2 learners.

We choose this study because of its similarity to the study of Janssen et al. (2015): both employ the picture-choice task, and both focus on the comprehension of case marking in Russian. These similarities will help us to make some informed predictions about the interpretation of case-marking cues, and test these predictions with our model. The main difference between the two experiments is the age of the subjects, which we can also take into account in our computational simulations by manipulating the overall amount of input the model is exposed to.

5.3 Computational model

The computational model we employ here is a novel version of the model used in earlier studies on monolingual and bilingual acquisition of argument structure constructions (Alishahi & Stevenson, 2010, 2008). Compared to the previous studies, here the model has been adapted to languages with free word order, as explained below. The model learns argument structure constructions from the input data.

5.3.1 Input to the model

Input representations

The input to the model consists of individual verb usages, which we call argument structure (AS) instances. Each AS instance comprises multiple independent features: lexical, semantic, and syntactic. Further, we make two important distinctions: between distributional features (FD) and symbolic features (FS), and between global (FG) and local features (FL). Consider an example instance in Table 5.1. Symbolic features carry values expressed by a single symbol (e.g., head predicate: *touch*; number of arguments: 2). In contrast, each value of a distributional feature is a set of elements (e.g., head properties: {ACTION, CAUSAL, MANIPULATE, PHYSICAL}). As for the global vs. local features, the former relate to the utterance or the described event as a whole (e.g., the

Table 5.1: An AS instance for the sentence *The snake touches the rooster.*

Feature	Type	Value
Head predicate	Global, symbolic	*touch*
Head properties	Global, distributional	{ACTION, CAUSAL, MANIPULATE, PHYSICAL}
Head position	Global, symbolic	2
Number of arguments	Global, symbolic	2
Arg.1	Local, symbolic	*snake*
Arg.2	Local, symbolic	*rooster*
Arg.1 case	Local, distributional	NOM
Arg.2 case	Local, distributional	GEN, ACC
Arg.1 lexical meaning	Local, distributional	{DIAPSIDE, REPTILE, ..., CAUSAL AGENT}
Arg.2 lexical meaning	Local, distributional	{CHICKEN, DOMESTIC FOWL, ..., CAUSAL AGENT}
Arg.1 role properties	Local, distributional	{ACTING, ANIMATE, ..., VOLITIONAL}
Arg.2 role properties	Local, distributional	{ANIMATE, CONCRETE, ..., TOUCHED}
Arg.1 preposition	Local, symbolic	N/A
Arg.2 preposition	Local, symbolic	N/A
Arg.1 position	Local, symbolic	1
Arg.2 position	Local, symbolic	2

head predicate), while the latter are tied to a particular participant of the event: e.g., an argument or its lexical meaning.

As we demonstrate in the next section, these two distinctions are important in the formal model. In particular, introducing the notion of local features helps us to simulate the learning of free word order languages in a more naturalistic manner. First, however, we briefly describe how the data sets for the model were obtained.

Data collection

In this study, we use four small data sets of child-directed speech: Russian, German, English, and French. All sentences are extracted from the respective corpora in the CHILDES database (MacWhinney, 2000), approximately 500 verb usages in each

language were manually annotated with the features listed in Table 5.1. The lexical meanings of noun arguments were automatically extracted from a lexical database WordNet (G. A. Miller, 1995). The annotation procedure and the resulting corpora are described in detail in chapter 2.

5.3.2 Learning process

Key components

During the learning, the model receives input instances one by one, and the learning consists in grouping them into clusters, which potentially correspond to argument structure constructions. The initial state of the model's knowledge is a single empty cluster. While the first instance is always placed into such an empty cluster, for any subsequent instance I each existing cluster C is considered, including an empty one. The goal is to find the "best" (most probable) cluster C_{best} for the encountered instance I.

$$C_{best}(I) = \underset{C}{\operatorname{argmax}}\, P(C|I) \tag{5.1}$$

The conditional probability in (5.1), $P(C|I)$, cannot be estimated directly; therefore, the Bayes rule is applied:

$$P(C|I) = \frac{P(C)P(I|C)}{P(I)} \tag{5.2}$$

The denominator in (5.2), which is the probability of the (given) instance, has the same value for all clusters and does not affect the decision. This is why it can be excluded from the computation:

$$P(C|I) \propto P(C)P(I|C) \tag{5.3}$$

Equation (5.3) has two components: the prior probability of a cluster, $P(C)$, and the conditional probability of the instance given the cluster, $P(I|C)$.

The prior is set to be proportional to the number of AS instances previously put into this cluster, $|C|$, which is normalized by the total number of instances encountered so far $(N+1)$, see equation (5.4). The idea is that frequent categories (clusters) are more entrenched than non-frequent ones: the learner can access frequent clusters easier, and is more likely to add the new instance into such clusters.

$$P(C) = \frac{|C|}{N+1}, \tag{5.4}$$

An empty cluster is also considered for each incoming AS instance, with potentially one member: the current instance.

The conditional probability in (5.3), $P(I|C)$, accounts for the degree of similarity between the new instance and each cluster. The main difference of the present model from its earlier versions relates to how such similarity is computed, which we explain in the next section.

Interpreting instances

In the previous versions of the model, the similarity between an instance and a cluster was compared in terms of each feature independently: how similar are the verb meanings in the instance and the cluster, the first arguments, the second arguments, etc. This general approach is preserved in this study. However, consider the following two sentences (46–47) and imagine that the model first encounters an instance based on sentence (46), places it into an appropriate cluster, and then encounters an instance based on sentence (47). Without an ability to "swap" the arguments for the purpose of comparing their similarity, the model would not be able to compare the first argument *giraffe* in (46) to the second argument *giraffe* in (47), and the two instances would most probably not be grouped together, despite having nearly identical meanings.

(46) Zhiraf-a vidit petuh-∅.
giraffe-M.ACC see:3SG rooster-M.NOM
'The rooster sees the giraffe.'

(47) Petuh-∅ vidit zhiraf-a.
rooster-M.NOM see:3SG giraffe-M.ACC
'The rooster sees the giraffe.'

This is why we need to ensure that the model is able to compute the similarity not only between the local features of the first argument in a new instance and in each cluster C, but also between features of arguments with different indexes: first to second, first to third, etc. Such a mechanism is essential for languages with free word order.

Therefore, multiple possible interpretations i of the instance I are considered in the model. Each interpretation i carries exactly the same feature values as I, but the indexes of the local features FL in i may be swapped. In simple terms, whenever the model encounters an instance extracted from the sentence (46), it considers its original order of arguments, but also the reversed one (47). This is not to say that this mechanism simulates what human learners do at the implementational level: it is unlikely that humans mentally swap the arguments to consider all the alternative word orders. However, humans must be able to see similarities between sentences such as (46) and (47), and this is argued to be reflected in the resulting cognitive representations: think of the notions of alternations and allostructions in construction grammar (Perek, 2015; Cappelle, 2006).

In formal terms, let us denote the value of a particular local feature in the interpretation i as FL_k^i, the value of the respective feature in the instance I as FL_k^I, and the set of all permutations for this feature $S(FL_k^I)$. Then the set of all possible interpretations $\mathbb{P}(I)$ can be defined as provided in (5.5).

$$\mathbb{P}(I) = \left\{ i : \forall FL_k^i \in S\left(FL_k^I\right), \forall FG_k^i = FG_k^I \right\} \tag{5.5}$$

This way, the model considers each possible argument order, and selects the one with the highest similarity to one of the existing clusters. This maximal similarity value is considered to be the resulting conditional probability, see equation (5.6).

$$P(I|C) = \max(\{P(i|C) : i \in \mathbb{P}(I)\}) \tag{5.6}$$

The overall similarity value between an interpretation i and a cluster C is taken to be a product of similarities of individual features, but the individual values for all the symbolic features FS are weighed by a factor w,[4] while the distributional features FD preserve their original similarity values (equation 5.7). This is necessary, because otherwise the symbolic features related to the sentence form (lexical arguments, argument positions, etc.) dominate the clustering process, and the model's decisions are informed mainly by the form of the instances, but not their meaning.

$$P(i|C) = \prod_{k=1}^{|FD^i|} P\left(FD_k^i|C\right) \left(\prod_{k=1}^{|FS^i|} P\left(FS_k^i|C\right) \right)^w \tag{5.7}$$

Finally, the independent similarities for symbolic and distributional features are computed differently, see (5.8–5.9).

$$P\left(FS_k^i|C\right) = \frac{\left|\left\{FS_k^i \middle| FS_k^i \in FS_k^C\right\}\right| + \lambda}{\left|FS_k^C\right| + \lambda |FS_k|} \tag{5.8}$$

In equation (5.8), the term $\left|\left\{FS_k^i \middle| FS_k^i \in FS_k^C\right\}\right|$ denotes how many times FS_k^i (the value of the feature FS_k observed in the interpretation i) occurs in the cluster C, and the term FS_k^C (the total number of occurrences of the target feature in C) serves as the normalizing factor. The smoothing parameter λ is introduced both in the numerator and the denominator, but in the latter case it is multiplied by the total number of different values of the target feature in the data set. This method would not be robust for calculating the similarity in the distributional features, because their values consist of sets, and the set equality is very unlikely to hold, so that $\left|\left\{FS_k^i \middle| FS_k^i \in FS_k^C\right\}\right| = 0$. This is why the method given in (5.9) is used:

$$P\left(FD_k^i|C\right) = \left(\prod_{e \in FD_k^i} P(e|C) \times \prod_{e \in FD_k \setminus FD_k^i} P(\neg e|C) \right)^{\frac{1}{|FD_k|}}, \tag{5.9}$$

where $P(e|C)$ and $P(\neg e|C)$ are computed in the same way as in (5.8), replacing FS_k^i with the respective element e, see equation (5.10):

$$P\left(e_k^i|C\right) = \frac{\left|\left\{e_k^i \middle| e_k^i \in FD_k^C\right\}\right| + \lambda}{\left|FD_k^C\right| + \lambda |FD_k|} \tag{5.10}$$

[4] The value of this factor is set empirically, together with the value of the smoothing parameter λ (see appendix B.4). In all the simulations presented here, we use $\lambda = 10^{-14}$ and $w = 0.2$. Altering the parameter values across different simulations could be seen as an implementation of individual cognitive differences between human speakers, although we do not explore this option in the present study. Note, however, that altering parameter values across different languages might not be compatible with the usage-based framework, because this would mean that each language contains explicit information about the usefulness of its individual features. Instead, the simulated learner should be able to infer this information during the learning process.

5.3.3 Simulated picture-choice task

At any point, the learning process can be paused, and the model is tested on the picture-choice task. The model receives a set of test stimuli, each of which includes a pair of alternatives (Table 5.2), and has to choose the correct one in each pair. Note that each alternative instance comprises all the features used in the input: lexical, syntactic, and semantic. The alternatives within each pair are identical, and the only difference is in the assignment of the argument roles. As it can be seen from Table 5.2, the role properties of the two arguments are swapped, to simulate what in human experiments is a pair of images with the participant roles reversed.

Given the two alternatives, the model computes their probability given the acquired knowledge, which can be expressed as the sum of the respective probabilities over all the acquired clusters:

$$P(I_A) = \sum_C P(I_A|C)P(C) \tag{5.11}$$

To compute the two probabilities in (5.11), we use the same methods as during the learning: equation (5.6) for computing the conditional probability $P(I_A|C)$, and equation (5.4) for the cluster's prior probability $P(C)$. After evaluating the probability of each alternative, the model selects the more probable one. As we mentioned earlier, CLI may be a factor affecting the model's choice. We next propose a measure of CLI.

5.3.4 Measuring the amount of CLI

The model accumulates evidence supporting each alternative from all the acquired clusters. At the same time, some clusters contribute to the decision substantially more than others, either because they are similar to the test instance, or because they are strongly entrenched in the model's knowledge. Besides, the amount of the non-target language instances in each acquired cluster differs: some clusters are based on the instances of a single language (L1 or L2), while others are "blended" – that is, based on data from both languages (see Figure 5.2). To summarize, there are two components that determine the amount of CLI given an instance I: the contribution of each cluster to the model's choice, and the number of the non-target language instances in the cluster.

If we denote the language of an instance I as $L(I)$, then the amount of CLI can be defined as follows:

$$CLI(I) = \sum_C P(I|C)P(C)\frac{|\{J|J \in C, L(J) \neq L(I)\}|}{|C|}, \tag{5.12}$$

where the last term denotes the proportion of instances from the non-target language in the cluster C.

In the picture-choice task, each pair has a correct alternative $I_{correct}$, and an incorrect alternative $I_{incorrect}$. Using equation (5.12), we can compute the amount of CLI independently for each alternative. In this particular task the two alternatives are competing, and the support from L1 for $I_{correct}$ can be seen as positive CLI, while the support from L1 for $I_{incorrect}$ is negative. This is why the best way to quantify the impact of CLI in

Table 5.2: A pair of test instances for *The rooster touches the snake*. For simplicity, an English translation of the Russian sentence is used.

Feature	Alternative 1	Alternative 2
Head predicate	*touch*	*touch*
Head properties	{ACTION, CAUSAL, MANIPULATE, PHYSICAL}	{ACTION, CAUSAL, MANIPULATE, PHYSICAL}
Head position	2	2
Number of arguments	2	2
Arg.1	*rooster*	*rooster*
Arg.2	*snake*	*snake*
Arg.1 case	NOM	NOM
Arg.2 case	ACC	ACC
Arg.1 lexical meaning	{CHICKEN, DOMESTIC FOWL, ..., CAUSAL AGENT}	{CHICKEN, DOMESTIC FOWL, ..., CAUSAL AGENT}
Arg.2 lexical meaning	{DIAPSIDE, REPTILE, ..., CAUSAL AGENT}	{DIAPSIDE, REPTILE, ..., CAUSAL AGENT}
Arg.1 role properties	**{ANIMATE, CONCRETE, TOUCHED}**	**{ACTING, ANIMATE, ..., VOLITIONAL}**
Arg.2 role properties	**{ACTING, ANIMATE, ..., VOLITIONAL}**	**{ANIMATE, CONCRETE, TOUCHED}**
Arg.1 preposition	N/A	N/A
Arg.2 preposition	N/A	N/A
Arg.1 position	1	1
Arg.2 position	2	2

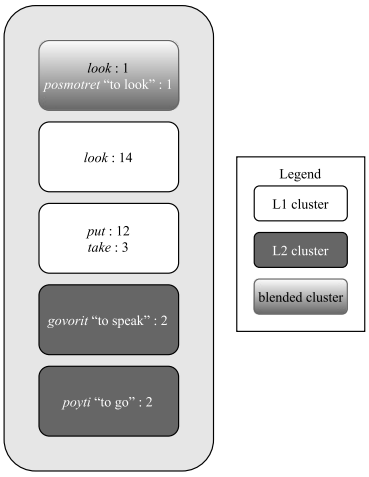


Figure 5.2: A subset of five clusters emerged in a bilingual English–Russian simulated learner. While each cluster normally consists of multiple features, in this figure only head predicates are shown for simplicity.

the picture-choice task is to measure the *difference* in the amount of CLI between the two alternatives:

$$\Delta CLI(I) = CLI(I_{correct}) - CLI(I_{incorrect}) \tag{5.13}$$

A positive value of $\Delta CLI(I)$ would mean that the positive effect of CLI prevails, while a negative value shows that CLI is damaging for the model's decision on a particular pair of instances.

5.4 Simulations and results

This section presents our computational simulations of the two target experiments. This is followed by two more simulations, which test our novel predictions regarding the comprehension of case-marking cues in additional language pairs.

5.4.1 Simulation set 1

In this experiment, we study whether our computational model performs similar to humans in the picture-choice task. Based on Janssen et al.'s (2015) results, we expect that the model will reach higher accuracy in the converging cue condition than in the conflicting cue condition. We also interpret the results in terms of CLI.

Simulation details

The 40 Russian stimuli from Janssen et al.'s (2015) experiment were obtained from the authors and annotated in the same way as our input data set. We had neither Hebrew

nor Dutch data to simulate the same language pairs as in the original experiments, yet the results of Janssen et al. were consistent across the two groups of bilinguals, which suggests that the findings generalize on other bilingual children, as long as they speak Russian and an SVO language without case marking. Among our data sets, English and French are such languages; therefore, we simulate English–Russian and French–Russian bilinguals, in addition to Russian age-matched and younger monolinguals.

Both in monolingual and bilingual simulations the model received a total of 400 AS instances (value established empirically): for monolinguals, these were Russian instances only, while for bilinguals the input included Russian and English/French instances in equal proportion. After that, the model in each condition performed the picture-choice task on the 40 test instances. To obtain the group of younger monolinguals, the simulated monolingual learners were additionally tested in the middle of the learning, after 200 training instances.

Results

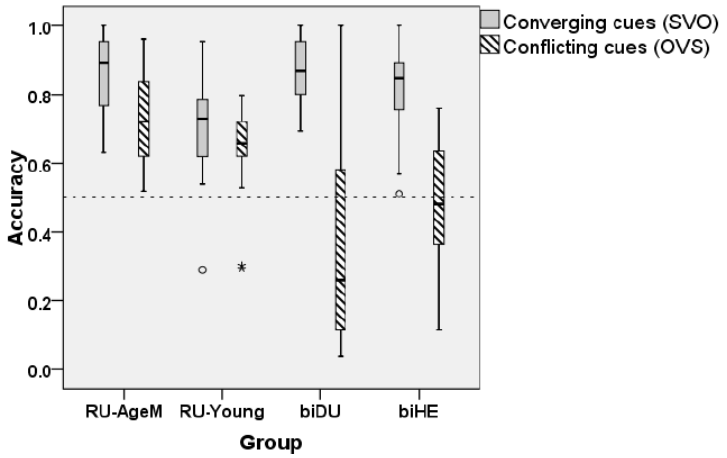
Figure 5.3 provides a visual comparison of our results vs. human data from Janssen et al. (2015). There are four groups in each figure: two groups of Russian monolinguals – age-matched and younger; and two groups of bilinguals – Dutch–Russian and Hebrew–Russian (in the original study), or French–Russian and English–Russian (in our simulations). Each group is tested in two conditions: on the stimuli with converging cues and with conflicting cues. The accuracy is measured as the ratio of the right choices to the total number of replies. We can observe the following similarities between the two studies:

1. All groups of learners in both studies perform high in the converging condition: see the gray bar plot in each pair.
2. Both younger and age-matched monolingual Russian learners (human as well as simulated) perform above chance in the conflicting condition, although not as high as in the converging condition: see the two pairs of bar plots on the left.
3. All bilingual learners perform either at chance or below chance in the conflicting condition: see the white bar plots in the two pairs on the right.

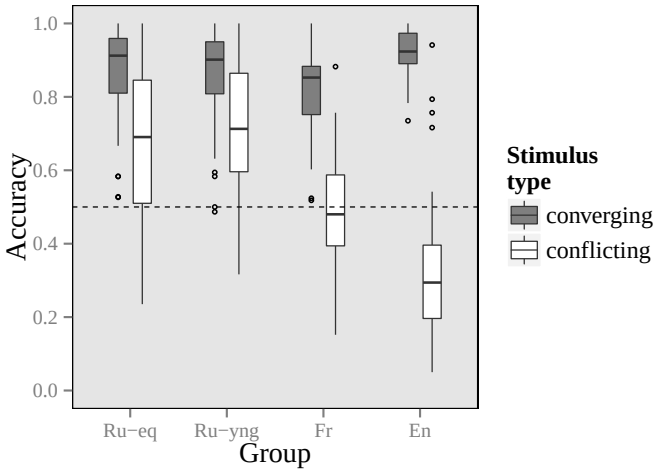
To investigate whether these similarities are statistically significant, we fit a logistic regression model to the data, which predicts the odds of making the right choice from three variables used by Janssen et al. (2015): group (age-matched Russian monolinguals vs. younger monolinguals vs. English bilinguals vs. French bilinguals), stimulus cue condition (converging vs. conflicting), and stimulus case contrast (nominative–accusative vs. nominative–dative), with all the interactions between these variables.[5] The summary is provided in Table 5.3.

When interpreting the results, it is important to keep in mind three points. First, the reference level in the table is the Russian monolingual age-matched group, conflicting cues and nominative–accusative case contrast. Second, to make the results more interpretable, we report them in terms of the probability of selecting the correct alternative

[5] We additionally tried fitting mixed-effects models to the data, but these did not converge.



(a) Original results of Janssen, Meir, Baker, and Armon-Lotem (2015). Reprinted from "On-line comprehension of Russian case cues in monolingual Russian and bilingual Russian-Dutch and Russian-Hebrew children", 2015, by B. Janssen, N. Meir, A. Baker, and S. Armon-Lotem. In E. Grillo & K. Jepson (Eds.), *Proceedings of the 39th Annual Boston University Conference on Language Development*, p. 273. Copyright 2015 by B. Janssen. Reprinted with permission.



(b) Results of our simulations.

Figure 5.3: Simulating the experiment of Janssen, Meir, Baker, and Armon-Lotem (2015).

Table 5.3: Summary of the regression model fitted to the data from our simulation of Janssen, Meir, Baker, and Armon-Lotem's (2015) experiment. Intercept corresponds to the probability of choosing the right alternative by the age-matched Russian monolingual group on the stimuli with conflicting cue type and nominative–accusative case contrast.

Variable	β	*SE*	p	$P(I_{correct})^*$
(Intercept)	1.67	0.10	$< .001$	.84
Group:En	-1.12	0.23	$< .001$	.63
Group:Fr	-0.91	0.22	$< .001$	.68
Group:Ru-yng	-0.12	0.29	.349	.82
Type:Conv	1.73	0.22	$< .001$	.97
Case:DAT	-1.12	0.21	$< .001$	.63
Group:En × Type:Conv	-0.33	0.26	.201	.88
Group:Fr × Type:Conv	-1.37	0.25	$< .001$	.75
Group:Ru-yng × Type:Conv	0.25	0.32	.431	.97
Group:En × Case:DAT	-1.24	0.27	$< .001$	.14
Group:Fr × Case:DAT	-0.41	0.24	.006	.32
Group:Ru-yng × Case:DAT	0.13	0.31	.399	.64
Type:Conv × Case:DAT	-0.86	0.24	$< .001$	.81
Group:En × Type:Conv × Case:DAT	4.10	0.31	$< .001$	.94
Group:Fr × Type:Conv × Case:DAT	3.08	0.29	$< .001$	.86
Group:Ru-yng × Type:Conv × Case:DAT	-0.36	0.35	.299	.79

* This variable shows the resulting probability of selecting the correct alternative in a particular condition: e.g., the value .88 in the line "Group:En × Type:Conv" means that the English group selects the correct alternative on a test stimulus with converging cues (and nominative–accusative contrast, which is the baseline) with the probability of 88%. Each $P(I_{correct})$ value is computed using an inverse-logit transformation on the value of the respective β-coefficient, and adding it up to the identically transformed baseline probability: intercept for the main effects, main effects for the two-way interactions, etc.

in a pair of instances, $P(I_{correct})$. Finally, only some pairwise comparisons between various factor levels are reported in the table: to obtain the missing comparisons, we use *lsmeans* package for *R* (Lenth, 2016).

First, there is a significant effect of type, which means that simulated Russian speakers interpret the nominative–accusative stimuli with conflicting cues less accurately than such stimuli with converging cues: $P(I_{correct}) = .84$ vs. .97. Our post-hoc pairwise comparisons confirm that this effect is significant in all the other group–case conditions.

More importantly, we observe a significant effect of group. The age-matched monolinguals perform significantly more accurately than English–Russian and French–Russian bilinguals on the nominative–accusative stimuli with conflicting cues: $P(I_{correct}) = .84$ vs. .63 and .68, respectively. The post-hoc comparisons yield the same effect for all the other types of stimuli, apart from the ones with converging cues and nominative–dative case contrast. Together with the main effect of case reported in the table, this suggests that the Russian monolinguals could not successfully acquire the nominative–dative cue contrast. This differs from the human subject results reported by Janssen et al. (2015). However, an analysis of the input data to our model explains this difference: the dative case occurs only 25 times in our Russian data, and not a single time in a noun-verb-noun sentence. Given such input, it is unsurprising that the model could not successfully acquire the nominative–dative cue contrast.

Another discrepancy between our results and the results of the original experiment is that we find no significant difference between age-matched and younger monolinguals, both for the reference type of stimuli ($P(I_{correct}) = .84$ vs. .82) and for the other types, as our post-hoc tests show. This may be due to the ceiling effect: the model might acquire the case-marking cue contrast right after the onset of the learning.

Despite the mentioned differences, our main finding in terms of the competition of the two cues, case-marking and word order is compatible with Janssen et al.'s (2015) results: OVS sentences are interpreted less accurately than SVO sentences, and this difference is most evident in bilingual learners. Given the competition of cues in our model, this result supports the explanation provided by the Competition Model. Next, we will investigate whether the results can be explained in terms of CLI.

Analysis of CLI

We use the ΔCLI measure introduced in section 5.3.4. Our main prediction concerns the bilinguals' interpretation of the OVS sentences: we expect the negative effect of CLI to prevail over its positive effect. This is why we first zoom in on the conflicting cue condition. The arithmetic mean of ΔCLI is negative in this condition for each group of bilinguals: -0.06 for the English group, and -0.05 for the French group. This is different from the converging cue condition, in which the corresponding values of ΔCLI are positive: 0.04 and 0.03. Although the difference is not large in absolute terms, the signs of the means are opposite, and the Mann–Whitney U test shows that the difference is statistically significant: $U = 2,079,000$, $p < .001$. The difference between the two types of stimuli is clearly visible in Figure 5.4: the average accuracy tends to be higher for those stimuli which yield more positive CLI. All together, this

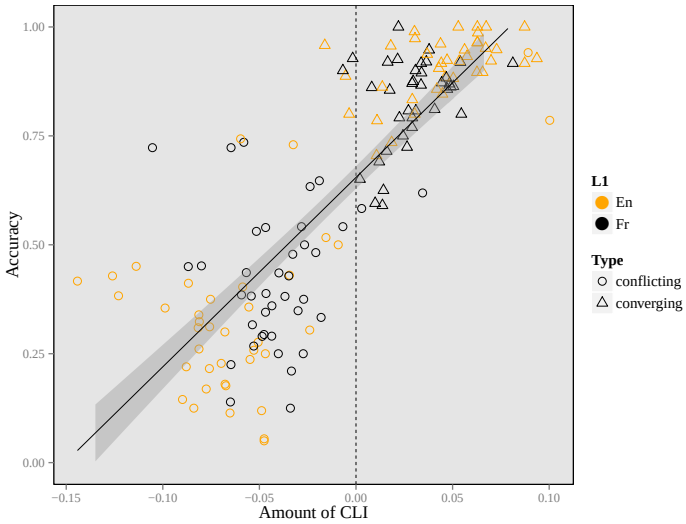


Figure 5.4: Average accuracy vs. amount of CLI per stimulus in simulation set 1, with a fitted linear regression line.

supports our prediction that the negative CLI prevails in OVS sentences, leading to their misinterpretation.

To test whether ΔCLI adds any explanatory power to the regression model reported in the previous section (Table 5.3), we updated the model by including various interactions between ΔCLI, group, type, and case. In the resulting model, the β-coefficients for the predictors and their interactions differed to a certain extent in their absolute values from those in the original model, but these differences were small and did not affect the main results – for brevity we do not report the full model. Most importantly, the amount of CLI had a significant effect on the accuracy of the two bilingual groups on the sentences with conflicting cues, judging by the respective β-coefficients. Also, the comparison between the two regression models, with and without ΔCLI, in terms of the corrected Akaike information criterion (AICc) demonstrated that the model which takes into account the amount of CLI predicted the data better: $\Delta\text{AICc} = 568$. This suggests that our ΔCLI measure is able to capture the amount of CLI, as well as its effect on the model's choice in the target task.

To summarize, the results of our simulation were similar to those reported by Janssen et al. (2015), although due to the lack of dative nouns in our input data the model could not successfully acquire the dative–nominative contrast. Taken into account the type of our computational model, this result supports the competition of cues as a plausible explanation for the misinterpretation of OVS sentences. Our analysis of CLI showed that the ΔCLI measure could serve as an additional independent predictor of the model's accuracy in the target task.

In the next experiment, we simulate a different population of learners, and further

Table 5.4: Availability, reliability, and validity of the case-marking cues in transitive sentences in our data sets.

	German			Russian		
Case	Availability	Reliability	Validity	Availability	Reliability	Validity
(Total)	.80	1.00	.80	1.00	1.00	1.00
NOM	.77	1.00	.77	.98	1.00	.98
ACC	.13	1.00	.13	.55	1.00	.55

investigate the role of CLI in the target task.

5.4.2 Simulation set 2

In our second set of simulations, we proceed with the experiment of Kempe and MacWhinney (1998). Just as in the previous section, we first test our model by simulating the picture-choice task in the two populations from the target experiment: adult L2 Russian learners and L2 German learners. Second, we investigate whether the impact of CLI on the comprehension of case-marking cues in Russian is manifested in these two populations. Ultimately, this set of simulations will also allow us to make more informed predictions about case-marking comprehension in other language pairs.

We start, however, with an additional data analysis. Kempe and MacWhinney (1998) report that the validity of the case-marking cues in Russian is higher than in German, which makes German case-marking cues more difficult to acquire and comprehend. Following their method (see section 5.2.3), we calculated the validity of case-marking and word order cues for all the transitive sentences in our data sets. The overall pattern (Table 5.4) is in line with what Kempe and MacWhinney report for their language samples, although the absolute values differ, probably due to the small number of target sentences in our data set (40 in Russian and 70 in German).

The validity of the case-marking cues, especially the accusative, is lower in German than in Russian – this is why we expect that our model will interpret Russian OVS sentences more successfully than German OVS sentences, just as the human participants in Kempe and MacWhinney's experiment.

Simulation details

We annotated the original stimuli available from Kempe and MacWhinney's (1998) study, using the same approach as for our input data sets. Recall that our data sets were obtained from child-directed speech, therefore the L1 input to our model in this experiment may not be as rich as the input that adult speakers are exposed to through the course of their life. Besides, the type of L2 input that adult learners receive differs from child-directed speech. Therefore, we use our data sets as an approximation of the input only, although they are representative in terms of the case marking in Russian and German transitive sentences.

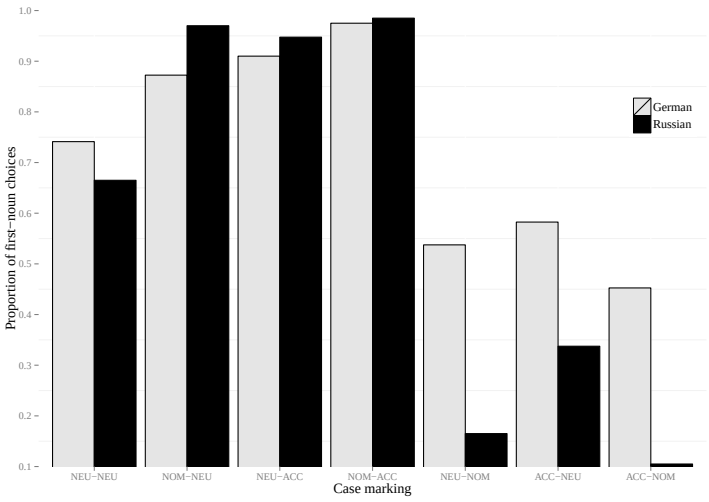


Figure 5.5: Results of our simulations of Kempe and MacWhinney's (1998) experiment.

The model was exposed to 600 English instances, followed by 600 instances of mixed input, in which English and Russian (or English and German) were contained in equal proportion. Note that these values are higher than in our previous simulation set, to better approximate adult L2 learning. After that, the model in each condition performed the picture-choice task on the 40 test instances.

Results

Figure 5.5 provides a visualization of our results. Each barplot shows how many times the SVO interpretation was chosen (first-noun-as-subject), normalized by the total number of (simulated) learners; there are seven groups of stimuli in total, depending on the case marking of the first and the second noun in the sentence. The first four groups (NEU-NEU, NOM-NEU, NEU-ACC, and NOM-ACC) represent the SVO pattern, and the other three the OVS pattern. If we compare this figure to Figure 5 in Kempe and MacWhinney's (1998) study (p. 563),[6] we can find the following similarities between the original study and our simulation:

1. In SVO sentences (four pairs of bar plots on the left), both Russian and German learners predominantly choose the first noun in the sentence as the agent.

2. In OVS sentences (three pairs of bar plots on the right), Russian learners tend to choose the second noun in the sentence as the agent, while German learners perform close to chance on this type of stimuli.

[6] This figure could not be reproduced here due to copyright issues.

Table 5.5: Summary of the regression model fitted to the data from our simulation of Kempe and MacWhinney's (1998) experiment. Intercept corresponds to the probability of choosing the right alternative by the German monolingual group on the OVS stimuli.

Variable	β	*SE*	p	$P(I_{correct})$
(Intercept)	-0.19	0.33	.756	.45
Group:Ru	3.04	0.44	$< .001$	.95
Type:SVO	5.96	1.01	$< .001$	1.00
Group:Ru $\times$ Type:SVO	-3.39	0.97	$< .001$	1.00

At the same time, the comparison of the two figures also reveals some differences between the two studies. Most importantly, human participants in Kempe and MacWhinney's (1998) study perform on SVO sentences with fully neutralized case-marking cues just as on the other SVO sentences, choosing the first noun as the agent in approximately 90% of cases. In contrast, our model exhibits a less clear preference on this type of stimuli: the proportion of first-noun choice is approximately 70% in each language. We believe it may be either due to the relatively small size of the input data that the model received compared to human speakers, or due to the model's insufficient attention to the word order cue in isolation.

Another difference relates to the relative accuracy on particular types of Russian OVS sentences. For Kempe and MacWhinney's participants, sentences with the neutralized–nominative case contrast were the most difficult to interpret among the three types of Russian OVS sentences. In contrast, our model performed worst on the accusative–neutralized case contrast. We see this difference as an artifact of the particular data sets used in our simulations.

To statistically test the difference in accuracy between the two types of stimuli (OVS vs. SVO sentences) and between the two languages (German vs. Russian), we fit a logistic mixed-effects model to the data, which predicts the odds of making the correct choice from the two mentioned variables and their interaction, with random intercepts over learners and stimuli, and with a random slope of the stimulus type over learners.[7] The model summary is presented in Table 5.5.

The results demonstrate a significant effect of language: Russian learners perform significantly more accurately than German learners on the OVS stimuli: $P(I_{correct}) = .95$ vs. .55, while there is no difference on SVO stimuli: $P(I_{correct})$ for both languages is close to 1. There is also a significant effect of sentence type: the performance of the German group on SVO sentences is significantly higher than on OVS sentences: $P(I_{correct}) = 1.00$ vs. .45. Our post-hoc analysis shows that the same effect is significant for Russian learners as well.

Additionally, we compared the performance of Russian and German simulated learners on each of the seven stimulus types shown in Figure 5.5. A logistic mixed-effects model was fitted to the data on each stimulus type with a fixed effect of language

[7] More complex models with other random slopes did not converge.

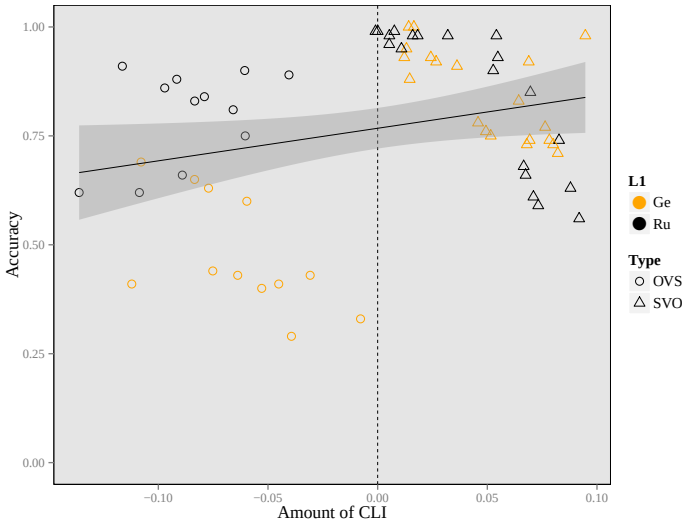


Figure 5.6: Average accuracy vs. amount of CLI per stimulus in simulation set 2, with a fitted linear regression line.

(German vs. Russian), a random intercept and a random slope of language over individual learners, and a random intercept over individual stimuli. The results demonstrated that the difference between Russian and German learners is only statistically significant in the three OVS types (the three pairs of bar plots on the right in Figure 5.5). These findings are in line with the results of Kempe and MacWhinney (1998) for the accuracy of case marking comprehension.

To conclude, the results support our prediction about the interpretation of SVO vs. OVS sentences. We proceed with the analysis of CLI in this set of simulations.

Analysis of CLI

Just as in the previous experiment, we investigate whether the choices made by our computational model can be explained in terms of CLI. We fit a regression model similar to the one described in the previous section, which includes ΔCLI and its interactions as additional predictors. The results demonstrate the effect of CLI in OVS sentences: a .1 increase in ΔCLI results in a .04 (German) or .01 (Russian) increase in the probability of making the correct choice for OVS sentences. This is also visualized in Figure 5.6: we see that the accuracy tends to be higher for the positive values of ΔCLI. The result shows that the CLI measure is highly predictive of the difference between subject groups in the target task.

At the same time, if we focus on OVS sentences and compare the ΔCLI values in Russian vs. German learners, there tends to be no difference: compare the X-coordinate of the OVS points (circles) in Figure 5.6 across the two colors. The Mann–Whitney U

test also demonstrates no support for the possible difference: $U = 1,310,400, p = .244$. This suggests that it is not the CLI that explains the difference in the interpretation of OVS sentences by German vs. Russian learners. Instead, this difference must be explained in terms of Russian-to-Russian or German-to-German influence: the higher ambiguity in the German case system, compared to the Russian system, leads to the observed difference in the model's performance on OVS sentences in Russian vs. German.

To summarize, in this set of simulations we demonstrated that our model produces results similar to the human data, when interpreting case-marking cues in Russian and German SVO and OVS sentences. The main discrepancy between our results and human subject results was observed in the interpretation of sentences with fully neutralized case-marking cues. As for the effect of CLI, it was manifested in this set of simulations just as in the previous one, but the amount of CLI could not explain the difference between the accuracy of Russian and German learners. Instead, we attribute this difference to the validity of case-marking cues, suggesting that our computational model is compatible with the Competition Model framework. In the next section we demonstrate how novel predictions can be made based on the outcomes of our two sets of simulations.

5.4.3 Novel simulations

We can now go beyond the replication setup and make predictions about the interpretation of case-marking cues in other bilingual populations. We make two specific predictions and run two additional sets of simulations to test them, followed by an analysis of the results in terms of CLI.

1. Janssen et al. (2015) in their study explain that their result, in particular the low accuracy on the conflicting sentences in bilinguals, may be "due to bilingualism in itself, ... or to the fact that the other language provided no support for case cues" (p. 276). At the same time, the presence or absence of case cues in the other language has been shown to be important: for example, L1 Italian speakers interpret L2 Spanish OVS sentences better than L1 English speakers (Isabelli, 2008). Similarly, we hypothesize that the knowledge of German with its rich case marking can be beneficial for the acquisition of Russian cases, and that German–Russian bilingual children would interpret Russian sentences more accurately than English–Russian or French–Russian bilinguals.

2. Kempe and MacWhinney (1998) demonstrate that case-marking cues are more difficult to acquire in German than in Russian, and our simulation set 2 validates this result applied to our model. It is therefore reasonable to hypothesize that monolingual German children would perform less accurately on OVS sentences in the picture-choice task when tested on German, compared to Russian monolingual children tested on Russian (i.e., as in the experiment of Janssen et al. and our first set of simulations). At the same time, bilingual French–German and English–German children are expected to perform poorly on OVS sentences,

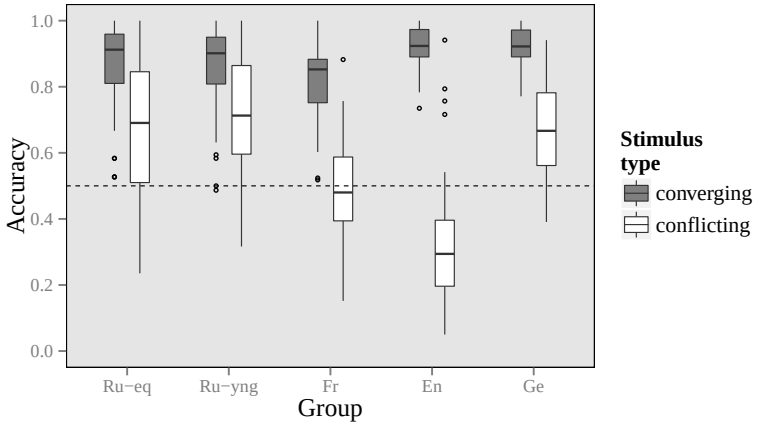


Figure 5.7: Accuracy of simulated German–Russian bilinguals against the other groups.

while Russian–German children may benefit from their knowledge of Russian and achieve higher accuracy compared to the two other groups of bilinguals.

Using our computational model, we run two additional simulations to test these hypotheses.

Bilingual German–Russian children

Following the setup described in section 5.4.1, we simulate an additional group within the same experiment: German–Russian bilinguals. This population of simulated learners is tested on the same Russian stimuli as the other four groups (Russian age-matched and younger monolinguals, and English–Russian and French–Russian bilinguals). The comprehension accuracy for the new group (utmost right plots) against the other groups is shown in Figure 5.7. It suggests that German–Russian bilingual learners have an advantage in this task over English–Russian and French–Russian learners. However, this claim is inconclusive without looking at the accuracy across the types of our stimuli. Such accuracy is plotted in Figure 5.8. We also statistically test the pairwise differences between the German–Russian group and the other groups: a logistic regression model similar to the one in the simulation set 1 (section 5.4.1) is fitted to the data for all the five groups, and all the pairwise contrasts between the German–Russian group vs. each of the other groups are analyzed using *lsmeans* package. The summary of the contrasts is provided in Table 5.6.

The results in Figure 5.8 and Table 5.6 show that the German–Russian group performs worse than the monolingual group on both types of nominative–accusative contrasts: the difference in accuracy in terms of least-square means, ΔLSM, equals 0.83 and 1.28, respectively. The relation is reversed on nominative-verb-dative sentences ($\Delta LSM = -1.36$): recall from the simulation set 1 that there were few datives in the input data, and the model could not successfully acquire the nominative–dative contrast.

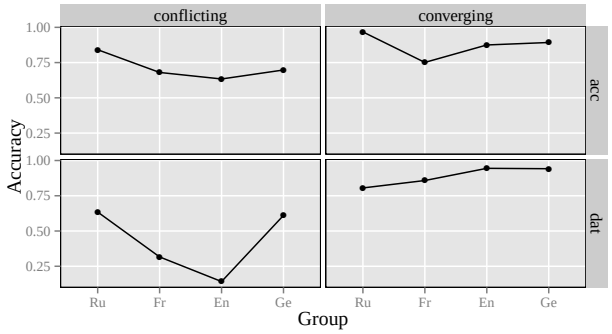


Figure 5.8: Average accuracy in each simulated group across the four types of stimuli (case × condition). For simplicity, younger Russian monolinguals are not shown: their accuracy did not differ significantly from that of the age-matched monolinguals.

Table 5.6: Summary of pairwise linear contrasts for accuracy and CLI using least-square means (LSM). "Ru" is the age-matched monolingual group, while "Ge", "En" and "Fr" denote the respective bilingual groups. The younger monolinguals are omitted for brevity.

			Accuracy			CLI		
Contrast[a]	Type	Case	Δ*LSM*	*SE*	p^{b}	Δ*LSM*	*SE*	*p*
Ru − Ge	conflicting	Acc	0.83	0.12	< .001	0.02	0.00	.001
En − Ge	conflicting	Acc	−0.29	0.11	.074	−0.03	0.00	< .001
Fr − Ge	conflicting	Acc	−0.08	0.11	.999	−0.02	0.00	< .001
Ru − Ge	converging	Acc	1.28	0.23	< .001	0.00	0.00	1.000
En − Ge	converging	Acc	−0.17	0.16	.962	0.01	0.00	.862
Fr − Ge	converging	Acc	−1.01	0.14	< .001	0.01	0.00	.942
Ru − Ge	conflicting	Dat	0.10	0.08	.907	−0.02	0.00	< .001
En − Ge	conflicting	Dat	−2.25	0.10	< .001	−0.10	0.00	< .001
Fr − Ge	conflicting	Dat	−1.22	0.08	< .001	−0.06	0.00	< .001
Ru − Ge	converging	Dat	−1.36	0.15	< .001	−0.05	0.00	< .001
En − Ge	converging	Dat	0.05	0.18	1.000	0.01	0.00	.011
Fr − Ge	converging	Dat	−0.98	0.16	< .001	−0.02	0.00	< .001

[a] German–Russian bilingual learners are at the second position in each contrast: negative Δ values are associated with the higher estimate of the respective coefficient in the German–Russian group.

[b] The *p*-values are adjusted for multiple comparisons (so-called multivariate *t*-probabilities) using the *mvt* method (Lenth, 2016).

Finally, there is no significant difference between the bilingual German–Russian and the monolingual group for dative-verb-nominative sentences ($\Delta LSM = 0.10$).

Speaking of the difference between German–Russian vs. the other two groups of bilinguals, we can see that the former perform substantially better than the other two groups on dative-verb-nominative sentences, and than the French–Russian group (but not English–Russian) on nominative-verb-accusative and nominative-verb-dative sentences. Because there are no significant differences for the other types of sentences, our hypothesis about the facilitatory effect of the German knowledge is supported only partially.

To investigate the contribution of CLI to this result, we can compare the differences in accuracy to the differences in the amount of CLI across different groups of learners. A linear mixed-effects model has been fitted to the data, predicting this time the amount of CLI (ΔCLI), and the pairwise contrasts were computed (see Table 5.6 on the right). We can see that the greatest difference in the amount of CLI is observed for dative-verb-nominative sentences ($\Delta LSM = -0.10$ and -0.06), which is the only type of stimuli on which the German–Russian group scores higher in accuracy than both other bilingual groups. This means that the amount of positive CLI for this type of stimuli is higher in German–Russian group than in French–Russian and English–Russian group, in line with our prediction.

As for the other types of stimuli, the amount of CLI in German is simply not high enough to facilitate the interpretation of Russian sentences: this can be demonstrated by plotting the average amount of CLI across the three groups of bilinguals, see Figure 5.9. Note that in the conflicting cue condition, the amount of CLI is always higher in the German group than in the other two groups. However, its absolute value is positive only for the dative-verb-nominative sentences, but not for accusative-verb-nominative. This explains why we observe no differences across the bilingual groups for this latter type of sentences.

To summarize, our simulated data only partially confirms our prediction that the knowledge of German can facilitate the interpretation of Russian case marking. The lack of the hypothesized effect can explained by the amount of CLI across different types of stimuli.

Bilingual and monolingual German children

Until this point, we have simulated the experiment of Janssen et al. (2015) using Russian sentences. Our final prediction, however, concerns case comprehension in German. We use the same setup as in simulation set 1, but this time simulating German monolingual children and three groups of bilinguals: French–German, English–German, and Russian–German. All the four groups are tested on German instances. Ideally, we would translate Janssen et al.'s stimuli into German, however many of such translated sentences would be fully neutralized in terms of their case-marking cues. To give an example, the sentence *Kukla lyubit zhirafa* "The doll loves the giraffe" would translate into German as *Die Puppe liebt die Giraffe*, where the case of both nouns can be interpreted as either accusative or nominative. Therefore, in this experiment we used a subset of Kempe and MacWhinney's German stimuli, 24 out of 32: the 8 fully

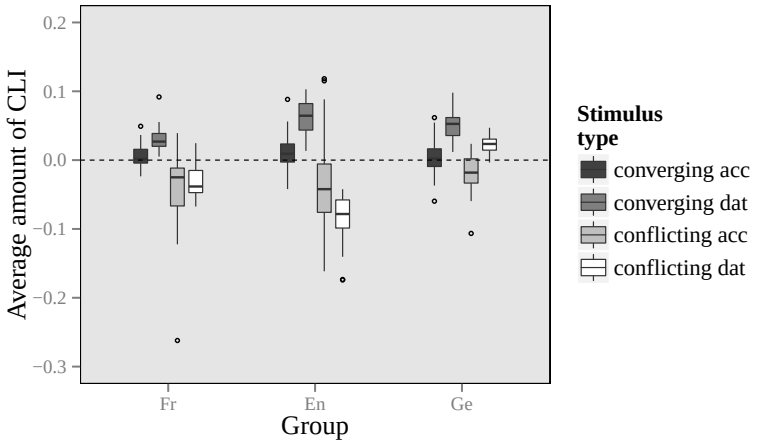


Figure 5.9: Amount of CLI (Δ*CLI*) per group, averaged over stimuli, in the Russian picture-choice task.

neutralized stimuli were eliminated.

The performance of the four groups is shown in Figure 5.10, while in Table 5.7 we also provide a summary of the logistic mixed-effects model fitted to the data. First of all, the results show that German monolinguals (the utmost left pair of plots) perform well in the converging cue condition, but close to chance in the conflicting cue condition. This is in line with the findings of Kempe and MacWhinney (1998) for L2 learners (which we simulated in experiment 2), and also with the existing data suggesting that German children are only able to interpret case marking in OVS sentences around the age of seven (Dittmar et al., 2008). Besides, this result is clearly different from the accuracy of simulated Russian monolinguals in our simulation set 1, who performed well in the conflicting cue condition. This supports our prediction about the accuracy of monolinguals on German vs. Russian OVS sentences.

As for the bilingual groups, both English–German and French–German bilinguals perform in the conflicting cue condition less accurately than monolinguals. Interestingly, our simulated Russian–German bilinguals perform significantly better than German monolinguals in this condition. While this is inconsistent with the view that bilinguals lag behind monolinguals in their language development (Schmitz, 2006; Hulk, 2004), bilingual children have been shown to acquire some grammatical features earlier than monolinguals (Pléh, Jarovinskij, & Balajan, 1987; Meisel, 1986). In our model, this may only happen if bilinguals benefit from positive CLI. To investigate this, we analyze the Δ*CLI* values for all the groups, focusing on the conflicting condition. The comparison is provided in Figure 5.11. Note the obvious difference across the groups in the conflicting condition, but not in the converging condition. The direction of this difference is as expected: the effect of CLI is positive in Russian–German bilinguals and negative in the other two groups.

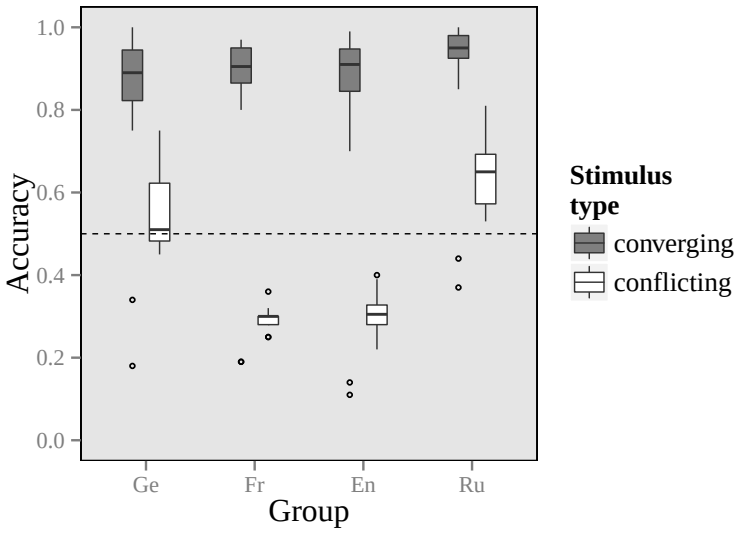


Figure 5.10: Accuracy per group in the German picture-choice task.

Table 5.7: Summary of the regression model fitted to the data from our simulation of the German picture-choice task. Intercept corresponds to the probability of choosing the right alternative by the German monolingual group in conflicting cue condition.

Variable	β	SE	p	$P(I_{correct})$
(Intercept)	0.23	0.38	.537	.56
Group:En	-1.05	0.09	$<.001$	.31
Group:Fr	-1.11	0.09	$<.001$	.29
Group:Ru	0.40	0.09	$<.001$	.65
Type:Conv	1.68	0.50	.001	.87
Group:En $\times$ Type:Conv	0.93	0.15	$<.001$	.86
Group:Fr $\times$ Type:Conv	1.18	0.15	$<.001$	.88
Group:Ru $\times$ Type:Conv	0.46	0.16	.005	.94

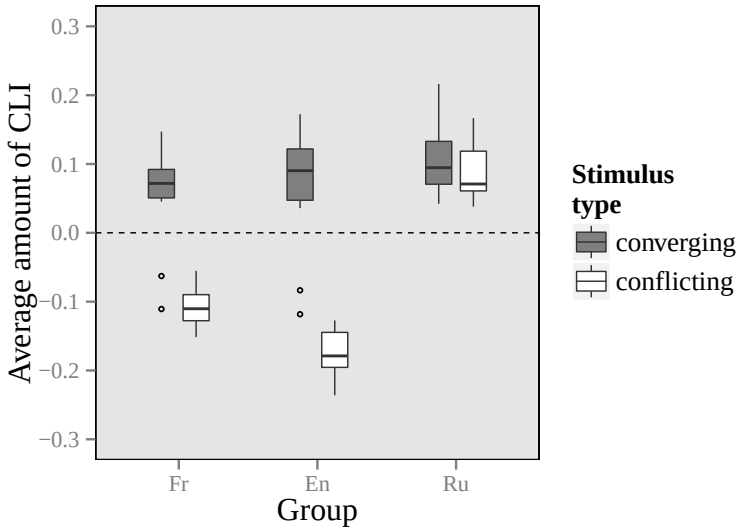


Figure 5.11: Amount of CLI (ΔCLI) per group, averaged over stimuli, in the German picture-choice task.

To summarize, the results of this simulation set support our prediction: the accuracy of simulated monolinguals on the sentences with conflicting cues is lower in German than in Russian. This is in line with Kempe and MacWhinney's (1998) findings and with the results of our simulation set 2, corroborating the idea in the Competition Model that the validity of case cues is important. The results also confirm our hypothesis about the performance of the bilingual groups on the German OVS sentences: while the simulated French–German and English–German bilinguals performed poorly on such sentences, Russian–German bilinguals benefited from positive CLI and achieved high accuracy.

5.5 Discussion

Our goal in this study was to demonstrate how the amount of CLI can be measured in a computational model, and how such a measure can be applied for explaining a particular phenomenon, the comprehension of case-marking cues.

5.5.1 Quantifying the effect of CLI

We introduced a measure of CLI and used it to quantify the CLI effect in the picture-choice task. This measure helped us to determine the contribution of CLI to the observed result. As it was demonstrated, the measure can be used both on the level of a particular group or condition (e.g., average amount of CLI in the interpretation of OVS

sentences by Russian–German bilinguals), and on the level of a particular test item (cf. Figure 5.4), in case the goal is to study the differences between individual sentences.

In this study, we only employed a particular linguistic task related to sentence interpretation, but it is perfectly possible to use this computational model for simulating other linguistic tasks: filling in verbs or prepositions, verb definition, verb selection, etc. (recall chapters 3–4). For all these tasks, the same type of measure (*CLI* or ΔCLI) can be used to quantify the amount of CLI and shed light on its role CLI in language comprehension and production.

5.5.2 CLI in case-marking cue comprehension

Speaking about the linguistic phenomenon of interest – case-marking cue comprehension – it was demonstrated in this study that our probabilistic computational model performed in the target task similar to human learners: early bilinguals in the experiment of Janssen et al. (2015) and L2 learners in the experiment of Kempe and MacWhinney (1998). In section 5.1.2 we outlined two general views on the role of CLI in case-marking cue comprehension by bilingual or L2 learners. Our results suggest that CLI is an important factor that affects this kind of comprehension. This explicitly contradicts VanPatten's (1996) First-Noun Principle: first, an effect of the amount of CLI was observed in all our simulations; second, the performance of simulated German–Russian speakers on Russian and German OVS sentences (recall our novel sets of simulations) was higher compared to bilinguals whose other language had no case-marking cues (i.e., French or English).

At the same time, recall from section 5.1.2 that there are at least two theories promoting the role of CLI: the Competition Model and the L1 Transfer Principle. In our model, various features compete with each other, which makes it compatible with the Competition Model. This similarity is also supported by the data of our simulation set 2: while the amount of CLI could not explain the difference in the performance of English–German vs. English–Russian learners on OVS sentences, our analysis of the cue validity in the German and Russian input data suggested that the model was sensitive to the cue validity, at least for the case-marking cues. This being said, our study supports MacWhinney's Competition Model as the explanation of the misinterpretation of OVS sentences. At the same time, the results do not necessarily challenge VanPatten's L1 Transfer Principle. To our knowledge, the cognitive mechanisms behind this principle have not been described in detail, and this is why it is challenging to verify of falsify this principle.

5.5.3 Additional factors

It is important to keep in mind that in the present study we did not consider all the cues that affect sentence comprehension. In particular, we did not take into account the pragmatics of the utterance, expressed in the intonational cues: such cues have been shown to be highly informative for the interpretation of participant roles by monolingual German children (Grünloh, Lieven, & Tomasello, 2011).

Another factor that was left out of our simulations is the lexical similarity between languages. In the study of Isabelli (2008), L1 Italian speakers could interpret L2 Spanish OVS sentences more accurately compared to L1 English speakers, and positive CLI was suggested to be responsible for this. However, as VanPatten (2015a) noticed, this may also be due to the better familiarity of L1 Italian speakers with L2 Spanish personal pronouns used in the experiment. This is why lexical similarity, including the potential effect of cognates, must be taken into account. In our computational model this effect could be captured only for the words spelled identically in the two target languages, such as *giraffe–Giraffe* in English and German. At the same time, there are more cognates between English/French and German, compared to Russian – the results in our simulations (in particular, simulation set 2) may have differed to a certain extent, had the effect of cognates been taken into account.

5.5.4 CLI in argument structure constructions

The present study also sheds some light on how CLI may occur at the representational level. Given the learning mechanism implemented in our computational model, as well as the type of CLI measure used, there are two ways for the CLI measure to obtain higher values. First, the learner may have the two languages separated in the existing constructions: that is, some constructions are based on L1 only, and others on L2 only. In order for the CLI value to be high, in this case the similarity between a test instance in the target L2 language and some of the existing L1 (non-target language) constructions must be rather high. This is possible, but improbable. A more likely alternative explanation is that some constructions are blended, as it was shown in Figure 5.2. Given a test instance in L2, the model sometimes makes its choice based on such blended constructions, hence the high CLI value. This supports the view that constructional representations may be shared across languages (Higby et al., 2016; Bernolet et al., 2013; Salamoura & Williams, 2007). At the same time, our preliminary analysis of constructional representations demonstrates that most constructions are based on a single language – either L1 or L2; and this explains why the amount of CLI in absolute terms was not high in the present study: looking back at Figures 5.4, 5.6, 5.9, 5.11, we can see that the the contribution of CLI to the learners' decisions hardly ever exceeded 30% in our simulations, and in most cases it was under 10%. This may explain why it is difficult to find such blended constructions using cross-linguistic comparisons (Wasserscheidt, 2014).

CHAPTER 6

General discussion

This thesis reported on three empirical studies that employed computational modeling to investigate the process of learning constructions in two languages (chapters 3–5). Additionally, chapter 2 described the corpora collected during this project and used as material for the computational simulations. While each individual study itself contains a theoretical discussion, here I summarize the studies and discuss their broad theoretical and methodological implications.

6.1 Overview

6.1.1 Summary of findings

In chapter 2, I described two corpora used in the simulations. The smaller multilingual corpus consists of English, German, Russian, and French data manually annotated with argument structure information, while the larger English–German corpus is compiled from the existing linguistic resources. The corpora contribute to the trend towards the integration of various linguistic resources (e.g., Lopez de Lacalle et al., 2016; Wu & Palmer, 2015; Palmer, 2009; Shi & Mihalcea, 2005). The combination of both syntactic and semantic information is an important advantage of the collected corpora. They were used for training the computational model in my empirical studies, as well as for generating some of the test data. They can be used for similar purposes in cognitive computational modeling, but also in natural language processing tasks, such as semantic role labeling or relation extraction.

The three modeling studies dealt with the effects of several quantitative input properties on the model's performance in various linguistic tasks. In chapter 3, I

focused on the role of such properties in both L1 and L2 learning. This study was inspired by the research of Nick Ellis and colleagues (N. C. Ellis et al., 2014a, 2014b; also Römer et al., 2015, 2014), who proposed a formal model predicting speakers' verb choice within a given construction from distributional and semantic variables in the linguistic input that speakers are exposed to: joint verb–construction frequency, strength of verb–construction mapping, and prototypicality of verb meaning. Computational modeling enabled me to overcome certain methodological challenges and provide a refined prediction model: the simulation results suggested that overall verb frequency was an additional factor affecting speakers' choice, while joint frequency and strength of mapping had a combined effect rather than independent. Importantly, this study demonstrated that L1 and L2 statistical learning could be simulated using the same approach. My comparison of the model's performance across the two languages yielded the differences quantitatively similar to those reported by Römer et al. (2014) for native vs. non-native human speakers.

In chapter 4, two global properties of the input were examined: the amount of L2 input and the time of L2 onset. I investigated the impact of these variables on the success of learning of argument structure constructions in L2, measured by five different linguistic tasks. The amount of L2 input was shown to predict the model's performance in all tasks. Importantly, it was the absolute amount that mattered, while the way the input was distributed across the overall learning trajectory was unimportant. The time of onset did not show any effect on the model's performance, in contrast with the findings in domains such as lexical or morphological learning, where the negative effect of the late onset was evident (Monner et al., 2013; Zhao & Li, 2010). I explained the lack of the negative effect by possible positive cross-linguistic influence between English and German – typologically close languages used in this set of simulations.

Chapter 5 was concerned with cross-linguistic influence: I proposed a method of measuring the amount of CLI in the computational model, and applied this method for studying case-marking comprehension in Russian and German, using French and English as additional languages. Russian and German use case-marking cues to discriminate between subject-verb-object (SVO) and object-verb-subject (OVS) sentences, and it is common among late learners of such languages to misinterpret OVS sentences as SVO. Two broad views on the role of CLI in this error exist. According to the First-Noun Principle (VanPatten, 1996), CLI has no effect, because speakers always assign the agentive (subject) role to the first noun or pronoun in a given sentence. An alternative account explains the error by CLI from learners' L1. My study defended the latter view: I demonstrated why CLI mattered in simulated learners by using a quantitative measure of CLI. As I argued, this result also suggested that the Competition Model (Morett & MacWhinney, 2013) was a more plausible explanation of the CLI than the L1 Transfer Principle (VanPatten, 2015b).

6.1.2 The broad picture

Before proceeding with the discussion of the broad implications of this thesis, I provide an integrative summary of its main original contributions.

- Implementing a probabilistic computational model of learning argument structure constructions in two languages (chapters 3–4), also compatible with the languages with free word order (chapter 5).
- Collecting a multilingual corpus of verb usages annotated with argument structure information (chapter 2).
- Carrying out three studies that contribute to the development of the statistical account of learning two languages (chapters 3–5).
- Refining the formal account of verb selection within argument structure constructions, both for L1 and L2 learning (chapter 3).
- Proposing a formal method to quantify cross-linguistic influence computationally (chapter 5).
- Comparing the predictive power of three measures reflecting the association strength between verbs and constructions (chapter 3).
- Advancing our understanding of the role of statistics and semantics in construction learning (chapter 3).
- Studying in isolation the impact of two variables – amount of L2 input and moment of L2 onset – on L2 learning (chapter 4).
- Demonstrating the role of CLI in the interpretation of subject-verb-object sentences (chapter 5).

I will provide more details on each point in the following sections. For now, it is important to keep in mind the broad picture. My studies are best situated within the usage-based approach to language learning, and statistical account of learning. A constructionist perspective was adopted, in particular that described in Goldberg's Construction Grammar, and her approach to argument structure. The modeling approach belongs to the general framework of probabilistic (Bayesian) modeling in cognitive science. In linguistic terms, my computational model also has much in common with the Unified Competition Model (MacWhinney, 2012, 2008): see especially chapter 5. Both this theory and my computational model employ the idea of having multiple cues (features) in the input, which compete with each other for determining speakers' linguistic decisions.

6.2 Theoretical implications

6.2.1 Statistical account of bilingual learning and use

The major contribution of this thesis relates to the development of a statistical account of construction learning and use in situations when the learner is exposed to more than a single language. This was achieved through the implementation of a probabilistic model

of learning argument structure constructions. This model rests on several assumptions and has some basic characteristics.

1. The emergence of abstract constructions in this model is simulated as a process of generalizing over individual usages, in line with the construction grammar account outlined, for example, in Goldberg et al. (2004), Goldberg (1995).

2. The processing of incoming usages relies on two factors: entrenchment of the existing constructions, in line with the respective theories in cognitive linguistics (Schmid, in press; Langacker, 1987); and similarity between usages and acquired constructions, following studies that demonstrate the role of similarity in categorization (Sloutsky, 2003; Hahn & Ramscar, 2001). At the same time, when it comes to entrenchment, it is important to mention that no forgetting mechanism is implemented, and the knowledge of the acquired constructions does not decay with time.

3. Both usages and constructions are represented as assemblies of various features – lexical, semantic, syntactic, and pragmatic. This is similar to theories such as the Competition Model (e.g., MacWhinney, 2008; E. Bates & MacWhinney, 1989), in which various features compete for informing the learner's linguistic decisions.

4. The model has direct access to the lexical, morphological, and semantic information in the input, which represents a situation when the learner has already acquired words and basic morphology by the moment s/he starts learning abstract constructions. This is clearly a simplistic assumption, made due to our focus on the domain of abstract construction learning alone – a common approach in computational modeling. In reality, human learners learn words and morphology in parallel with syntax (e.g., Lieven & Tomasello, 2008).

5. L1 and L2 usages are not explicitly labeled as such, although some of the features (in particular, lexical) do carry language-specific information. This allows the model to place similar L1 and L2 usages into the same construction, following the assumption that languages share their storage resource and may form shared constructional representations (e.g., Abutalebi & Green, 2007).

6. Similarly, when making the decisions about language use, the model equally considers all constructions, irrespective of whether they are language-specific or "blended", and makes the decision probabilistically. In terms of statistical learning, this means that the evidence in favor of one or another linguistic choice is collected from all the constructions emerged in the model's repertoire. This mechanism is grounded in the experimental findings of Hartsuiker et al. (2016), Higby et al. (2016), etc., which demonstrate how linguistic representations from both languages are employed in actual use.

Speaking of *learning*, the computational model in my studies could form reasonable argument structure generalizations, relying on the mechanisms of probabilistic

(Bayesian) learning alone. This supports some of the described assumptions (especially points 1–3 and 5). In short, the studies in this thesis (chapters 3–5) clearly demonstrate that the statistical learning mechanism *may* be sufficient for successful bilingual (or L2) learning, at least when it comes to the acquisition of argument structure constructions.

Turning to language *use*, through which the model's language proficiency was assessed, the purely statistical approach to language use (points 5–6 above) yielded reasonable behavior of the model in comprehension tasks, such as word ordering, verb definition, role comprehension (chapter 4), and case-marking comprehension (chapter 5). For example, in the latter study the model performed similar to human learners. At the same time, in production tasks, such as filling in verbs or prepositions (chapter 4), the model produced both L1 and L2 lexemes in the L2 test. This is why in chapter 3 L1 verbs generated by the model in the L2 test were excluded from the analysis. On the one hand, this suggests that the purely statistical approach is not enough to simulate language production in bilinguals: mechanisms inhibiting the activated lexemes from the non-target language (e.g., Kroll et al., 2008; Green, 1998) are needed. On the other hand, in tasks such as filling in verbs the choice of L1 vs. L2 verbs may rely on the knowledge of lexically specific collocations or word co-occurrence patterns, which, again, can be acquired with statistical mechanisms (Webb, Newton, & Chang, 2013). Finally, note that mixing L1 and L2 lexemes is not uncommon in bilingual speakers: think of the widespread phenomenon of insertional code-switching, when both languages are used within the same sentence (e.g., Auer, 2014). Code-switching most frequently occurs in colloquial speech, and language statistics may be successful in simulating this type of language use in bilinguals. All together, this supports the single-system view on language acquisition and use (cf. assumptions 5–6 above), commonly adopted in usage-based linguistics (Ortega, 2015; N. C. Ellis et al., 2014a; MacWhinney, 2012, 2008; N. C. Ellis, 2006b).

6.2.2 Age/order effect

The study in chapter 4 contributes to the discussion about the age/order effect in acquisition. While in other domains, such as lexis or morphology, the late onset of a language has a pronounced negative effect on learning (Monner et al., 2013; Zhao & Li, 2010), this is not so for English and German argument structure constructions in my study. On the one hand, this may be an artifact of the high amount of positive CLI between English and German, as I argue in chapter 4 (and the simulations in chapter 5 support this claim). On the other hand, it is important to keep in mind that the time of onset in terms of statistical learning is a merely distributional variable: even after a long period of exposure to the L1 the statistical learner may preserve the ability to create new representations (what is called plasticity in associative learning). In this case, L2 constructions do not "parasitize" on the respective L1 constructions, and the negative onset effect is not observed. Other mechanisms than purely statistical learning may be responsible for this effect. To give an example of a mechanism which is not captured in a purely statistical learner, selective attention may negatively affect L2 learning: the learner gets accustomed to relying on certain cues in L1, while these cues may be less informative in L2. This is known as the transfer of cue strengths

in the Competition Model (MacWhinney, 1992), and reflected in the phenomena of overshadowing and blocking in cognitive theories of associative learning (N. C. Ellis, 2006b). One possibility to simulate these phenomena would be to set the weights of each feature (similar to what I did in chapter 5) at the onset of L2 learning and allow for their adaptive adjustment depending on the properties of L2 input.

6.2.3 Statistics and semantics in language learning

The input data that the computational model was trained on in my studies contained various types of features: lexical, syntactic, semantic, and (in chapter 5) pragmatic. Respectively, the emergent representations were assemblies of various features, too. This reflects the idea in construction grammar that constructions are form–meaning pairings. All features equally contributed to the probabilistic learning process (although this was different in chapter 5, where a weight was assigned to each feature). At the same time, there is a discussion in the literature whether the role of semantics in learning is independent of that of so-called language statistics: the distributional properties of linguistic input. One view is that the effects of semantics are implicitly captured when one considers the distributional properties of linguistic forms: forms cannot be dissociated from their meanings, which makes the effects of language statistics and semantics intertwined or at least highly confounded (Ninio, 1999a); this may be why the two independent effects are not always observed (Theakston et al., 2004). According to the other view, both statistics and semantics affect the learning independently (e.g., Ambridge, Bidgood, Pine, et al., 2015; N. C. Ellis et al., 2014a). My study in chapter 3 supported this latter view. Note, however, that semantics in my computational model is learned probabilistically – that is, through language statistics, just as the other features. This suggests that the effect of semantics can, in fact, be captured by looking at distributional properties, but these must be properties of meanings, not forms.

6.2.4 Cross-linguistic influence in constructions

This thesis also has implications for research on cross-linguistic influence. Specifically, the study in chapter 5 contributed to our understanding of how CLI occurs at the representational level: it supported the view that CLI is a result of "blending" multiple languages within some constructions (Higby et al., 2016; Bernolet et al., 2013; Salamoura & Williams, 2007), although most constructions in the repertoire acquired by the model were still language-specific. At the same time, the non-selective access, or cross-language activation (Kroll, Bobb, & Wodniecka, 2006; Marian & Spivey, 2003), was one of the assumptions in my computational model, and it is well possible that in some cases the non-selective access alone might have been responsible for the high amounts of CLI, without the need to have blended constructions.

6.3 Methodological implications

6.3.1 Computational modeling

This thesis has an important methodological component: computational modeling has not been commonly used for studying how two or more languages are acquired. One of the starting points of this thesis was that we can advance our understanding of bilingual and L2 learning through the use of computational modeling: the high control over all the variables provided by a computational model resolves the problem of inherent variability in human learners. My studies demonstrated that computational modeling, indeed, could provide useful insights on various aspects of bilingual and L2 learning: the role of statistics and semantics, age/order effect, and CLI. More specifically, the computational model I employed in this thesis enabled me to study in isolation the impact of such variables as the amount of input and the time of L2 onset, which are confounded in human speakers (chapter 4). It also allowed me to measure the amount of CLI not by analyzing language use, as it is common in human subject research, but by looking inside the "black box" of actual linguistic representations, which is nearly impossible to do with human participants (chapter 5).

Compared to the earlier studies that employed a computational model with the same learning mechanism (Alishahi & Stevenson, 2010, 2008), in the present thesis I adapted the model for simulating the learning of two languages. Also, different test tasks were employed, and the model was trained on novel data sets. Finally, for the study in chapter 5 the learning mechanism was adapted to accommodate the learning of languages with relatively free word order, such as Russian. This latter methodological improvement may serve as a starting point for new cross-linguistic research with this computational model.

Speaking about the studies in this thesis in relation to human subject research, it is also essential to keep in mind George Box's quote that "all models are wrong but some are useful". In our case, while the computational model simulates the process of statistical language learning and use, it must not be seen as a perfect equivalent of a human speaker. On the contrary: whenever human data yield a pattern which is not supported in my studies (think of the missing effect of the time of onset in chapter 4 or of the rather low effect of semantics in chapter 3), it means that this particular model (or, more generally, the statistical learning account alone) cannot account for this pattern, and additional mechanisms must be at play. Compared to humans, my model can be metaphorically described as an "idealized" speaker: to name a few of its features, the model does not suffer from forgetting, it receives little noise in the input, it processes each and every incoming instance. Many of these issues can be resolved within the same modeling framework: for example, Alishahi and Stevenson, 2008 showed that the model can successfully learn constructions from the noisy input as well. At the same time, presenting additional simulations in each study would shift the focus of my thesis away from SLA and bilingualism.

Comparing the results of computational simulations obtained from many different models could yield a more comprehensive evaluation of a particular model as the one employed in this thesis, and also lead to a better understanding of the effective

mechanisms in bilingual learning. At the same time, it is at least equally important to compare simulated results to existing human data – the approach I took in this thesis. The fact that the model in my studies yielded certain patterns similar to human learners (chapters 3–5), supports some of the theoretical assumptions that the model rests on: those related to the learning mechanisms, and the existence of blended constructions (recall the discussion in section 6.2.1). Other assumptions, for example the learner's prior knowledge of lexis and morphology, or the lack of explicit mechanisms to sort out the activated representations from the non-target language, are obvious simplifications that must be addressed in the future.

6.3.2 Quantitative approach to language

This thesis strongly promotes a quantitative approach to bilingual and L2 learning. This is, of course, not to diminish the importance of qualitative research in SLA: on the contrary, I recognize that some phenomena in language learning can be best described qualitatively. However, the field of SLA emerged as an applied discipline, and one problem emphasized since long time ago was the lack of quantitative methodologies (Brown, 1991). A recent special issue of *Language Learning* on quantitative reasoning in SLA suggests that the problem persists (Norris, Ross, & Schoonen, 2015). Studies in my thesis contribute to strengthening the quantitative approach to SLA. First of all, the method of probabilistic computational modeling relies on the distributional properties of language and is inherently quantitative. Second, chapter 3 demonstrates how the use of statistical mixed-effects models can help us to account for individual variation between speakers, and how model comparison can provide evidence in favor of one or another theory. These statistical methods can be applied just as successfully in SLA research with human subjects (e.g., Linck & Cunnings, 2015). Third, studies in this thesis propose quantitative measures for such phenomena as the amount of input, the time of language onset (chapter 4), and the amount of CLI (chapter 5).

Speaking about quantitative measures, another contribution of this thesis relates to the discussion in cognitive linguistics on the existing measures of association strength between two linguistic units (e.g., Pecina, 2010; Wiechmann, 2008). Chapter 3 contains the relevant overview: in short, there is no agreement on which measure can predict human data best. My study suggests that for estimating the association strength between verbs and constructions, joint verb–construction frequency is better than other measures. At the same time, it is important to keep in mind that this measure in my study was used in combination with the overall verb frequency, so that taking into account both of them may be necessary.

6.3.3 Individual variation

Individual variation sometimes constitutes a major problem for studies in SLA, due to enormous variability even in a seemingly homogeneous sample of learners (R. Ellis, 2004): there are numerous learner variables (e.g., mother tongue, history of language learning and use) and learning variables (e.g., learning setting, type of instruction), as I explain in the introduction. The use of computational modeling helped me to either

eliminate or keep under control most sources of variation: aptitude, motivation, L1 background, moment of onset, learning setting, amount of input, etc. At the same time, it has been argued that the emergent linguistic representations depend on the exact learner's experiences with the language (e.g., Misyak & Christiansen, 2012). Therefore, in my studies each simulated learner was exposed to an individual input sample, to simulate a population of different statistical learners and prevent the situation in which the emergent constructions are a sheer artifact of a particular input sequence.

This way, a certain type of variation between learners was preserved, and this variation had to be accounted for in the statistical analysis, for example using mixed-effects models with a random effect of individual learners. Without this step, one can only report the effects observed on the level of populations, but not of individual learners. There is no guarantee that the effects parallel each other on the two levels (e.g., Verhagen & Mos, 2016). In my study (chapter 3), taking the individual variation into account did not invalidate the main findings, yet the effect sizes were found to be different.

6.4 Future work

The studies in this thesis outline a number of general directions for future research. First, they invite for more research on bilingual and L2 learning of lexis *and* constructions at the same time. In particular, this is important for our understanding of the similarities and differences between the learning of lexis and abstract constructions in bilingual speakers. At this point, it is unclear why the findings for lexis (e.g., in terms of the effect of the time of onset) are not always applicable to constructions, and vice versa. This is especially important for clarifying the common view in construction grammar on the existence of the syntax–lexicon continuum (e.g., Broccias, 2012; Boas, 2010). In particular, I did not focus on the development of fixed expressions ("chunks") in the computational model, although the emergence of such expressions is perfectly possible: it occurs when exactly the same instance is repeated in the input over and over again, yielding a highly entrenched cluster containing only this instance.

Second, in the present thesis I only focused on bottom-up statistical learning. However, top-down learning (explicit instruction) is an essential component in many population of L2 learners, especially in classroom settings (e.g., N. C. Ellis, 2015; DeKeyser, 2008). One important direction for developing a comprehensive computational account of SLA is to implement a mechanism of providing the model with explicit "instructions", for example by forcing several linguistic usages to be placed into the same construction. Speaking more broadly, no distinction was drawn in my studies between early and late L2 learners, apart from the difference in the moment of L2 onset (and, respectively, the amount of prior L1 input, see chapter 4). This was due to the focus of my thesis: to test how much of L2 learning can be explained by statistical learning alone (cf. section 1.2). Clearly, the cognitive system of an adult L2 learner differs from that of a child, and this must have consequences for the exact mechanisms involved in the learning process. A future ideal model, as I mentioned in the introduction, will account for such differences.

Third, the syntactic representations used by my computational model differed across the studies in this thesis. Most notably, the study in chapter 5 represented each syntactic pattern as an assembly of several features, such as the positions of the head verb and its arguments. This may or may not be realistic in terms of human sentence processing. We do not know yet how exactly speakers decompose utterances into their smaller constituents in comprehension, and how these are joined to form utterances in production. Using the most recent data from the areas of language production and comprehension may help to make the computational model more cognitively plausible in this respect. At the same time, the learning model itself can inform the mentioned fields: if simulation results replicate human data on language acquisition, then the employed feature structure is likely to be cognitively plausible. However, this requires a direct comparison of the model's performance on different sets of features.

Fourth, I implicitly assumed that semantic representations are universal across languages: semantic features in my data sets were extracted from existing lexical resources such as WordNet, which is a common approach in computational linguistics and cognitive modeling. While this may be a reasonable approximation on the global scale of language learning, languages do not always encode meanings in a universal way (Bowerman & Choi, 2001). Therefore, for examining how a particular L2 linguistic unit (e.g., a construction) is acquired, it is important to account for potential differences between the semantic representations of the target unit in L1 vs. L2 (Beekhuizen & Stevenson, 2015).

Finally, the results reported on in this thesis come from computational simulations with a single model. I would strongly suggest that the results of the simulations presented here are compared in the future against similar results generated by other computational models (Barak, Goldberg, & Stevenson, in press, is an example of such a study). As for this thesis, it is the first serious exploration of how computational modeling can be applied to the study of bilingual and second language construction learning. I hope it lays the groundwork for future research in this field.

Bibliography

Abbot-Smith, K. & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275–290. doi:10.1515/tlr.2006.011

Abutalebi, J. & Green, D. W. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics*, *20*, 242–275. doi:10.1016/j.jneuroling.2006.10.003

Ågren, M., Granfeldt, J., & Thomas, A. (2014). Combined effects of age of onset and input on the development of different grammatical structures: A study of simultaneous and successive bilingual acquisition of French. *Linguistic Approaches to Bilingualism*, *4*, 462–493. doi:10.1075/lab.4.4.03agr

Akbik, A., Chiticariu, L., Danilevsky, M., Li, Y., Vaithyanathan, S., & Zhu, H. (2015). Generating high quality proposition banks for multilingual semantic role labeling. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long papers)* (pp. 397–407). Retrieved from http://www.aclweb.org/anthology/P15-1039

Akhtar, N. & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology*, *33*, 952–965. doi:10.1037/0012-1649.33.6.952

Albright, A. & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*, 119–161. doi:10.1016/s0010-0277(03)00146-x

Alishahi, A. & Fazly, A. (2010). Integrating syntactic knowledge into a model of cross-situational word learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 2452–2457). Austin, TX: Cognitive Science Society.

Alishahi, A. & Pyykkönen, P. (2011). The onset of syntactic bootstrapping in word learning: Evidence from a computational study. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Alishahi, A. & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*, 789–834. doi:10.1080/03640210801929287

Alishahi, A. & Stevenson, S. (2010). A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, *25*, 50–93. doi:10.1080/01690960902840279

Ambridge, B. (2013). How do children restrict their linguistic generalizations? An (un-)grammaticality judgment study. *Cognitive Science*, *37*, 508–543. doi:10.1111/cogs.12018

Ambridge, B., Bidgood, A., Pine, J. M., Rowland, C. F., & Freudenthal, D. (2015). Is passive syntax semantically constrained? Evidence from adult grammaticality judgment and comprehension studies. *Cognitive Science*, *40*, 1435–1459. doi:10.1111/cogs.12277

Ambridge, B., Bidgood, A., Twomey, K. E., Pine, J. M., Rowland, C. F., & Freudenthal, D. (2015). Preemption versus entrenchment: Towards a construction-general solution to the problem of the retreat from verb argument structure overgeneralization. *PLoS ONE*, *10*, 1–20. doi:10.1371/journal.pone.0123723

Ambridge, B. & Blything, R. P. (2015). A connectionist model of the retreat from verb argument structure overgeneralization. *Journal of Child Language*. Advance online publication. doi:10.1017/S0305000915000586

Ambridge, B. & Brandt, S. (2013). Lisa filled water into the cup: The roles of entrenchment, pre-emption and verb semantics in German speakers' L2 acquisition of English locatives. *Zeitschrift für Anglistik und Amerikanistik*, *61*, 245–263. doi:10.1515/zaa-2013-0304

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*, 239–273. doi:10.1017/s030500091400049x

Ambridge, B., Pine, J. M., & Rowland, C. F. (2012). Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition*, *123*, 260–279. doi:10.1016/j.cognition.2012.01.002

Ambridge, B., Pine, J. M., Rowland, C. F., Freudenthal, D., & Chang, F. (2014). Avoiding dative overgeneralisation errors: Semantics, statistics or both? *Language, Cognition and Neuroscience*, *29*, 218–243. doi:10.1080/01690965.2012.738300

Ambridge, B., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, *21*, 174–193. doi:10.1016/j.cogdev.2005.09.003

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261–295. doi:10.1016/s0022-5371(83)90201-3

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429. doi:10.1037/0033-295x.98.3.409

Ansaldo, A. I., Marcotte, K., Scherer, L., & Raboyeau, G. (2008). Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *Journal of Neurolinguistics*, *21*, 539–557. doi:10.1016/j.jneuroling.2008.02.001

Auer, P. (2014). Language mixing and language fusion: When bilingual talk becomes monolingual. In J. Besters-Dilger, C. Dermarkar, S. Pfänder, & A. Rabus (Eds.), *Congruence in contact-induced language change: Language families, typological resemblance, and perceived similarity* (pp. 294–334). Berlin: Walter de Gruyter.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511801686

Bai, X. & Xue, N. (2016). Generalizing the semantic roles in the Chinese Proposition Bank. *Language Resources and Evaluation*, *50*, 643–666. doi:10.1007/s10579-016-9342-y

Baker, C. F. (2012). FrameNet, current collaborations and future goals. *Language Resources and Evaluation*, *46*, 269–286. doi:10.1007/s10579-012-9191-2

Barak, L., Fazly, A., & Stevenson, S. (2012). Modeling the acquisition of mental state verbs. In R. Levi & D. Reitter (Eds.), *Proceedings of the 2012 Workshop on Cognitive Modeling and Computational Linguistics (CMCL-2012)* (pp. 1–10). Retrieved from http://www.aclweb.org/anthology/W13-2606

Barak, L., Fazly, A., & Stevenson, S. (2013a). Acquisition of desires before beliefs: A computational investigation. In J. Hockenmaier & S. Riedel (Eds.), *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL-2013)* (pp. 231–240). Retrieved from http://www.aclweb.org/anthology/W13-3525

Barak, L., Fazly, A., & Stevenson, S. (2013b). Modeling the emergence of an exemplar verb in construction learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1815–1820). Austin, TX: Cognitive Science Society.

Barak, L., Goldberg, A. E., & Stevenson, S. (in press). Comparing computational cognitive models of generalization in a language acquisition task. In K. Duh & X. Carreras (Eds.), *Proceedings of the 2012 Workshop on Cognitive Modeling and Computational Linguistics (CMCL-2012)*. Retrieved from http://www.cs.utoronto.ca/~suzanne/papers/EMNLP16BarakGoldbergStevenson.pdf

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. doi:10.1016/j.jml.2012.11.001

Bar-Shalom, E. & Snyder, W. (1996). Optional infinitives in Russian and their implications for the pro-drop debate. In M. Lindseth & S. Franks (Eds.), *Formal approaches to Slavic linguistics: The Indiana Meeting* (pp. 38–47). Ann Arbor, MI: Michigan Slavic Publications.

Bartoń, K. (2016). Package 'MuMIn': Multi-model inference. Retrieved from https://cran.r-project.org/web/packages/MuMIn/MuMIn.pdf

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. doi:10.18637/jss.v067.i01

Bates, E. & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 3–76). Cambridge: Cambridge University Press.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., … Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, *59*, 1–26. doi:10.1111/j.1467-9922.2009.00533.x

Beekhuizen, B. (2015). *Constructions emerging: A usage-based model of the acquisition of grammar*. Utrecht: LOT.

Beekhuizen, B. & Stevenson, S. (2015). Crowdsourcing elicitation data for semantic typologies. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 202–207). Austin, TX: Cognitive Science Society.

Behrens, H. (2006). The input–output relationship in first language acquisition. *Language and Cognitive Processes*, *21*, 2–24. doi:10.1080/01690960400001721

Belke, E., Brysbaert, M., Meyer, A. S., & Ghyselinck, M. (2005). Age of acquisition effects in picture naming: Evidence for a lexical-semantic competition hypothesis. *Cognition*, *96*, B45–B54. doi:10.1016/j.cognition.2004.11.006

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2013). From language-specific to shared syntactic representations: The influence of second language proficiency on syntactic sharing in bilinguals. *Cognition*, *127*, 287–306. doi:10.1016/j.cognition.2013.02.005

Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 109–127). New York, NY: Oxford University Press.

Blom, E. (2010). Effects of input on the early grammatical development of bilingual children. *International Journal of Bilingualism*, *14*, 422–446. doi:10.1177/1367006910370917

Blommaert, J. & Backus, A. M. (2013). *Repertoires revisited: 'Knowing language' in superdiversity*. London: King's College.

Blumenthal-Dramé, A. (2012). *Entrenchment in usage-based theories: What corpus data do and do not reveal about the mind*. Berlin: Walter De Gruyter. doi:10.1515/9783110294002

Blything, R. P., Ambridge, B., & Lieven, E. V. M. (2014). Children use statistics and semantics in the retreat from overgeneralization. *PLoS ONE*, *9*, 1–11. doi:10.1371/journal.pone.0110009

Boas, H. C. (2010). The syntax–lexicon continuum in Construction Grammar: A case study of English communication verbs. *Belgian Journal of Linguistics*, *24*, 54–82. doi:10.1075/bjl.24.03boa

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*, 1–47. doi:10.1037/0033-295x.89.1.1

Bolker, B. & R Development Core Team. (2016). Package 'bbmle': Tools for general maximum likelihood estimation. Retrieved from https://cran.r-project.org/web/packages/bbmle/bbmle.pdf

Born, R. (1985). Error types and negative transfer in compositions of third, fourth, and fifth semester German students. *Die Unterrichtspraxis/Teaching German*, *18*, 246–253. doi:10.2307/3530457

Bowerman, M. & Choi, S. (2001). Shaping meanings for language: Universal and language-specific in the acquisition of semantic categories. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 475–511). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620669.018

Boyd, J. K. & Goldberg, A. E. (2009). Input effects within a constructionist framework. *The Modern Language Journal*, *93*, 418–429. doi:10.1111/j.1540-4781.2009.00899.x

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., … Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, *2*, 597–620. doi:10.1007/s11168-004-7431-3

Broccias, C. (2012). The syntax-lexicon continuum. In T. Nevalainen & E. Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 735–747). New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780199922765.013.0061

Broeder, P. & Plunkett, K. (1994). Connectionism and second language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 421–453). London: Academic Press.

Brown, J. D. (1991). Statistics as a foreign language—Part 1: What to look for in reading statistical language studies. *TESOL Quarterly*, *25*, 569–586. doi:10.2307/3587077

Bryl, V., Tonelli, S., Giuliano, C., & Serafini, L. (2012). A novel FrameNet-based resource for the semantic web. In S. Ossowski & P. Lecca (Eds.), *Proceedings of the 27th Annual ACM Symposium on Applied Computing* (pp. 360–365). New York, NY: Association for Computing Machinery.

Brysbaert, M. & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition*, *13*, 992–1011. doi:10.1080/13506280544000165

Budanitsky, A. & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In S. Harabagiu, D. Moldovan, W. Peters, M. Stevenson, & Y. Wilks (Eds.), *Proceedings of the Workshop on WordNet and other lexical resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*. Retrieved from ftp://learning.cs.utoronto.ca/public_html/public_html/pub/gh/Budanitsky+Hirst-2001.pdf

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S., & Pinkal, M. (2006). The SALSA corpus: A German corpus resource for lexical semantics. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)* (pp. 969–974). Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer Science & Business Media. doi:10.1007/b97636

Bybee, J. (2003). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*, 711–733. doi:10.1353/lan.2006.0186

Bybee, J. (2008). Usage-based grammar and second language acquisition. In P. Robinson & N. C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 216–236). New York, NY: Routledge.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511750526

Bybee, J. & Thompson, S. (1997). Three frequency effects in syntax. In M. L. Juge & J. L. Moxley (Eds.), *Proceedings of the 23rd Annual Meeting of the Berkeley Linguistics Society: General session and parasession on pragmatics and grammatical structure* (pp. 378–388). doi:10.3765/bls.v23i1.1293

Cain, K. (2007). Syntactic awareness and reading ability: Is there any evidence for a special relationship? *Applied Psycholinguistics*, *28*, 679–694. doi:10.1017/s0142716407070361

Cappelle, B. (2006). Particle placement and the case for "allostructions". *Constructions*, *Special Volume 1*. Retrieved from http://journals.linguisticsociety.org/elanguage/constructions/article/view/22.html

Carreras, X. & Màrquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In C. Brew & D. Radev (Eds.), *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)* (pp. 152–164). Retrieved from http://www.aclweb.org/anthology/W/W05/W05-0620.pdf

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. doi:10.1037/0033-2909.132.3.354

Chang, N. (2008). *Constructing grammar: A computational model of the emergence of early constructions*. (Unpublished doctoral dissertation, University of California, Berkeley, CA). Retrieved from http://www1.icsi.berkeley.edu/%C3%B1chang/pubs/Chang-diss.pdf

Chater, N. & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*, 335–344. doi:10.1016/j.tics.2006.05.006

Clahsen, H. (2007). Psycholinguistic perspectives on grammatical representations. In S. Featherstone & W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base* (pp. 97–132). Berlin: Mouton de Gryuter.

Cook, V. (2002). Background to the L2 user. In V. Cook (Ed.), *Portraits of the L2 user* (pp. 1–28). Clevedon: Multilingual Matters.

Croft, W. (2001). *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198299554.001.0001

Cuppini, C., Magosso, E., & Ursino, M. (2013). Learning the lexical aspects of a second language at different proficiencies: A neural computational study. *Bilingualism: Language and Cognition*, *16*, 266–287. doi:10.1017/S1366728911000617

Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, *2*, 219–253. doi:10.1075/lab.2.3.01dab

De Angelis, G. & Selinker, L. (2001). Interlanguage transfer and competing linguistic systems in the multilingual mind. In J. Cenoz, B. Hufeisen, & U. Jessner (Eds.), *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives* (pp. 42–58). Clevedon: Multilingual Matters.

DeKeyser, R. M. (2008). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313–348). Malden, MA: Blackwell Publishing.

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning*, *63*, 52–67. doi:10.1111/j.1467-9922.2012.00737.x

Denhovska, N., Serratrice, L., & Payne, J. (2016). Acquisition of second language grammar under incidental learning conditions: The role of frequency and working memory. *Language Learning*, *66*, 159–190. doi:10.1111/lang.12142

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*, 108–127. doi:10.1016/j.newideapsych.2007.02.002

Dijkstra, T. & Van Heuven, W. J. (1998). The BIA model and bilingual word recognition. In J. Grainger & A. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 189–225). Mahwah, NJ: Lawrence Erlbaum Associates.

Dimroth, C., Rast, R., Starren, M., & Watorek, M. (2013). Methods for studying the acquisition of a new language under controlled input conditions: The VILLA project. *EUROSLA Yearbook*, *13*, 109–138.

Dittmar, M., Abbot-Smith, K., Lieven, E. V. M., & Tomasello, M. (2008). German children's comprehension of word order and case marking in causative sentences. *Child Development*, *79*, 1152–1167. doi:10.1111/j.1467-8624.2008.01181.x

Divjak, D. (2008). On (in)frequency and (un)acceptability. In B. Lewandowska-Tomaszczyk (Ed.), *Corpus Linguistics, Computer Tools and Applications – State of the Art* (pp. 213–233). Frankfurt: Peter Lang.

Divjak, D. & Caldwell-Harris, C. L. (2015). Frequency and entrenchment. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics* (pp. 53–75). Berlin: Walter de Gruyter. doi:10.1515/9783110292022-004

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*, 547–619. doi:10.1353/lan.1991.0021

Duran, M. S. & Aluísio, S. M. (2012). Propbank-Br: a Brazilian Treebank annotated with semantic role labels. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, ... P. Stelios (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)* (pp. 1862–1867). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/index.html

Durrant, P. & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*, 163–188. doi:10.1177/0267658309349431

Ellis, A. W. & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1103–1123. doi:10.1037/0278-7393.26.5.1103

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*, 143–188. doi:10.1017/s0272263102002024

Ellis, N. C. (2006a). Cognitive perspectives on SLA: The associative-cognitive CREED. *AILA Review*, *19*, 100–121. doi:10.1075/aila.19.08ell

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

Ellis, N. C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use. In S. T. Gries & D. Divjak (Eds.), *Frequency effects in language learning and processing* (Vol. 1, pp. 7–34). Berlin: Walter de Gruyter. doi:10.1515/9783110274059.7

Ellis, N. C. (2015). Implicit AND explicit learning of languages: Their dynamic interface and complexity. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 1–24). Amsterdam: John Benjamins Publishing Company. doi:10.1075/sibil.48

Ellis, N. C. & Cadierno, T. (2009). Constructing a second language: Introduction to the Special Section. *Annual Review of Cognitive Linguistics*, *7*, 111–139. doi:10.1075/arcl.7.05ell

Ellis, N. C. & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*, 187–221. doi:10.1075/arcl.7.08ell

Ellis, N. C. & Larsen-Freeman, D. (2006). Language emergence: Implications for Applied Linguistics—Introduction to the Special Issue. *Applied Linguistics*, *27*, 558–589. doi:10.1093/applin/aml028

Ellis, N. C. & Larsen-Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning*, *59*, 90–125. doi:10.1111/j.1467-9922.2009.00537.x

Ellis, N. C. & O'Donnell, M. B. (2012). Statistical construction learning: Does a Zipfian problem space ensure robust language learning. In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 265–304). Boston: De Gruyter Mouton. doi:10.1515/9781934078242.265

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014a). Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*, *4*, 405–431. doi:10.1075/lab.4.4.01ell

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014b). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality. *Cognitive Linguistics*, *25*, 55–98. doi:10.1515/cog-2013-0031

Ellis, N. C. & Sagarra, N. (2010). The bounds of adult language acquisition. *Studies in Second Language Acquisition*, *32*, 553–580. doi:10.1017/s0272263110000264

Ellis, R. (2004). Individual differences in second language learning. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 525–551). Malden, MA: Blackwell Publishing. doi:10.1002/9780470757000.ch21

Ervin-Tripp, S. M. (1974). Is second language like the first. *TESOL Quarterly*, *8*, 111–127. doi:10.2307/3585535

European Commission. (2012). *Europeans and their languages: Report. (Special Eurobarometer 386)*. Retrieved from http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. (Unpublished doctoral thesis, Universität Stuttgart, Stuttgart, Germany). Retrieved from http://elib.uni-stuttgart.de/bitstream/11682/2573/1/Evert2005phd.pdf

Fang, S.-Y., Zinszer, B. D., Malt, B. C., & Li, P. (2016). Bilingual object naming: A connectionist model. *Frontiers in Psychology*, *7*. Advance online publication. doi:10.3389/fpsyg.2016.00644

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*, 1017–1063. doi:10.1111/j.1551-6709.2010.01104.x

Fellbaum, C. & Baker, C. F. (2013). Comparing and harmonizing different verb classifications in light of a semantic annotation task. *Linguistics*, *51*, 707–728. doi:10.1515/ling-2013-0025

Flege, J. E. (2008). Give input a chance. In T. Piske & M. Young-Scholten (Eds.), *Input matters in SLA* (pp. 175–190). Bristol: Multilingual Matters.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, *41*, 78–104. doi:10.1006/jmla.1999.2638

Forsberg, F. & Fant, L. (2010). Idiomatically speaking: Effects of task variation on formulaic language in highly proficient user of L2 Spanish. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 47–70). London: Continuum.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585. doi:10.1111/j.1467-9280.2009.02335.x

French, R. M. (1998). A simple recurrent network model of bilingual memory. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 368–373). Mahwah, NJ: Lawrence Erlbaum Associates.

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2015). Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions. *Cognition*, *143*, 61–76. doi:10.1016/j.cognition.2015.05.027

Gass, S. M. (1987). The resolution of conflicts among competing systems: A bidirectional perspective. *Applied Psycholinguistics*, *8*, 329–350. doi:10.1017/S0142716400000369

Gasser, M. (1990). Connectionism and universals of second language acquisition. *Studies in Second Language Acquisition*, *12*, 179–199. doi:10.1017/s0272263100009074

Geeraerts, D., Grondelaers, S., & Bakema, P. (1994). *The structure of lexical variation: Meaning, naming, and context*. Berlin: Mouton de Gruyter. doi:10.1515/9783110873061

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, *115*, 43–67. doi:10.1016/j.actpsy.2003.11.002

Gillund, G. & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67. doi:10.1037/0033-295x.91.1.1

Gilquin, G. (2006). The place of prototypicality in corpus linguistics: Causation in the hot seat. In S. T. Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis* (pp. 159–192). Berlin: Mouton de Gruyter. doi:10.1515/9783110197709.159

Gilquin, G. (2010). Language production: A window to the mind? In H. Götzsche (Ed.), *Memory, mind and language* (pp. 89–102). Newcastle upon Tyne: Cambridge Scholars Publishing.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Goldberg, A. E. (1998). Patterns of experience in patterns of language. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure* (Vol. 1, pp. 203–219). Mahwah, NJ: Lawrence Erlbaum.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199268511.001.0001

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*, 289–316. doi:10.1515/cogl.2004.011

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2005). The role of prediction in construction-learning. *Journal of Child Language*, *32*, 407–426. doi:10.1017/s0305000904006798

Gomez, R. L. & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109–135. doi:10.1016/S0010-0277(99)00003-7

Gor, K. & Long, M. H. (2009). Input and second language processing. In W. Ritchie & T. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 445–472). Bingley: Emerald.

Grauberg, W. (1971). An error analysis in German of first-year university students. In G. Perren & J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 257–263). London: The University Press.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*, 67–81. doi:10.1017/s1366728998000133

Greven, S. & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, *97*, 773–789. doi:10.1093/biomet/asq042

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next... *International Journal of Corpus Linguistics*, *18*, 137–166. doi:10.1075/ijcl.18.1.09gri

Gries, S. T. (2015). More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics*, *26*, 505–536. doi:10.1515/cog-2014-0092

Gries, S. T. & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*, 228–255. doi:10.1111/lang.12119

Gries, S. T., Hampe, B., & Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, *16*, 635–676. doi:10.1515/cogl.2005.16.4.635

Gries, S. T. & Wulff, S. (2005). Do foreign language learners also have constructions? *Annual Review of Cognitive Linguistics*, *3*, 182–200. doi:10.1075/arcl.3.10gri

Gries, S. T. & Wulff, S. (2009). Psycholinguistic and corpus-linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, *7*, 163–186. doi:10.1075/arcl.7.07gri

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*, 1–27. doi:10.18637/jss.v017.i01

Grosjean, F. (1998). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, *1*, 131–149. doi:10.1017/S136672899800025X

Grosjean, F. (2010). *Bilingual: Life and reality*. Cambridge, MA: Harvard University Press. doi:10.4159/9780674056459

Grünloh, T., Lieven, E. V. M., & Tomasello, M. (2011). German children use prosody to identify participant roles in transitive sentences. *Cognitive Linguistics*, *22*, 393–419. doi:10.1515/cogl.2011.015

Haegeman, L. (1994). *Introduction to Government and Binding Theory* (2nd ed.). Oxford: Blackwell.

Hahn, U. & Ramscar, M. J. A. (2001). Conclusion: Mere similarity? In M. J. A. Ramscar & U. Hahn (Eds.), *Similarity and categorization* (pp. 257–272). Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198506287.003.0013

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., … Zhang, Y. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In S. Stevenson & X. Carreras (Eds.), *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009): Shared Task* (pp. 1–18). Retrieved from http://www.aclweb.org/anthology/W09-1201

Hajič, J., Martí, M. A., Marquez, L., Nivre, J., Štěpánek, J., Padó, S., & Straňák, P. (2012). 2009 CoNLL Shared Task Part 1: LDC2012T03 [DVD]. Philadelphia, PA: Linguistic Data Consortium.

Hall, J. K., Cheng, A., & Carlson, M. T. (2006). Reconceptualizing multicompetence as a theory of language knowledge. *Applied Linguistics*, *27*, 220–240. doi:10.1093/applin/aml013

Hamrick, P. & Rebuschat, P. (2011). How implicit is statistical learning? In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 362–382). Boston: De Gruyter Mouton. doi:10.1515/9781934078242.365

Hanson, S., Aroline, E., & Carlson, M. T. (2014). The roles of first language and proficiency in L2 processing of Spanish clitics: Global effects. *Language Learning*, *64*, 310–342. doi:10.1111/lang.12050

Hartsuiker, R. J., Beerts, S., Loncke, M., Desmet, T., & Bernolet, S. (2016). Cross-linguistic structural priming in multilinguals: Further evidence for shared syntax. *Journal of Memory and Language*, *90*, 14–30. doi:10.1016/j.jml.2016.03.003

Hernandez, A. E. & Li, P. (2007). Age of acquisition: Its neural and computational mechanisms. *Psychological Bulletin*, *133*, 638–650. doi:10.1037/0033-2909.133.4.638

Higby, E., Vargas, I., Pérez, S., Ramirez, W., Varela, E., Campoverde, G., … Obler, L. K. (2016). The bilingual's mental grammar system: Language-specific syntax is shared by both languages. Poster presented at Cognitive Neuroscience Society Annual Meeting, April 4, New York, NY.

Hockley, W. E. & Cristi, C. (1996). Tests of the separate retrieval of item and associative information using a frequency-judgment task. *Memory & Cognition*, *24*, 796–811. doi:10.3758/bf03201103

Hoff, E. & Naigles, L. R. (2002). How children use input to acquire a lexicon. *Child Development*, *73*, 418–433. doi:10.1111/1467-8624.00415

Hudson, R. (1995). Does English really have case? *Journal of Linguistics*, *31*, 375–392. doi:10.1017/s0022226700015644

Hulk, A. (2004). The acquisition of the French DP in a bilingual context. In P. Prévost & J. Paradis (Eds.), *Language acquisition and language disorders* (pp. 243–274). Amsterdam: John Benjamins Publishing Company. doi:10.1075/lald.32.12hul

Hulstijn, J. H. (1997). Second language acquisition research in the laboratory. *Studies in Second Language Acquisition*, *19*, 131–143. doi:10.1017/s0272263197002015

Isabelli, C. A. (2008). First Noun Principle or L1 Transfer Principle in SLA? *Hispania*, *91*, 465–478. doi:10.2307/20063732

Izura, C., Pérez, M. A., Agallou, E., Wright, V. C., Marín, J., Stadthagen-González, H., & Ellis, A. W. (2011). Age/order of acquisition effects and the cumulative learning of foreign words: A word training study. *Journal of Memory and Language*, *64*, 32–58. doi:10.1016/j.jml.2010.09.002

Jäkel, O. (2010). Working with authentic ELT discourse data: The Flensburg English Classroom Corpus. In R. Vogel & S. Sahel (Eds.), *NLK Proceedings 2010* (pp. 65–76). Bielefeld: Universität Bielefeld.

Janssen, B., Meir, N., Baker, A., & Armon-Lotem, S. (2015). On-line comprehension of Russian case cues in monolingual Russian and bilingual Russian-Dutch and Russian-Hebrew children. In E. Grillo & K. Jepson (Eds.), *Proceedings of the 39th Annual Boston University Conference on Language Development* (pp. 266–278). Somerville, MA: Cascadilla Press.

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in them interlanguage lexicon. *Language Learning*, *50*, 245–309. doi:10.1111/0023-8333.00118

Jarvis, S. & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. New York, NY: Routledge. doi:10.4324/9780203935927

Jia, G. & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics*, *24*. doi:10.1017/s0142716403000079

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution*, *5*, 944–946. doi:10.1111/2041-210x.12225

Jordens, P. (1977). Rules, grammatical intuitions and strategies in foreign language learning. *Interlanguage Studies Bulletin*, *2*, 5–76.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, *131*, 684–712. doi:10.1037/0033-2909.131.5.684

Kelly, M. H., Bock, J. K., & Keil, F. C. (1986). Prototypicality in a linguistic context: Effects on sentence structure. *Journal of Memory and Language*, *25*, 59–74. doi:10.1016/0749-596x(86)90021-5

Kemmer, S. & Barlow, M. (2000). Introduction: A usage-based conception of language. In S. Kemmer & M. Barlow (Eds.), *Usage-based models of language* (pp. 7–28). Stanford, CA: CSLI Publications.

Kempe, V. & MacWhinney, B. (1998). The acquisition of case marking by adult learners of Russian and German. *Studies in Second Language Acquisition*, *20*, 543–587. doi:10.1017/s0272263198004045

Kilborn, K. & Cooreman, A. (1987). Sentence interpretation strategies in adult Dutch–English bilinguals. *Applied Psycholinguistics*, *8*, 415–431. doi:10.1017/S0142716400000394

Kim, S., O'Grady, W., & Cho, S. (1995). The acquisition of case and word order in Korean: A note on the role of context. *Language Research*, *31*, 687–695.

Kipper Schuler, K. (2006). *VerbNet: A broad-coverage, comprehensive verb lexicon*. (Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, PA). Retrieved from http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf

Kiran, S., Grasemann, U., Sandberg, C., & Miikkulainen, R. (2013). A computational account of bilingual aphasia rehabilitation. *Bilingualism: Language and Cognition*, *16*, 325–342. doi:10.1017/s1366728912000533

Kozhevnikov, M. & Titov, I. (2013). Cross-lingual transfer of semantic role labeling models. In P. Fung & M. Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers)* (pp. 1190–1200). Retrieved from http://www.aclweb.org/anthology/P13-1117

Kroll, J. F., Bobb, S. C., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*, *128*, 416–430. doi:10.1016/j.actpsy.2008.02.001

Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition*, *9*, 119–135. doi:10.1017/s1366728906002483

Küchenhoff, H. & Schmid, H.-J. (2015). Reply to "More (old and new) misunderstandings of collostructional analysis: On Schmid & Küchenhoff" by Stefan Th. Gries. *Cognitive Linguistics*, *26*, 537–547. doi:10.1515/cog-2015-0053

Kunze, C. (2000). Extension and use of GermaNet, a lexical-semantic database. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. Retrieved from http://www.lrec-conf.org/proceedings/lrec2000/pdf/369.pdf

Küpper-Tetzel, C. E. (2014). Understanding the distributed practice effect: Strong effects on weak theoretical grounds. *Zeitschrift für Psychologie*, *222*, 71–81. doi:10.1027/2151-2604/a000168

Lambon Ralph, M. A. & Ehsan, S. (2006). Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: Empirical and computational evidence. *Visual Cognition*, *13*, 928–948. doi:10.1080/13506280544000110

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research*, *24*, 35–63. doi:10.1177/0267658307082981

Lee, J. F. & Malovrh, P. A. (2009). Linguistic and non-linguistic factors affecting OVS processing of accusative and dative case pronouns by advanced L2 learners of Spanish. In J. Collentine, M. García, B. Lafford, & F. Marcos-Marín (Eds.), *Selected proceedings of the 11th Hispanic Linguistics Symposium* (pp. 105–116). Somerville, MA: Cascadilla Proceedings Project.

Lenth, R. V. (2016). Using **lsmeans**. Retrieved from https://cran.r-project.org/web/packages/lsmeans/vignettes/using-lsmeans.pdf

Lewis, M. B., Gerhand, S., & Ellis, H. D. (2001). Re-evaluating age-of-acquisition effects: Are they simply cumulative-frequency effects? *Cognition*, *78*, 189–205. doi:10.1016/s0010-0277(00)00117-7

Lewis, M. & Steedman, M. (2013). Unsupervised induction of cross-lingual semantic relations. In T. Baldwin & A. Korhonen (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 681–692). Retrieved from http://www.aclweb.org/anthology/D13-1064

Li, P. (2009). Lexical organization and competition in first and second languages: Computational and neural mechanisms. *Cognitive Science*, *33*, 629–664. doi:10.1111/j.1551-6709.2009.01028.x

Li, P. (2013). Computational modeling of bilingualism: How can models tell us more about the bilingual mind? *Bilingualism: Language and Cognition*, *16*, 241–245. doi:10.1017/S1366728913000059

Li, P. & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. *Advances in Psycholody*, *134*, 59–85. doi:10.1016/s0166-4115(02)80006-1

Li, P., Zhao, X., & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science*, *31*, 581–612. doi:10.1080/15326900701399905

Lieven, E. V. M. (2010). Input and first language acquisition: Evaluating the role of frequency. *Lingua*, *120*, 2546–2556. doi:10.1016/j.lingua.2010.06.005

Lieven, E. V. M. & Tomasello, M. (2008). Children's first language acquisition from a usage-based perspective. In P. Robinson & N. C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 168–196). New York, NY: Routledge.

Linck, J. A. & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207. doi:10.1111/lang.12117

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Glenview, IL: Scott Foresman.

Litkowski, K. (2004). Senseval-3 task: Automatic labeling of semantic roles. In R. Mihalcea & P. Edmonds (Eds.), *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (pp. 9–12). Retrieved from http://web.eecs.umich.edu/~mihalcea/senseval/senseval3/proceedings/pdf/litkowski1.pdf

Long, M. H. (1990). The least a second language acquisition theory needs to explain. *TESOL Quarterly*, *24*, 649–666. doi:10.2307/3587113

Lopez de Lacalle, M., Laparra, E., Aldabe, I., & Rigau, G. (2016). Predicate Matrix: Automatically extending the semantic interoperability between predicate resources. *Language Resources and Evaluation*, *50*, 263–289. doi:10.1007/s10579-016-9348-5

MacWhinney, B. (1992). Transfer and competition in second language learning. *Advances in Psychology*, *83*, 371–390. doi:10.1016/s0166-4115(08)61506-x

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2008). A unified model. In P. Robinson & N. C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 341–371). New York, NY: Routledge.

MacWhinney, B. (2010). Computational models of child language learning: An introduction. *Journal of Child Language*, *37*, 477–485. doi:10.1017/s0305000910000139

MacWhinney, B. (2012). The logic of the unified model. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 211–227). London: Routledge. doi:10.4324/9780203808184.ch13

MacWhinney, B. (2015). Multidimensional SLA. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 19–48). Berlin: De Gruyter Mouton. doi:10.1515/9783110378528-004

Madan, C. R., Glaholt, M. G., & Caplan, J. B. (2010). The influence of item properties on association-memory. *Journal of Memory and Language*, *63*, 46–63. doi:10.1016/j.jml.2010.03.001

Maki, W. S. & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*, 598–603. doi:10.3758/pbr.15.3.598

Marcus, M., Kim, G., Marcinkiewicz, M. A., Macintyre, R., Bies, A., Ferguson, M., … Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. In C. J. Weinstein (Ed.), *Proceedings of the 1994 ARPA Human Language Technology Workshop* (pp. 114–119). San Francisco, CA: Morgan Kaufmann. doi:10.3115/1075812.1075835

Marian, V. & Spivey, M. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition*, *6*, 97–115. doi:10.1017/s1366728903001068

Màrquez, L., Carreras, X., Litkowski, K. C., & Stevenson, S. (2008). Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, *34*, 145–159. doi:10.1162/coli.2008.34.2.145

Matusevych, Y., Alishahi, A., & Backus, A. M. (2013). Computational simulations of second language construction learning. In V. Demberg & R. Levi (Eds.), *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)* (pp. 47–56). Retrieved from http://www.aclweb.org/anthology/W13-2606

Matusevych, Y., Alishahi, A., & Backus, A. M. (2015a). Distributional determinants of learning argument structure constructions in first and second language. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1547–1552). Austin, TX: Cognitive Science Society.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111. doi:10.1016/S0010-0277(01)00157-3

McDonald, J. L. (1987). Sentence interpretation in bilingual speakers of English and Dutch. *Applied Psycholinguistics*, *8*, 379–413. doi:10.1017/S0142716400000382

McDonough, K. & Nekrasova-Becker, T. (2012). Comparing the effect of skewed and balanced input on English as a foreign language learners' comprehension of the double-object dative construction. *Applied Psycholinguistics*, *35*, 419–442. doi:10.1017/s0142716412000446

McRae, K., Ferretti, T. R., & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, *12*, 137–176. doi:10.1080/016909697386835

Meisel, J. M. (1986). Word order and case marking in early child language. Evidence from simultaneous acquisition of two first languages: French and German. *Linguistics*, *24*, 123–184. doi:10.1515/ling.1986.24.1.123

Mermillod, M., Bonin, P., Méot, A., Ferrand, L., & Paindavoine, M. (2012). Computational evidence that frequency trajectory theory does not oppose but emerges from age-of-acquisition theory. *Cognitive Science*, *36*, 1499–1531. doi:10.1111/j.1551-6709.2012.01266.x

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society*, *7*, 283–284. doi:10.3758/bf03337190

Meunier, F. & Littre, D. (2013). Tracking learners' progress: Adopting a dual 'corpus cum experimental data' approach. *The Modern Language Journal*, *97*, 61–76. doi:10.1111/j.1540-4781.2012.01424.x

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 39–41. doi:10.1145/219717.219748

Miller, M. (1979). *The logic of language development in early childhood*. Berlin: Springer-Verlag.

Mimica, I., Sullivan, M., & Smith, S. (1994). An on-line study of sentence interpretation in native Croatian speakers. *Applied Psycholinguistics*, *15*, 237–261. doi:10.1017/S0142716400005348

Misyak, J. B. & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*, 302–331. doi:10.1111/j.1467-9922.2010.00626.x

Monaghan, J. & Ellis, A. W. (2002). What exactly interacts with spelling–sound consistency in word naming? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 183–206. doi:10.1037/0278-7393.28.1.183

Monner, D., Vatz, K., Morini, G., Hwang, S.-O., & DeKeyser, R. M. (2013). A neural network model of the effects of entrenchment and memory development on grammatical gender learning. *Bilingualism: Language and Cognition*, *16*, 246–265. doi:10.1017/S1366728912000454

Montrul, S. A. (2008). *Incomplete acquisition in bilingualism: Re-examining the age factor*. Amsterdam: John Benjamins Publishing Company. doi:10.1075/sibil.39

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

Morgenstern, A. & Parisse, C. (2012). The Paris corpus. *Journal of French Language Studies*, *22*, 7–12. doi:10.1017/S095926951100055X

Morgenstern, A., Parisse, C., Sekali, M., Bourdoux, F., & Caet, S. (2004). French Paris Corpus [Electronic database]. Retrieved from http://childes.psy.cmu.edu/data/French/Paris.zip

Moyer, A. (2004). *Age, accent, and experience in second language acquisition: An integrated approach to critical period inquiry*. Clevedon: Multilingual Matters.

Moyer, A. (2005). Formal and informal experiential realms in German as a foreign language: A preliminary investigation. *Foreign Language Annals*, *38*, 377–387. doi:10.1111/j.1944-9720.2005.tb02224.x

Muñoz, C. (2011). Input and long-term effects of starting age in foreign language learning. *International Review of Applied Linguistics in Language Teaching*, *49*, 113–133. doi:10.1515/iral.2011.006

Muñoz, C. & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, *44*, 1–35. doi:10.1017/s0261444810000327

Murre, J. M. (2005). Models of monolingual and bilingual language acquisition. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 154–169). New York, NY: Oxford University Press.

Naigles, L. R. & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25*, 95–120. doi:10.1017/s0305000997003358

Nastase, V., Nakov, P., Seaghdha, D. O., & Szpakowicz, S. (2013). Semantic relations between nominals. *Synthesis lectures on human language technologies*, *6*, 1–119. doi:10.2200/S00489ED1V01Y201303HLT019

Newell, A. & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Lawrence Erlbaum.

Ninio, A. (1999a). Model learning in syntactic development: Intransitive verbs. *International Journal of Bilingualism*, *3*, 111–130. doi:10.1177/13670069990030020301

Ninio, A. (1999b). Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language*, *26*, 619–653. doi:10.1017/s0305000999003931

Norris, J. M., Ross, S. J., & Schoonen, R. (2015). Improving second language quantitative research. *Language Learning*, *65*, 1–8. doi:10.1111/lang.12110

Oller, J. W. & Inal, N. (1971). A cloze test of English prepositions. *TESOL Quarterly*, *5*, 315–326. doi:10.2307/3585498

Onishi, K. H., Murphy, G. L., & Bock, K. (2008). Prototypicality in sentence production. *Cognitive Psychology*, *56*, 103–141. doi:10.1016/j.cogpsych.2007.04.001

Onnis, L. (2011). The potential contribution of statistical learning to second language acquisition. In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 203–235). Boston: De Gruyter Mouton. doi:10.1515/9781934078242.203

Ortega, L. (2015). Usage-based SLA: A research habitus whose time has come. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 353–373). Berlin: De Gruyter Mouton. doi:10.1515/9783110378528-004

Ortega, L. & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26–45. doi:10.1017/S0267190505000024

O'Shannessy, C. (2011). Competition between word order and case-marking in interpreting grammatical relations: A case study in multilingual acquisition. *Journal of Child Language*, *38*, 763–792. doi:10.1017/s0305000910000358

Palmberg, R. (1976). A select bibliography of error analysis and related topics. *Interlanguage Studies Bulletin*, *1*, 340–389.

Palmer, M. (2009). SemLink: Linking PropBank, VerbNet and FrameNet. In A. Rumshisky & N. Calzolari (Eds.), *Proceedings of the Fifth International Conference on Generative Approaches to the Lexicon* (pp. 9–15). Stroudsburg, PA: Association for Computational Linguistics.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*, 71–106. doi:10.1162/0891201053630264

Palmer, M., Gildea, D., & Xue, N. (2010). Semantic role labeling. *Synthesis lectures on human language technologies*, *3*, 1–103. doi:10.2200/S00239ED1V01Y200912HLT006

Palmer, M., Ryu, S., Choi, J., Yoon, S., & Jeon, Y. (2006). Korean Propbank: LDC2006T03 [Web Download]. Philadelphia, PA: Linguistic Data Consortium.

Paradis, J. & Grüter, T. (2014). Introduction to "Input and experience in bilingual development". In T. Grüter & J. Paradis (Eds.), *Input and experience in bilingual development* (pp. 1–14). Amsterdam: John Benjamins Publishing Company. doi:10.1075/tilar.13.01int

Paradis, J., Nicoladis, E., Crago, M., & Genesee, F. (2011). Bilingual children's acquisition of the past tense: A usage-based approach. *Journal of Child Language*, *38*, 554–578. doi:10.1017/s0305000910000218

Pavlenko, A. & Jarvis, S. (2002). Bidirectional transfer. *Applied linguistics*, *23*, 190–214. doi:10.1093/applin/23.2.190

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, *44*, 137–158. doi:10.1007/s10579-009-9101-4

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity - measuring the relatedness of concepts. In G. Ferguson & D. McGuinnes (Eds.), *Proceedings of the 19th National Conference on Artificial Intelligence* (pp. 1024–1025). Retrieved from http://www.aaai.org/Papers/AAAI/2004/AAAI04-160.pdf

Perek, F. (2015). *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives*. Amsterdam: John Benjamins Publishing Company. doi:10.1075/cal.17

Perfors, A. & Navarro, D. J. (2011). What Bayesian modelling can tell us about statistical learning: What it requires and why it works. In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 383–408). Boston: De Gruyter Mouton. doi:10.1515/9781934078242.383

Perruchet, P. & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233–238. doi:10.1016/j.tics.2006.03.006

Pinker, S. (2013). *Learnability and cognition: The acquisition of argument structure* (New ed.). Cambridge, MA: MIT Press.

Pinker, S. & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*, 456–463. doi:10.1016/s1364-6613(02)01990-3

Plant, C., Webster, J., & Whitworth, A. (2011). Category norm data and relationships with lexical frequency and typicality within verb semantic categories. *Behavior Research Methods*, *43*, 424–440. doi:10.3758/s13428-010-0051-y

Pléh, C., Jarovinskij, A., & Balajan, A. (1987). Sentence comprehension in Hungarian-Russian bilingual and monolingual preschool children. *Journal of Child Language*, *14*, 587–603. doi:10.1017/s0305000900010308

Poibeau, T., Villavicencio, A., Korhonen, A., & Alishahi, A. (2013). Computational modeling as a methodology for studying human language learning. In A. Villavicencio, T. Poibeau, A. Korhonen, & A. Alishahi (Eds.), *Cognitive aspects of computational language acquisition* (pp. 1–25). Heidelberg: Springer. doi:10.1007/978-3-642-31863-4_1

Protassova, E. (2004). Russian Protassova Corpus [Electronic database]. Retrieved from http://childes.psy.cmu.edu/data/Slavic/Russian/Protassova.zip

Pulvermüller, F., Cappelle, B., & Shtyrov, Y. (2013). Brain basis of meaning, words, constructions, and grammar. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of Construction Grammar* (pp. 397–416). Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780195396683.013.0022

Pulvermüller, F. & Knoblauch, A. (2009). Discrete combinatorial circuits emerging in neural networks: A mechanism for rules of grammar in the human brain? *Neural Networks*, *22*, 161–172. doi:10.1016/j.neunet.2009.01.009

Rappoport, A. & Sheinman, V. (2005). A second language acquisition model using example generalization and concept categories. In W. G. Sakas, A. Clark, J. Cussens, & A. Xanthos (Eds.), *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition* (pp. 45–52). Retrieved from http://aclweb.org/anthology/W/W05/W05-0506.pdf

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*, 595–626. doi:10.1111/lang.12010

Rebuschat, P. & Williams, J. N. (2012). Introduction: Statistical learning and language acquisition. In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 1–12). Boston: De Gruyter Mouton. doi:10.1515/9781934078242.1

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2014). Second language learner knowledge of verb–argument constructions: Effects of language transfer and typology. *The Modern Language Journal*, *98*, 952–975. doi:10.1111/modl.12149

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2015). Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions. In N. Groom, M. Charles, & S. John (Eds.), *Corpora, grammar and discourse: In honour of Susan Hunston* (pp. 43–71). Amsterdam: John Benjamins. doi:10.1075/scl.73.03rom

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605. doi:10.1016/0010-0285(75)90024-9

Rothman, J. (2011). L3 syntactic transfer selectivity and typological determinacy: The typological primacy model. *Second Language Research*, *27*, 107–127. doi:10.1177/0267658310386439

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., & Scheffczyk, J. (2006). FrameNet II: Extended Theory and Practice. Retrieved from https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. *Current Directions in Psychological Science*, *12*, 110–114. doi:10.1111/1467-8721.01243

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. doi:10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621. doi:10.1006/jmla.1996.0032

Salamoura, A. & Williams, J. N. (2007). Processing verb argument structure across languages: Evidence for shared representations in the bilingual lexicon. *Applied Psycholinguistics*, *28*, 627–660. doi:10.1017/s0142716407070348

Santesteban, M. & Costa, A. (2006). Does L1 syntax affect L2 processing? A study with highly proficient early bilinguals. In F. F. Beatriz & L. M. Itziar (Eds.), *Andolin Gogoan: Essays in honour of Professor Eguzkitza* (pp. 817–834). Vitoria-Gasteiz: Universidad del País Vasco / Euskal Herriko Unibertsitatea.

Saturno, J. (2015). Perceptual prominence and morphological processing in initial second language acquisition. In A. De Dominicis (Ed.), *pS-prominenceS: Prominences in linguistics: Proceedings of the International Conference* (pp. 76–95). Viterbo: DISUCOM PRESS.

Schaner-Wolles, C. (1989). Strategies in acquiring grammatical relations in German: Word order or case marking. *Folia Linguistica*, *23*, 131–156.

Schmid, H.-J. (2007). Entrenchment, salience, and basic levels. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of Cognitive Linguistics* (pp. 117–138). Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199738632.013.0005

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system. In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 101–133). Berlin: Walter de Gruyter. doi:10.1515/9783110226423.101

Schmid, H.-J. (in press). Introduction: A framework for understanding linguistic entrenchment and its psychological foundations in memory and automatization. In H.-J. Schmid (Ed.), *Entrenchment, memory and automaticity: The psychology of linguistic knowledge and language learning*.

Schmid, H.-J. & Küchenhoff, H. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics*, *24*. doi:10.1515/cog-2013-0018

Schmitz, K. (2006). Indirect objects and dative case in monolingual German and bilingual German/Romance language acquisition. In D. Hole, A. Meinunger, & W. Abraham (Eds.), *Datives and other cases: Between argument structure and event structure* (pp. 239–268). Amsterdam: John Benjamins Publishing Company. doi:10.1075/slcs.75.11sch

Shaoul, C., Baayen, R. H., & Westbury, C. F. (2014). *N*-gram probability effects in a cloze task. *The Mental Lexicon*, *9*, 437–472. doi:10.1075/ml.9.3.04sha

Shen, D. & Lapata, M. (2007). Using semantic roles to improve question answering. In J. Eisner (Ed.), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 12–21). Retrieved from http://www.aclweb.org/anthology/D/D07/D07-1002.pdf

Shi, L. & Mihalcea, R. (2005). Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In A. Gelbukh (Ed.), *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 100–111). Berlin: Springer. doi:10.1007/978-3-540-30586-6_9

Shook, A. & Marian, V. (2013). The bilingual language interaction network for comprehension of speech. *Bilingualism: Language and Cognition*, *16*, 304–324. doi:10.1017/S1366728912000466

Siyanova-Chanturia, A., Conklin, K., & Van Heuven, W. J. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 776–784. doi:10.1037/a0022531

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, *7*, 246–251. doi:10.1016/s1364-6613(03)00109-8

Smolík, F. (2015). Word order and information structure in Czech 3-and 4-year-olds' comprehension. *First Language*, *35*, 237–253. doi:10.1177/0142723715596098

Stefanowitsch, A. & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*, 209–243. doi:10.1075/ijcl.8.2.03ste

Stewart, N. & Ellis, A. W. (2008). Order of acquisition in learning perceptual categories: A laboratory analogue of the age-of-acquisition effect? *Psychonomic Bulletin & Review*, *15*, 70–74. doi:10.3758/pbr.15.1.70

Subirats, C. (2013). Frames, constructions, and metaphors in Spanish FrameNet. In I. Verdaguer, N. J. Laso, & D. Salazar (Eds.), *Biomedical English: A corpus-based approach* (pp. 185–210). Amsterdam: John Benjamins. doi:10.1075/scl.56.10sub

Sun, R., Jiang, J., Fan, Y., Hang, T., Tat-seng, C., & Kan, C. M.-y. (2005). Using syntactic and semantic relation analysis in question answering. In E. Voorhees & L. Buckland (Eds.), *Proceedings of the 14th Text REtrieval Conference*. Retrieved from http://trec.nist.gov/pubs/trec14/papers/nus.qa.pdf

Surdeanu, M., Harabagiu, S., Williams, J., & Aarseth, P. (2003). Using predicate-argument structures for information extraction. In E. W. Hinrichs & D. Roth (Eds.), *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 8–15). doi:10.3115/1075096.1075098

Surdeanu, M., Johansson, R., Màrquez, L., Meyers, A., & Nivre, J. (2009). 2008 CoNLL Shared Task Data: LDC2009T12 [Web Download]. Philadelphia, PA: Linguistic Data Consortium.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., & Nivre, J. (2008). The CoNLL 2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In A. Clark & K. Toutanova (Eds.), *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL-2008)* (pp. 159–177). Retrieved from http://www.aclweb.org/anthology/W08-2121

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152. doi:10.1017/s0305000900004608

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2004). Semantic generality, input frequency and the acquisition of syntax. *Journal of Child Language*, *31*, 61–99. doi:10.1017/s0305000903005956

Thomas, M. (1998). Distributed representations and the bilingual lexicon: One store or two? In J. A. Bullinaria, G. Houghton, & D. W. Glasspool (Eds.), *Proceedings*

of the Fourth Neural Computation and Psychology Workshop: Connectionist representations (pp. 240–253). doi:10.1007/978-1-4471-1546-5_19

Thomas, M. & van Heuven, W. J. (2005). Models of monolingual and bilingual language acquisition. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 202–225). New York, NY: Oxford University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In M. Hearst & M. Ostendorf (Eds.), *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 173–180). Retrieved from http://aclweb.org/anthology/N/N03/N03-1033.pdf

Tucker, G. R. (2002). A global perspective on bilingualism and bilingual education: Implications for New Jersey educators. *Journal of Iberian and Latin American Literary and Cultural Studies*, *2*(2). Retrieved from http://www.libraries.rutgers.edu/rul/projects/arachne/vol2_2tucker.html

Tyler, A. (2012). *Cognitive linguistics and second language learning: Theoretical basics and experimental evidence*. New York, NY: Routledge. doi:10.4324/9780203876039

Ullman, M. T. (2015). The declarative/procedural model. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 135–158). New York, NY: Routledge.

Unsworth, S. & Blom, E. (2010). Comparing L1 children, L2 children and L2 adults. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 201–222). Amsterdam: John Benjamins. doi:10.1075/lllt.27.12uns

van der Plas, L., Apidianaki, M., & Chen, C. (2014). Global methods for cross-lingual semantic role and predicate labelling. In J. Hajic & J. Tsujii (Eds.), *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical papers* (pp. 1279–1290). Retrieved from http://www.aclweb.org/anthology/C14-1121

VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Norwood, NJ: Ablex Publishing Corporation.

VanPatten, B. (2012). Input processing. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 268–281). London: Routledge. doi:10.4324/9780203808184.ch16

VanPatten, B. (2015a). Foundations of processing instruction. *International Review of Applied Linguistics in Language Teaching*, *53*, 91–109. doi:10.1515/iral-2015-0005

VanPatten, B. (2015b). Input processing in adult SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 113–134). New York, NY: Routledge.

Verhagen, V. & Mos, M. (2016). Stability of familiarity judgments: Individual variation and the invariant bigger picture. *Cognitive Linguistics*, *27*, 307–344. doi:10.1515/cog-2015-0063

von Stutterheim, C. (2004). German Caroline Corpus [Electronic database]. Retrieved from http://childes.psy.cmu.edu/data-xml/Germanic/German/Caroline.zip

Wasserscheidt, P. (2014). Constructions do not cross languages: On cross-linguistic generalizations of constructions. *Constructions and Frames*, *6*, 305–337. doi:10.1075/cf.6.2.07was

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*, 91–120. doi:10.1111/j.1467-9922.2012.00729.x

Weinreich, U. (1968). *Languages in Contact: Findings and Problems* (6th ed.). The Hague: Mouton Publishers. doi:10.1515/9783110802177

Wesche, M. B. & Paribakht, T. S. (2000). Reading-based exercises in second language vocabulary learning: An introspective study. *The Modern Language Journal*, *84*, 196–213. doi:10.1111/0026-7902.00062

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, *4*, 253–290. doi:10.1515/cllt.2008.011

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*, 165–209. doi:10.1016/j.cogpsych.2007.04.002

Wu, S. & Palmer, M. (2015). Improving Chinese-English PropBank alignment. In C. Zong & M. Sun (Eds.), *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 74–82). Retrieved from http://www.aclweb.org/anthology/W15-1012

Yang, J., Shu, H., McCandliss, B. D., & Zevin, J. D. (2013). Orthographic influences on division of labor in learning to read Chinese and English: Insights from computational modeling. *Bilingualism: Language and Cognition*, *16*, 354–366. doi:10.1017/S1366728912000296

Year, J. & Gordon, P. (2009). Korean speakers' acquisition of the English ditransitive construction: The role of verb prototype, input distribution, and frequency. *The Modern Language Journal*, *93*, 399–417. doi:10.1111/j.1540-4781.2009.00898.x

Yoshimura, Y. & MacWhinney, B. (2010). The use of pronominal case in English sentence interpretation. *Applied Psycholinguistics*, *31*, 619–633. doi:10.1017/s0142716410000160

You, L. & Liu, K. (2005). Building Chinese FrameNet database. In C. Zong & M. Sun (Eds.), *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'05)* (pp. 301–306). Beijing: Bupt Publishing House. doi:10.1109/NLPKE.2005.1598752

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., & Palmer, M. (2010). The Revised Arabic PropBank. In N. Xue & M. Poesio (Eds.), *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 222–226). Retrieved from http://www.aclweb.org/anthology/W10-1836

Zevin, J. D. & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*, 1–29. doi:10.1006/jmla.2001.2834

Zhao, X. & Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *International Journal of Bilingual Education and Bilingualism*, *13*, 505–524. doi:10.1080/13670050.2010.488284

Zhao, X. & Li, P. (2013). Simulating cross-language priming with a dynamic computational model of the lexicon. *Bilingualism: Language and Cognition*, *16*, 288–303. doi:10.1017/s1366728912000624

APPENDIX A

Features used for annotation

A.1 Verb semantic features

Feature	Associated event
ACTION	an action occurs
APPEAR	someone or something appears, created or produced
APPROACH	someone or something approaches physically
AUDIAL	a sound can be heard
CAUSAL	a change is caused
COMMUNICATIVE	verbal communication occurs
CONSUME	an eating or drinking event
CONTINUE	the current action is continued
DECORATE	something is being decorated
DRESS	a piece of clothes is put on or taken off
EXPLAIN	something is being explained
HURT	something causes physical pain
INTERACTION	an interaction between humans/animals
MANIPULATE	an object is being manipulated
MENTAL	a mental activity occurs
MOVE	something is moving or being moved
PERCEPTIVE	physical perception occurs
PHYSICAL	a physical action or state is described
PICTURE	a picture is being taken
PLAYFUL	a play occurs
POSSESS	an object changes its possessor
POSTURE	a physical posture is described
PRODUCE	something is created or produced
SEEK	something is being looked for
STATE	a current state of things is described

SWITCHING	something is switched on or off
TRANSPORT	public transport is being used
VISUAL	visual perception occurs
WALK	someone moves by walking
WILL	something is being wanted

A.2 Argument semantic role features

ABSTRACT
ACTING
ACTIVITY
AFFECTED
AFFECTING
ANIMATE
BECOME
BENEFICIARY
BODYPART
CAUSE
CHANGE
CHANGE-LOCATION
COMING
COMPREHENDING
CONCRETE
CONSUMED
CONSUMING
CONSUMPTION
DIRECTION
EMOTIONAL
GIVEN
GIVING
GOAL
INANIMATE
LISTENING
LOCATION
LOOKED
LOOKING
MANIPULATED
MANNER
MESSAGE
MOVE
MOVING
ORIGINATING
PATH
PERCEIVED
PERCEIVING
PERIOD
PLAYING
POSSESSING
PRODUCE
PRODUCED
PRODUCING
PROPERTY
PUT
PUTTING
READ
READING
RECEIVING
SEEING
SEEKING
SEEN
SITTING
SLEEP
SLEEPING
SOUGHT
SOUND
SOURCE
SPEAKING
STAYING
STRANGE
SUBSTANCE
TAKEN
TAKING
TEXTUAL
TICKLE
TOOL
TOPIC
VOLITIONAL
WALKING
WANTED

APPENDIX B

The formal model

B.1 Basic notations

In this appendix, the following notations are used: C – a construction; I – an argument structure instance, S – the feature set used by the model. The feature set consists of a number of features F_k:

$$S = \{F_1, F_2, F_3, ..., F_n\} \tag{B.1}$$

Each feature $F_k \in S$ is represented by multiple values in a data set, which we denote using the feature cardinality $|F_k|$. Some features by definition take single string values, while other features are defined as sets of elements, e.g.:

$$F_k = \{abandon, about, accept, ..., wrong\} \tag{B.2}$$

An instance I is, in fact, a unique combination of specific values (F_k^I) of all features $F_k \in S$:

$$I = \{F_1^I, F_2^I, F_3^I, ..., F_n^I\} \tag{B.3}$$

Each construction C also has a combination of values (F_k^C) of each $F_k \in S$ associated with it. However, each element $e \in F_k^C$ may occur in F_k^C multiple times. In other words, F_k^C is a multiset, and $|e_i|$ denotes the number of occurrences of e_i in F_k^C.

B.2 Learning

The learner processes instances one by one: N denotes the number of instances encountered by a certain moment of time. For a given instance I, the model looks for the most

probable construction C_{best}:

$$C_{best}(I) = \underset{C}{\operatorname{argmax}}\, P(C|I) \tag{B.4}$$

In (B.4), C ranges over all the constructions learned so far, as well as a potential new construction. The conditional probability in (B.4) can be estimated using the Bayes rule:

$$P(C|I) = \frac{P(C)P(I|C)}{P(I)} \tag{B.5}$$

Since the denominator in (B.5) is the same for all constructions, it can be dropped when comparing the probabilities of constructions:

$$P(C|I) \propto P(C)P(I|C) \tag{B.6}$$

In (B.6), there are two factors that determine which construction the new instance is added to: prior probability $P(C)$ and conditional probability $P(I|C)$. $P(C)$ is proportional to the frequency of C in the previously encountered input – in other words, the number of instances that C is based on:

$$P(C) = \frac{|C|}{N+1}, \tag{B.7}$$

For the potential (new) construction C_0 the frequency is initially assigned to 1, to avoid zero values in the multiplicative formula (6):

$$P(C_0) = \frac{|1|}{N+1}, \tag{B.8}$$

The conditional probability captures the similarity between the encountered instance I and a construction C. The features are assumed to be independent, and the overall similarity is a product of the similarities in terms of each feature. Considering (3), this can be noted as follows:

$$P(I|C) = \prod_{k=1}^{|F^I|} P\left(F_k^I \middle| C\right) \tag{B.9}$$

For features which take a single (string) value, such as the head verb, this probability is computed via a smoothed maximum likelihood estimator:

$$P\left(F_k^I \middle| C\right) = \frac{\left|\left\{F_k^I \middle| F_k^I \in F_k^C\right\}\right| + \lambda}{\left|F_k^C\right| + \lambda\,|F_k|} \tag{B.10}$$

In (B.10), $\left|\left\{F_k^I \middle| F_k^I \in F_k^C\right\}\right|$ denotes the number of occurrences of F_k^I in the multiset F_k^C, while λ is a smoothing parameter, whose value is set as described below in this Appendix. Note that for a new construction $\left|\left\{F_k^I \middle| F_k^I \in F_k^C\right\}\right| = |F_k^C| = 0$.

For features with a set value such as the semantic properties of the verb and the arguments, the method given in (B.10) is too strict, because any two sets of properties (e.g., lexical meaning properties) are unlikely to be fully identical, so that

$\left|\left\{F_k^I \middle| F_k^I \in F_k^C\right\}\right|$ would always equal zero. Therefore, the conditional probability for each set feature is calculated as follows (cf. Alishahi & Pyykkönen, 2011):

$$P\left(F_k^I \middle| C\right) = \left(\prod_{e \in F_k^I} P(e|C) \times \prod_{e \in F_k \setminus F_k^I} P(\neg e|C)\right)^{\frac{1}{|F_k|}} \tag{B.11}$$

In (B.11), F_k is the superset of all values of the respective feature in the data set. Then, $F_k \setminus F_k^I$ denotes all potential elements in F_k which do not occur in F_k^I. The probabilities $P(e|C)$ and $P(\neg e|C)$ are computed as given in (B.10), to obtain the probability of each element e occurring or not occurring, respectively, in C.

B.3 Testing

At certain intervals, the language proficiency (L2 proficiency, in our case) of the model is tested on a number of test tasks. Each task contains in total T test instances. To create a test instance I_{test}, a value of a single feature F_x in I is masked (F_x^I), so that the resulting I_{test} is incomplete:

$$I_{test} = I \setminus F_x^I \tag{B.12}$$

The model, then, has to predict all the values that could be used in place of F_x^I, and their probabilities, given I_{test}. We denote the enumerated set of predicted values as $F_x^{predicted}$. The probability of each $F_x^j \in F_x^{predicted}$ is calculated as follows:

$$P\left(F_x^j \middle| I_{test}\right) = \sum_C P\left(F_x^j \middle| C\right) P(C|I_{test}) \tag{B.13}$$

The right part in (B.13) is the sum over all acquired constructions. $P\left(F_x^j \middle| C\right)$ is computed as given in (B.10), while $P(C|I_{test})$, again, can be transformed using the Bayes rule:

$$P(C|I_{test}) = \frac{P(C)P(I_{test}|C)}{P(I_{test})} \tag{B.14}$$

Dropping the constant denominator in (B.14) yields:

$$P(C|I_{test}) \propto P(C)P(I_{test}|C) \tag{B.15}$$

The two probabilities in the right part of (B.15) are computed using equations (B.7) and (B.9), respectively.

The model's accuracy in a test task is computed differently for single-value features and for set-value features. For single-value features, the original value F_x^I is looked up in the enumerated set of predicted values $F_x^{predicted}$, and the probability of F_x^I in this set is used as the model's accuracy for a specific instance in the task:

$$Accuracy(F_x^I, F_x^{predicted}) = P\left(F_x^I \middle| F_x^{predicted}\right) \tag{B.16}$$

The overall accuracy in the task is the average over all the instances:

$$OverallAccuracy(T) = \frac{1}{T}\sum_{j=1}^{T} Accuracy\left(F_{xj}^I, F_{xj}^{predicted}\right) \tag{B.17}$$

For set-value features, the accuracy for a single test instance is estimated by comparing the enumerated set $F_x^{predicted}$ to the original set value F_x^I. This is done by using average precision (AP), a standard measure in information retrieval, where a set of relevant items are expected to appear at the top of a ranked list of results. The average precision is usually defined via so called precision at a rank k:

$$Precision\left(k, F_x^I, F_x^{predicted}\right) = \frac{1}{k}\sum_{j=1}^{k} 1_{F_x^{predicted}}\left(F_{xj}^I\right), \tag{B.18}$$

where $1_{F_x^{predicted}}$ is a characteristic function of the set $F_x^{predicted}$, the image of $1_{F_x^{predicted}}$ is $\{0,1\}$. Given (B.18), the AP (and, respectively, the accuracy) is defined as follows:

$$AP\left(F_x^I, F_x^{predicted}\right) = \frac{1}{|F_x^I|}\sum_{k=1}^{\left|F_x^{predicted}\right|} Precision\left(k, F_x^I, F_x^{predicted}\right) \times 1_{F_x^{predicted}}\left(F_{xk}^I\right) \tag{B.19}$$

Again, the overall accuracy in the task is the average over all the instances, as given in (B.17).

B.4 Parameter setting

The model has a smoothing parameter λ mentioned in equation (B.10). It determines the default probability of F_k^I in a construction C when $\left|\left\{F_k^I \middle| F_k^I \in F_k^C\right\}\right| = |F_k^C| = 0$. The value of λ is determined empirically: its lower bound depends on the numbers of values of all features F_k in the data set and can be approximated as $\prod_k \frac{1}{F_k}$, which in our case equals to 10^{-17}. Setting λ to 10^{-17} would likely result in creating a new construction for each novel instance. To ensure this is not the case, we set λ to a moderate value of 10^{-9}. This way, the number of constructions formed by the model at the end of learning varied from 89 to 210, depending on the experiment, with an average of 158. A more elaborated explanation of the parameter setting is provided by Alishahi and Stevenson (2008, Appendix B).

The version of the model described in chapter 5 has an extra parameter w, which serves as a weight factor for each value of all symbolic features. An important function of this parameter is that it determines how easily the model can swap arguments to consider alternative argument orders: recall examples (46–47) in chapter 5.3.2. When the model encounters a new instance, it can either put it into a new cluster C_0 or into an

existing cluster C_x. Based on the equations in section 5.3.2, the probabilities of these two options in this version of the model are computed as provided in (B.20–B.21).

$$P(C_0|I) \propto \frac{1}{N+1} \prod_{k=1}^{|FD^i|} \frac{1}{|FD_k|} \left(\prod_{k=1}^{|FS^i|} \frac{1}{|FS_k|} \right)^w \tag{B.20}$$

$$P(C_x|I) \propto \frac{|x|}{N+1} \prod_{k=1}^{|FD^i|} \left(\prod_{e \in FD_k^i} P(e|C) \times \prod_{e \in FD_k \setminus FD_k^i} P(\neg e|C) \right)^{\frac{1}{|FD_k|}} \times \\ \times \left(\prod_{k=1}^{|FS^i|} \frac{\left|\left\{FS_k^i | FS_k^i \in FS_k^C\right\}\right| + \lambda}{\left|FS_k^C\right| + \lambda |FS_k|} \right)^w \tag{B.21}$$

For the ease of the future computation in this particular estimation of w we assume that for any i in (B.21) holds $FD_k^i = FD_k$, in which case (B.21) can be rewritten as follows in (B.22):

$$P(C_x|I) \propto \frac{|x|}{N+1} \prod_{k=1}^{|FD^i|} \frac{\left|\left\{FD_k^i | FD_k^i \in FD_k^C\right\}\right| + \lambda}{\left|FD_k^C\right| + \lambda |FD_k|} \times \\ \times \left(\prod_{k=1}^{|FS^i|} \frac{\left|\left\{FS_k^i | FS_k^i \in FS_k^C\right\}\right| + \lambda}{\left|FS_k^C\right| + \lambda |FS_k|} \right)^w \tag{B.22}$$

To ensure that the argument swapping functions in a sensible manner, we need to define two conditions:

1. The values of all the features (FD and FS), except the "argument position" features, are the same in the new instance and a cluster C_x. In this case, we would like the arguments to be swapped, and the new instance to be added to C_x, even if C_x consists of only one instance, $|C_x| = 1$. In other words,

$$P(C_x|I) > P(C_0|I). \tag{B.23}$$

2. The values of some important features, for example "preposition", differ in the new instance and a cluster C_x. In this case, we do not want to force the argument swap. Instead, the new instance should be put into a new cluster, even if C_x is highly entrenched, let us take $|C_x| = 100$. In other words,

$$P(C_x|I) < P(C_0|I). \tag{B.24}$$

Based on the two conditions, we can compute the extrema of w by solving the two inequalities (B.23–B.24):

$$\frac{ln\left(\frac{\lambda\prod_{k=1}^{|FD^i|}|FD_k|}{2\lambda+1}\right)}{ln\left(\frac{(3\lambda+1)^3}{\lambda^3\prod_{k=1}^{|FS^i|}|FS_k|}\right)} < w < \frac{ln\left(\prod_{k=1}^{|FD^i|}|FD_k|\right)}{ln\left(\frac{(3\lambda+1)^3}{\lambda^3\prod_{k=1}^{|FS^i|}|FS_k|}\right)} \tag{B.25}$$

Substituting the actual values from our data sets into (B.25), and considering that $\lambda = 10^{-14}$ in this set of simulations, we can establish that for any data set containing a pair of language samples from our manually annotated corpus $0.05 < w < 0.7$. Running simulations with various values of w within this range yields an acceptable value of 0.2, which is used throughout the reported simulations.

Summary

Learning foreign languages is not at all uncommon in today's global world. For scholars who carry out experimental or observational studies with bilinguals or second language learners, obtaining a sample of participants seems to be a relatively easy task. Yet if one takes a closer look at the issue, (s)he will discover substantial differences between learners even in a seemingly homogeneous group. Each person has particular learning experiences, abilities, motivations, etc. – the number of variables is huge. This variability creates a problem for traditional research on bilingualism and second language acquisition with human participants. Chapter 1 of my thesis introduces this and related theoretical issues.

In various fields, such as cognitive science or first language acquisition, the method of cognitive computational modeling has been successfully used to overcome this problem. At the same time, computational modeling studies in the fields of bilingualism and second language acquisition have been scarce. Computational models help researchers to eliminate unwanted sources of variation and facilitate a focus on the phenomena of interest. Clearly, computational models are not a replacement for experimental or observational studies of humans. This is why any novel predictions based on computational simulation alone need to be verified with human participants. However, models are extremely useful when it comes to studying general cognitive mechanisms or tendencies common for all learners, irrespective of their personalities.

This thesis demonstrates how the method of cognitive computational modeling can be used in research on second language acquisition and bilingualism. I use a particular computational model which simulates the process of learning linguistic constructions (argument structure constructions, to be more precise) from bilingual input. Chapter 2 provides details on the type of input used in the simulations, and on the steps taken to prepare the input corpora. Two novel data sets are presented: one of them is based on a combination of features extracted from existing English and German linguistic resources, while the other one contains multilingual data (English, French, German, and Russian) manually annotated during this project.

The learning process simulated in my model does not replicate human learning in all its complexity. Instead, it instantiates a particular mechanism of statistical learning,

on which humans are believed to rely in their acquisition of languages. In simple terms, the mechanism of statistical learning consists in noticing regularities in how different units (words, syntactic patterns, etc.) co-occur with each other in the input, and in using these regularities to form linguistic representations in the mind. This is a sketch of how the computational model in this thesis works. After gaining constructional knowledge from the input this way, the model is tested on carefully designed linguistic tasks. This general approach is adopted here to study the role of several variables.

In chapter 3, I study how the distribution of verbs in the input and their semantic properties affect the choice of verbs in a given linguistic construction. This chapter is based on earlier experiments carried out with human participants, and the computational model is used to refine the existing account predicting the verb choice. For example, I demonstrate which distributional variables may be important, and which may not be.

Chapter 4 focuses on two variables often discussed in the field of second language acquisition: the amount of second language input and the time of its onset. These two variables are nearly always confounded in human speakers: the later a person starts to learn a foreign language, the less input (s)he will have received by a certain age. Computational modeling allows me to disentangle the effect of the two variables, and the simulation results predict that the late start is not necessarily associated with worse performance, when it comes to the knowledge of English and German argument structure constructions.

The phenomenon of cross-linguistic influence is central to second language learning, and it is studied in chapter 5. I propose a method of measuring the amount of cross-linguistic influence in the computational model. This method is then used to study the comprehension of linguistic cases (e.g., the accusative) in languages with relatively free word order: Russian and German. This study demonstrates how computational modeling can be used to test alternative theories explaining a particular type of linguistic behavior.

The reported studies are carried out within the usage-based framework, which is widely adopted in cognitive linguistics. Chapter 6 summarizes the studies and describes their theoretical and methodological implications. Overall, the work contributes to our understanding of the role that statistical learning plays as a mechanism in bilingual and second language acquisition, and the extent to which this mechanism accounts for various forms of linguistic behavior commonly observed in human participants.

List of publications

Journal articles (included in the thesis):

1. Matusevych, Y., Alishahi, A., & Backus, A. M. (2015b). The impact of first and second language exposure on learning second language constructions. *Bilingualism: Language and Cognition*. Advance online publication. doi:10.1017/S1366728915000607
2. Matusevych, Y., Alishahi, A., & Backus, A. M. (2016b). Modelling verb selection within argument structure constructions. *Language, Cognition and Neuroscience*. Advance online publication. doi:10.1080/23273798.2016.1200732
3. Matusevych, Y., Alishahi, A., & Backus, A. M. (2016a). A multilingual corpus of verb usages annotated with argument structure information. Manuscript submitted for publication.
4. Matusevych, Y., Alishahi, A., & Backus, A. M. (2016c). Quantifying cross-linguistic influence with a computational model: a study of case marking comprehension. Manuscript submitted for publication.

Related conference papers (not included in the thesis):

1. Matusevych, Y., Alishahi, A., & Backus, A. M. (2013). Computational simulations of second language construction learning. In V. Demberg & R. Levi (Eds.), *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)* (pp. 47–56). Retrieved from http://www.aclweb.org/anthology/W13-2606
2. Matusevych, Y., Alishahi, A., & Backus, A. M. (2014). Isolating second language learning factors in a computational study of bilingual construction acquisition. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 988–994). Austin, TX: Cognitive Science Society.
3. Matusevych, Y., Alishahi, A., & Backus, A. M. (2015a). Distributional determinants of learning argument structure constructions in first and second language. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1547–1552). Austin, TX: Cognitive Science Society.

Other work published during the PhD project:

1. Matusevych, Y., Backus, A. M., & Reynaert, M. (2013). Do we teach the real language?: An analysis of patterns in textbooks of Russian as a foreign language. *Dutch Journal of Applied Linguistics*, *2*, 224–241. doi:10.1075/dujal.2.2.07mat
2. Matusevych, Y., Alishahi, A., & Vogt, P. (2013). Automatic generation of naturalistic child–adult interaction data. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2996–3001). Austin, TX: Cognitive Science Society.

TiCC PhD Series

1. Pashiera Barkhuysen. Audiovisual prosody in interaction. Promotores: M. G. J. Swerts, E. J. Krahmer. Tilburg, 3 October 2008.
2. Ben Torben-Nielsen. Dendritic morphology: Function shapes structure. Promotores: H. J. van den Herik, E. O. Postma. Copromotor: K. P. Tuyls. Tilburg, 3 December 2008.
3. Hans Stol. A framework for evidence-based policy making using IT. Promotor: H. J. van den Herik. Tilburg, 21 January 2009.
4. Jeroen Geertzen. Dialogue act recognition and prediction. Promotor: H. Bunt. Copromotor: J. M. B. Terken. Tilburg, 11 February 2009.
5. Sander Canisius. Structured prediction for natural language processing. Promotores: A. P. J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.
6. Fritz Reul. New architectures in computer chess. Promotor: H. J. van den Herik. Copromotor: J. W. H. M. Uiterwijk. Tilburg, 17 June 2009.
7. Laurens van der Maaten. Feature extraction from visual data. Promotores: E. O. Postma, H. J. van den Herik. Copromotor: A. G. Lange. Tilburg, 23 June 2009 (cum laude).
8. Stephan Raaijmakers. Multinomial language learning. Promotores: W. Daelemans, A. P. J. van den Bosch. Tilburg, 1 December 2009.
9. Igor Berezhnoy. Digital analysis of paintings. Promotores: E. O. Postma, H. J. van den Herik. Tilburg, 7 December 2009.
10. Toine Bogers. Recommender systems for social bookmarking. Promotor: A. P. J. van den Bosch. Tilburg, 8 December 2009.
11. Sander Bakkes. Rapid adaptation of video game AI. Promotor: H. J. van den Herik. Copromotor: P. Spronck. Tilburg, 3 March 2010.
12. Maria Mos. Complex lexical items. Promotor: A. P. J. van den Bosch. Copromotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010.
13. Marieke van Erp. Accessing natural history. Discoveries in data cleaning, structuring, and retrieval. Promotor: A. P. J. van den Bosch. Copromotor: P. K. Lendvai. Tilburg, 30 June 2010.
14. Edwin Commandeur. Implicit causality and implicit consequentiality in language comprehension. Promotores: L. G. M. Noordman, W. Vonk. Copromotor: R. Cozijn. Tilburg, 30 June 2010.

15. Bart Bogaert. Cloud content contention. Promotores: H. J. van den Herik, E. O. Postma. Tilburg, 30 March 2011.
16. Xiaoyu Mao. Airport under control. Promotores: H. J. van den Herik, E. O. Postma. Copromotores: N. Roos, A. Salden. Tilburg, 25 May 2011.
17. Olga Petukhova. Multidimensional dialogue modelling. Promotor: H. Bunt. Tilburg, 1 September 2011.
18. Lisette Mol. Language in the hands. Promotores: E. J. Krahmer, A. A. Maes, M. G. J. Swerts. Tilburg, 7 November 2011 (cum laude).
19. Herman Stehouwer. Statistical language models for alternative sequence selection. Promotores: A. P. J. van den Bosch, H. J. van den Herik. Copromotor: M. M. van Zaanen. Tilburg, 7 December 2011.
20. Terry Kakeeto-Aelen. Relationship marketing for SMEs in Uganda. Promotores: J. Chr. van Dalen, H. J. van den Herik. Copromotor: B. A. Van de Walle. Tilburg, 1 February 2012.
21. Suleman Shahid. Fun & Face: Exploring non-verbal expressions of emotion during playful interactions. Promotores: E. J. Krahmer, M. G. J. Swerts. Tilburg, 25 May 2012.
22. Thijs Vis. Intelligence, politie en veiligheidsdienst: Verenigbare grootheden? Promotores: T. A. de Roos, H. J. van den Herik, A. C. M. Spapens. Tilburg, 6 June 2012.
23. Nancy Pascall. Engendering technology empowering women. Promotores: H. J. van den Herik, M. Diocaretz. Tilburg, 19 November 2012.
24. Agus Gunawan. Information access for SMEs in Indonesia. Promotor: H. J. van den Herik. Copromotores: M. Wahdan, B. A. Van de Walle. Tilburg, 19 December 2012.
25. Giel van Lankveld. Quantifying individual player differences. Promotores: H. J. van den Herik, A. R. Arntz. Copromotor: P. Spronck. Tilburg, 27 February 2013.
26. Sander Wubben. Text-to-text generation using monolingual machine translation. Promotores: E. J. Krahmer, A. P. J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.
27. Jeroen Janssens. Outlier selection and one-class classification. Promotores: E. O. Postma, H. J. van den Herik. Tilburg, 11 June 2013.
28. Martijn Balsters. Expression and perception of emotions: The case of depression, sadness and fear. Promotores: E. J. Krahmer, M. G. J. Swerts, A. J. J. M. Vingerhoets. Tilburg, 25 June 2013.
29. Lisanne van Weelden. Metaphor in good shape. Promotor: A. A. Maes. Copromotor: J. Schilperoord. Tilburg, 28 June 2013.
30. Ruud Koolen. Need I say more? On overspecification in definite reference. Promotores: E. J. Krahmer, M. G. J. Swerts. Tilburg, 20 September 2013.
31. J. Douglas Mastin. Exploring infant engagement, language socialization and vocabulary development: A study of rural and urban communities in Mozambique. Promotor: A. A. Maes. Copromotor: P. A. Vogt. Tilburg, 11 October 2013.
32. Philip C. Jackson. Jr. Toward human-level artificial intelligence: Representation and computation of meaning in natural language. Promotores: H. C. Bunt, W. P. M. Daelemans. Tilburg, 22 April 2014.
33. Jorrig Vogels. Referential choices in language production: The role of accessibility. Promotores: A. A. Maes, E. J. Krahmer. Tilburg, 23 April 2014.

34. Peter de Kock. Anticipating criminal behaviour. Promotores: H. J. van den Herik, J. C. Scholtes. Copromotor: P. Spronck. Tilburg, 10 September 2014.
35. Constantijn Kaland. Prosodic marking of semantic contrasts: Do speakers adapt to addressees? Promotores: M. G. J. Swerts, E. J. Krahmer. Tilburg, 1 October 2014.
36. Jasmina Marić. Web communities, immigration and social capital. Promotor: H. J. van den Herik. Copromotores: R. Cozijn, M. Spotti. Tilburg, 18 November 2014.
37. Pauline Meesters. Intelligent blauw. Promotores: H. J. van den Herik, T. A. de Roos. Tilburg, 1 December 2014.
38. Mandy Visser. Better use your head: How people learn to signal emotions in social contexts. Promotores: M. G. J. Swerts, E. J. Krahmer. Tilburg, 10 June 2015.
39. Sterling Hutchinson. How symbolic and embodied representations work in concert. Promotores: M. M. Louwerse, E. O. Postma. Tilburg, 30 June 2015.
40. Marieke Hoetjes. Talking hands: Reference in speech, gesture and sign. Promotores: E. J. Krahmer, M. G. J. Swerts. Tilburg, 7 October 2015.
41. Elisabeth Lubinga. Stop HIV. Start talking?: The effects of rhetorical figures in health messages on conversations among South African adolescents. Promotores: A. A. Maes, C. J. M. Jansen. Tilburg, 16 October 2015.
42. Janet Bagorogoza. Knowledge management and high performance: The Uganda financial institutions models for HPO. Promotores: H. J. van den Herik, B. van der Walle, Tilburg, 24 November 2015.
43. Hans Westerbeek. Visual realism: Exploring effects on memory, language production, comprehension, and preference. Promotores: A. A. Maes, M. G. J. Swerts. Copromotor: M. A. A. van Amelsvoort. Tilburg, 10 February 2016.
44. Matje van de Camp. A link to the past: Constructing historical social networks from unstructured data. Promotores: A. P. J. van den Bosch, E. O. Postma. Tilburg, 2 March 2016.
45. Annemarie Quispel. Data for all: How designers and laymen use and evaluate information visualizations. Promotor: A. A. Maes. Copromotor: J. Schilperoord. Tilburg, 15 June 2016.
46. Rick Tillman. Language matters: The influence of language and language use on cognition. Promotores: M. M. Louwerse, E. O. Postma. Tilburg, 30 June 2016.
47. Ruud Mattheij. The eyes have it. Promotores: E. O. Postma, H. J. van den Herik. Copromotor: P. H. M. Spronck. Tilburg, 5 October 2016.
48. Marten Pijl. Tracking of human motion over time. Promotores: E. H. L. Aarts, M. M. Louwerse. Copromotor: J. H. M. Korst. Tilburg, 14 December 2016.
49. Yevgen Matusevych. Learning constructions from bilingual exposure: Computational studies of argument structure acquisition. Promotor: A. M. Backus. Copromotor: A. Alishahi. Tilburg, 19 December 2016.