Expanding the Pipeline: A Prolegomenon to Modeling Multilingual Grammars in LFG

Lara Schwarz
The Pennsylvania State University

Michael T. Putnam
The Pennsylvania State University

Proceedings of the LFG'17 Conference

University of Konstanz

Miriam Butt, Tracy Holloway King (Editors)

2017

CSLI Publications

pages 368–386

http://csli-publications.stanford.edu/LFG/2017

Keywords: bilingualism, code-switching, architecture, Optimality Theory

Schwarz, Lara, & Putnam, Michael T. (2017). Expanding the Pipeline: A Prolegomenon to Modeling Multilingual Grammars in LFG. In Butt, Miriam, & King, Tracy Holloway (Eds.): *Proceedings of the LFG'17 Conference, University of Konstanz* (pp. 368–386). Stanford, CA: CSLI Publications.

Abstract

In light of evidence from cognitive neuroscience that both source grammars are simultaneously active in the mind of a bilingual, we discuss the ramifications this has on the modeling of outputs from bilingual grammars, especially those that contain elements from multiple source grammars (i.e., *code-switching*). Here we provide a sketch of the architectural assumptions necessitated in light of these findings. To best model these structures, we introduce an expanded pipeline architecture that builds upon the foundation of previous work by Asudeh and Toivonen (2015). Similar to previous work integrating violable constraints from Optimality Theory (OT) (Prince & Smolensky, 2008) into LFG's parallel correspondence architecture (Bresnan, 2000; Sells, 2001a, 2001b), we augment this architecture with gradient, probabilistic mapping functions between the independent levels of grammar as initially suggested by Goldrick, Putnam, and Schwarz (2016a).

1 Introduction

In this paper, we introduce and discuss adjustments to the pipeline architecture Asudeh and Toivonen (2015), which is common in some versions of LFG, in order to improve its applicability to both monolingual and bi-/multilingual grammars. Since Grosjean (1989), it has become widely acknowledged that bilinguals are not the sum of two monolinguals. This poses an interesting challenge for the field of linguistics. How do we reconcile monolingual production with bilingual–or multilingual[1]–production under the assumption that both mono- and multilinguals are utilizing identical resources, when the two can differ, and when some linguistic phenomena can only be observed in bilingual data? Here, we take a closer look at a small sample of instances of bilingual code-switching, and discuss the implications of such utterances on the pipeline.

Code-switching, or code-mixing, is the phenomenon of bilingual dialogue in which speakers switch between their languages. Previous research has firmly established that code-switching does not occur haphazardly, and that the resultant structures are regulated by linguistic and cognitive constraints (Aguirre Jr., 1980; Poplack, 1980). Additional evidence also provides support for the position that rather than unique, "third grammar" constraints for code-switching, these outputs can best be understood as the result of the interaction between a bilingual's source grammars and knowledge about language in general (Cantone, 2005; Lederberg & Morales, 1985; MacSwan, 2014a, 2014b; Mahootian, 1993; Pfaff, 1979). As such, we operate on the assumption that bilinguals utilize the same faculties as monolinguals, and as linguists we seek to improve our formal models to account for

[0]We would like to thank the reviewers for their comments, as well as the attendees of LFG17 for their valuable feedback on an earlier version of this paper. Special thanks go to Stephen Jones for an insightful discussion leading to a significantly revised and improved final product. The usual disclaimers apply.

[1]We will focus on bilingual production in this paper, yet see no reason why the principles put forth here should not extend to multilinguals.

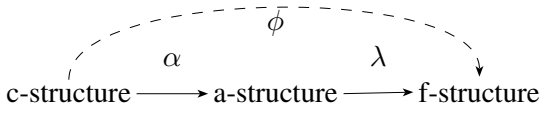

Figure 1: Classic c- to a- to f-structure pipeline

these grammars. An important challenge in this research program is to develop a model that does not assume operations and constraints that are specific to bilingual grammars, while simultaneously allowing for production and comprehension to be gradient, rather than strictly categorical.

Although LFG's existing architecture comes close to achieving this goal in its current state, we introduce minor adjustments to bring it in line with recent psycholinguistic insight into the cognitive architecture. Our focus here will be on supplementing the existing Correspondence Architecture with a means of introducing gradience in the mapping functions and permitting competition to extend through the pipeline.

In the pipeline version of the Correspondence Architecture, information flows from form to meaning through a series of mapping functions. Conventionally, LFG represents a sentence as a constituent structure, a functional structure, and the ϕ mapping function, through which the two structures correspond. According to Butt, Dalrymple, and Frank (1997), the role of the argument structure is incorporated into the representation of the sentence, and ϕ became the sum of two new functions: λ, the correspondence between argument and functional structure, and α, the correspondence between constituent and argument structure (Figure 1). Typically, a one-to-one correspondence is assumed, and any optionality is quickly eliminated. For instance, a ditransitive verb that participates in the dative alternation has two possible c-structures. While the semantics may be the same for both the double-object and oblique constructions, the argument structures are not, and once one a-structure has been selected, only one c-structure is possible.

Here we augment the version of the pipeline correspondence architecture introduced immediately above with violable, weighted mapping functions connecting independent levels of representation. We propose that, in situations of optionality, all options remain residually active. While one option may be preferred and subsequently gain momentum through the pipeline, the dispreferred options can still impact the process and potentially be selected as the optimal output. In this paper, we discuss the need for stochastic mapping functions, in the spirit of similar explorations of integrating LFG with Stochastic OT (Bresnan, Cueni, Nikitina, & Baayen, 2007) and demonstrate how Gradient Symbolic Computation (GSC) (Smolensky, Goldrick, & Mathis, 2014) can fill that need, and bring us closer to a Correspondence Architecture that can accommodate both monolingual and bilingual production. Adopting the adjustments to the pipeline we propose in this paper, brings the model in line with recent theories of language processing such as Christiansen and Chater's (2016) *Chunk-and-Pass* processing strategy.

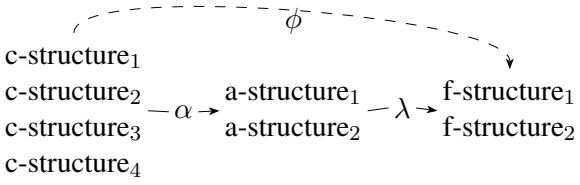

Figure 2: Revised pipeline for language production

This adjusted architecture (see Figure 2) is motivated by recent psycholinguistic research that has shed light on how a bilingual's languages interact and impact each other (Kroll & Gollan, 2014). There is strong evidence that bilinguals' source grammars are co-activated, and inhibitory control is necessary to produce the preferred language. Nonetheless, the dispreferred language is never truly "switched off," and this leads to differences in monolingual and bilingual production and comprehension in a rage of domains, including syntactic and discourse-pragmatics (Sorace, 2011). A wealth of psycholinguistic research shows the parallel nature of the cognitive architecture underlying the language faculty, and in particular work on cognates (Costa, Caramazza, & Sebastian-Galles, 2000; Starreveld, de Groot, Rossmark, & Van Hell, 2014) demonstrates the close relationship between different levels of the grammar (phonology, semantics, lexicon, etc.), and how these levels and languages share resources.

This paper adheres to the follow structure: Section 2 profiles a brief overview of key findings on parallel activation and extended competition in language and cognition. In Section 3, we introduce how residual activation can be modeled in our adjusted pipeline architecture, with Section 4 providing supporting empirical evidence demonstrating the basic desiderata of our proposal. We conclude this paper in Section 5.

2 Parallel Activation and Extended Competition

Research on language processing abounds with evidence in favor of the view that bilinguals co-activate both languages even when only one is in use (Dijkstra, 2005; Kroll, Dussias, Bogulski, & Valdes Kroff, 2012; Kroll & Gollan, 2014; J. Morales, Gómez-Ariza, & Bajo, 2016). Non-selective activation, the fact that both languages are active to some degree even when they are not necessarily needed, has been observed in studies focusing on the lexicon (Lemhöfer & Dijkstra, 2004) and in grammatical properties such as grammatical gender (L. Morales, Paolieri, & Bajo, 2011). In the following, we briefly present collected evidence of parallel activation in both bilingual and monolingual language processing.

2.1 Bilingual competition

Evidence of parallel activation can be seen in the Cognate Facilitation Effect (Costa et al., 2000; Starreveld et al., 2014), in which lexical items that share meaning and phonological form cross-linguistically are are more quickly retrieved. While much research has investigated lexical access in bilinguals, simultaneous activation of both source grammars extends beyond the lexicon, to phonological (Balukas & Koops, 2015), morphosyntactic (Lipski, 2015, 2017), and syntactic (Goldrick et al., 2016a; Kootstra, van Hell, & Dijkstra, 2010) information. Work with multimodal bilinguals, such as speakers of American Sign Language and English, has shown how structural information, such as grammatical markers, is shared (Pyers & Emmorey, 2008), a phenomenon which is difficult to observe in unimodal bilinguals. However, code-switching data may provide evidence of co-activated, competing structural information.

Following seminal work by Muysken (2000), code-switching takes three different forms. **Insertion** is a type of code switching frequently referred to as borrowing. Insertion is defined as the insertion of material from one language into the structure of another (Muysken, 2000, p. 3). In Example (1), the prepositional phrase *in a state of shock* is inserted into the Spanish clausal structure as a unit.

(1) yo anduve *in a state of shock* por dos dias

'I walked in a state of shock for two days'
(Spanish-English insertion, (Pfaff, 1979, p. 296))

Another type of code switching is **alternation**. Alternation is when a bilingual switches between structures from languages (Muysken, 2000, p. 3). Example (2) is an example of Spanish/English alternation, where a switch is made at the coordinating conjunction.

(2) andales pues *and do come again*

'That's alright then, and do come again'
(Spanish-English alternation, (Gumperz & Hernandez-Chavez, 1971, p. 312))

The last type of code switching is **congruent lexicalization**. According to Muysken (2000), in this type of code switching,"material from different lexical inventories" (p. 3) are congruently lexicalized "into a shared grammatical structure" (p. 3). In Example (3), the verb phrase *to give* was congruently lexicalized, leading to a doubling of the verb. In Example (4), we see a doubling of the preposition in a congruently lexicalized prepositional phrase.

(3) they gave me a research grant *koḍutaa*
they gave me a research grant give3.PL.PAST
'They gave me a research grant.'
(English-Tamil congruent lexicalization, (Sankoff, Poplack, & Vanniarajan, 1990, p. 93))

(4) mutta se oli *kidney*-sta *to aorta*-an
but it was kidney-from to aorta-to
'But it was from the kidney to the aorta.'
(Finnish-English adposition doubling, (Poplack, Wheeler, & Westwood, 1989, p. 405))

Such portmanteau constructions are rare and may be viewed as errors, which arise when the bilingual's inhibitory control mechanisms do not prevent the language switch nor the doubled phrase head. This breakdown of inhibitory control sheds light on the linguistic process, and Chan (2015) and Goldrick et al. (2016a) argue that portmanteau constructions are evidence of structural information being co-activated in unimodal bilinguals. Goldrick et al. (2016a) suggest that sometimes the most harmonious and efficient output involves blends where both source grammars can contribute structural attributes.

2.2 Monolingual competition

Competition is not a bilingual phenomenon. Melinger, Branigan, and Pickering (2014) survey types of competition in monolinguals, ranging from lexical to syntactic, a parallel to competition in bilinguals. Nascent research on the role of typological proximity may play in connection with the development of bidialectalism has thus far revealed the conflict between these two systems share certain affinities with bilingualism (Altenberg, 1991; Castro, Rothman, & Westergaard, 2017; Grohmann, Kambanaros, Leivada, & Rowe, 2016; Gürel, 2008). The primary difference between the conflict manifest by bidialectalism vs. bilingualism appears to be the (lack of) overlap between elements from both competing grammars.

A classic example of monolingual syntactic competition is the dative alternation in English (i.e., *give flowers to Anna* vs. *give Anna flowers*). A ditransitive verb can take one of two argument structures, a double object construction or an oblique construction. This is competition in the argument structure, that has an affect on the syntactic structure. Similarly, phrasal verbs exhibit optionality in the positioning of the verb particle. While no occurrences of blended monolingual double object constructions have been attested in the literature, to the best of our knowledge, there are a number of syntactic blends involving phrasal verbs, such as *Would you turn on the light on?* (Melinger et al. 2014, p. 672, cited from Fay, 1980). Such examples exhibit the same surface evidence visible in the portmanteaus of structural competition that is resolved late in the pipeline. This is argued to confirm what Melinger et al. (2014). refer to as *extended competition*. This term refers to the fact that competition occurs throughout the language production

process. When a semantic concept carries with it syntactic optionality, this competition remains unresolved until the stage in the language production process where the structure is built. In Section 3.2 we return to this phenomena to illustrate how these data can be analyzed along the lines of our extended pipeline architecture.

2.3 Parallel Architecture and Inhibitory Control

In light of this research on co-activation, the need for a parallel architecture such as LFG's to model bilingual grammar becomes apparent. Additionally, bilinguals must employ some sort of control filter in order to select the appropriate grammar that will also block out and avoid intrusion from the alternative source grammar (Green, 1998; Green & Abutalebi, 2013; Green & Wei, 2014). Monolinguals make use of these mechanisms as well, albeit to a lesser degree, when processing optionality in argument structure, constituent structure, or even dialectal variation. Leaving aside the debate of whether such control and selection mechanisms are due to domain-specific or domain-general processes (although there is clearly a preference in the literature for the latter, but see Gollan & Goldrick, 2016 for counterarguments), we assume an inhibitory model of control in bilingual grammar and cognition (Abutalebi & Green, 2007; Van Heuven, Schriefers, Dijkstra, & Hagoort, 2008). These inhibitory models "argue that selection occurs at a late locus, once lexical candidates are active in both languages, and it depends on the *competition level*" J. Morales et al. (2016, p. 274). Therefore, we need a model of language production that allows all relevant pieces of linguistic information to cumulatively impact the selection of the surface string.

3 Probability and gradience in the pipeline

Two notable and non-trivial challenges that arise in attempts to model and predict hybrid outputs are the following: First, how can we best define and quantify the notion of (neural) *activity* and how they relate to some sort of *competition level*? Clearly, given the wide gamut of individual differences found in bi- and multilingual grammars, notions such as cognitive control, lexical robustness, and the frequency that these individuals use both/all languages, especially in code-switching contexts, undoubtedly play an important role in determining these factors. A review of such factors can be found in Schwieter and Ferreira (2016). Second, as noted above, this competition seems to extend far beyond lexical items and arguably involves all domains of linguistic structure to various degrees. What notation system of grammatical information or competence can capture both these predominantly discrete and gradient effects simultaneously? In what follows, we employ an LFG model of grammar that that is parallel in design, with correspondences between each respective level mediated by weighted mapping functions.

LFG lends itself to modeling the demonstrated parallel nature of the cognitive linguistic architecture. However, a one-to-one correspondence between infor-

mation structures in the pipeline is difficult to reconcile with the psycholinguistic evidence of co-activation and instances of congruent lexicalization. To adjust the pipeline to accommodate bilingual production, and to allow for the level of gradience that results in various code-switching phenomena, we must adjust the representation of the mapping functions. As we have learned, a bilingual co-activates their source grammars and, despite inhibitory control, cannot completely shut off the language they are not currently using. This means that elements from both source grammars are active simultaneously. By taking a closer look at the examples of portmanteau constructions and the dative alternation in English, we illustrate how applying a probabilistic version of the pipeline architecture can account for these data in a straightforward way with the only significant change being in the functions mapping levels of structure to one another, rather than in the ontology of the pipeline itself.

3.1 The underlying structure of portmanteaus

To begin, we return to the portmanteau construction, as it is the most clear evidence of the parallel activation of linguistic information up until the point of utterance.

(5) they gave me a research grant *koḍutaa*
they gave me a research grant give.3.PL.PAST
'They gave me a research grant.'
(English-Tamil verb doubling, Sankoff et al. (1990, p. 93))

Closer inspection of the portmanteau in (5) reveals that the subcategorization frame for both versions of the verb are identical. Additionally, the phrase structure of the doubled element has a single mirrored difference. In a review of portmanteau constructions, Chan (2015) notes that "portmanteaus emerge in language-pairs in which head-complement order is different for a particular phrase" (p. 105). English verb phrases are left-headed VO constructions, while Tamil verb phrases are right-headed OV constructions. Furthermore, both verbs match in a number of features. Both agree with a third person plural subject, and both verbs are marked for past tense. A detailed analysis of portmanteaus must account for these facts. In LFG terms, we therefore have a single a-structure and a single f-structure, onto which the c-structures map.

However, the presence of the doubled verb raises an interesting question that impacts the pipeline. How do we deal with the two conflicting c-structures that end up merged in the final utterance? Following from that, at what point are lexical items selected that could increase the preference for one c-structure over the other? For instance, selecting English as the preferred language prior to selection of the c-structure should increase the preference for the SVO c-structure and vice versa, if Tamil is selected. The merged c-structures that appear in portmanteau constructions suggest that this choice is made late in the pipeline, and that both English and Tamil lexical items and c-structures remain in competition with one another until

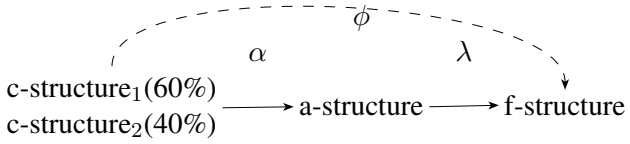

Figure 3: Preliminary revised c- to a- to f-structure pipeline for bilingual production

the very end. Therefore, we assume that there are two competing c-structures, and each c-structure has a probability of being uttered (see Figure 3). For illustrative purposes, we have assign arbitrary probabilities to the individual c-structures in Figure 3. These probabilities are influenced by a number of factors, such as the speaker's linguistic mode, the matrix language, or syntactic priming.

We reach this assumption by establishing a number of foundational concepts. Importantly, the semantic representation of an event is shared in the mind of a bilingual (Kroll, Van Hell, Tokowicz, & Green, 2010). The verb "to give" in our portmanteau example carries with it the same meaning, from the action to the participants, in both English and Tamil. This means that the f-structure is also shared between English and Tamil, as the grammatical functions are identical. In this exercise, we assume that the decision to use a double object construction, rather than an oblique, has been made somewhere further up the pipeline. To yield a shared f-structure, the a-structure must be identical, and therefore shared as well. Under these assumptions, the only location in the pipeline where optionality is possible is the c-structure, and English and Tamil have differing c-structures with a degree of overlap. We therefore have two differing c-structures attempting to map onto the shared a-structure, and thereby the shared f-structure. The mapping of both c-structures to a single f-structure is represented in Figure 4. Here we see that the terminal nodes in the c-structures compete to map to the grammatical functions.

The mapping function must therefore be able to capture the competition in these structures and the multiple output possibilities that exist, albeit with different probable degrees of occurrence. Here, we draw attention to the analysis put forward by Goldrick et al. (2016a), as a possible approach to representing the c-to-f-structure mapping function. In their analysis, they employ Gradient Symbolic Computation (Smolensky et al., 2014) to the process of resolving the word order conflict between English and Tamil. Gradient Symbolic Computation is a version of Harmonic Grammar that is capable of modeling not only the probabilistic distribution of attested outputs, but also the gradience within (bilingual) language outputs. Constraints are weighted for each language specifically, and these weights combine to act on a language-general level. Through language-specific weights, Goldrick et al. (2016a) represent the strength of each c-structure's link to the f-structure, and the combined language-general weight illustrates how the two languages impact one another. The weights can fluctuate, similarly to Stochastic OT, given a number of factors. Importantly, though, in the evaluation of candidates,

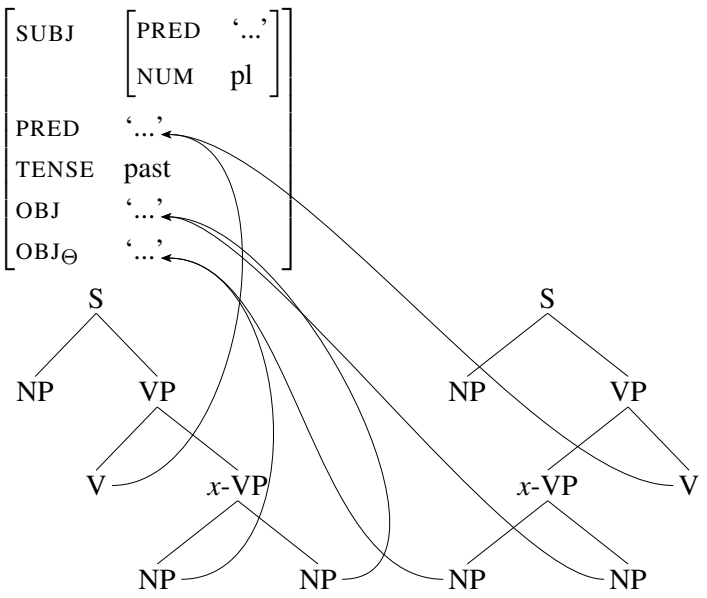

Figure 4: c-to-f structure mapping (ϕ) in bilingual production

no one candidate is deemed optimal. Instead, candidates are assigned a harmony value, and, based on this harmony value, are assigned a probability of being uttered. Any candidate with a non-zero probability may be produced by the speaker.

3.2 The underlying structure of the dative alternation

Returning to monolingual competition, we resume our discussion of the dative alternation. In the case of the dative alternation, there is a direct mapping between each subcategorization frame and a corresponding c-structure. The competition between the double object construction and the oblique construction lies not in the c-structure but instead in the a-structure, and therefore further up the pipeline. Otherwise, the f-structure would not contain the grammatical information necessary to map with the c-structure.

Consider the following monolingual variants of the portmanteau analyzed in Section 3.1. Both (6) and (7) are equally valid for conveying the desired meaning. Yet, one option may be more preferred than the other. This has been studied through a number of lenses, but most relevant here is the work done by Bresnan (2007); Bresnan and Nikitina (2009) and Bresnan et al. (2007). Through corpus studies, Bresnan and colleagues identified patterns in ditransitive verbs that grouped them into preferred tendencies to take either the double object construction, or a prepositional phrase. For instance, for verbs that signify a transfer of possession, such as the verb from our portmanteau, *to give*, English is "heavily

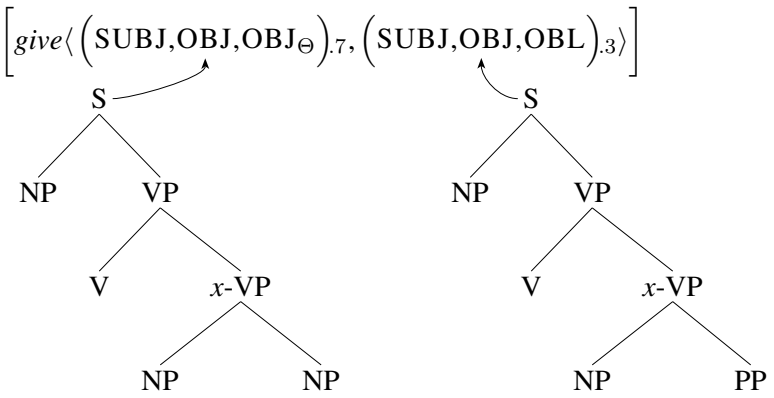

Figure 5: c- to a-structure mapping (α)

biased toward the dative NP construction" (Bresnan & Nikitina, 2009, p. 14).

(6) they gave me a research grant

(7) they gave a research grant to me

This preference for the dative NP construction could be represented as an optimization process in Stochastic OT, as in Bresnan and Nikitina (2009). We build on the probabilistic nature of Gradient Symbolic Computation and propose instead that, in the a-structure, each potential subcategorization frame has a certain probability of occurrence, where the probabilities add up to a total of 1. A representation of this can be seen in Figure 5.

3.3 Reassembling the pipeline

With the insights we have gained by examining competition in both bilingual and monolingual language production, we must now bring all of the pieces together. For the sake of space constraints, we do not elaborate further on the the GSC-treatment of the mapping functions and refer the reader to Goldrick et al. (2016a) and Goldrick, Putnam, and Schwarz (2016b) for a more detailed, illustrative treatment of portmanteaus. Instead, we focus here on the conceptualization of the new pipeline and the principles that underlie it, on the example of the portmanteau construction that also involves the dative alternation, Example (5).

The revised pipeline comprises a number of links; here specifically the mapping of c-to-a-structure, a-to-f-structure, c-to-f-structure, and the cumulative effect of competition in each link. In reassembling the new pipeline, we build upon the evidence of parallel activation, extended competition, and gradience. In what follows, we will model and discuss each link, before discussing the expanded pipeline as a whole and its implications.

Arguments	Probability
DoubleObject$_{en}$	.7
Oblique$_{en}$	.3

Figure 6: English a-structure with probabilities

Arguments	Probability
DoubleObject$_{ta}$	.6
Oblique$_{ta}$	.4

Figure 7: Tamil a-structure with probabilities

We begin with the piece of the pipeline closest to the semantics, and furthest from the utterance, the mapping function λ, which maps the a-structure to the f-structure. As stated above, the verb "to give" participates in the dative alternation in both English and Tamil (Sundaresan, 2006). Therefore, there are two possible argument structures for both English and Tamil. From the Bresnan and Nikitina (2009) study on the dative alternation, we know that the double object construction is preferred in English for this specific verb (Figure 6). A similar study has, to our knowledge, not been performed for Tamil. For illustrative purposes, we therefore will assume that Tamil also has a preference for the double object construction, albeit not as strong as the English preference (Figure 7). A monolingual contends with two options, while a bilingual contends with all four, though they may influence on another. By coming together in the mind of a hypothetical perfectly balanced bilingual, the preferences for each potential argument structure is impacted (Figure 8).

Each potential argument structure corresponds with a specific f-structure. We therefore have two competing f-structures in both the monolingual and bilingual speaker. The preference for the double object construction in our hypothetical speaker translates to a preference for its corresponding f-structure. Importantly, keeping in mind the evidenc of extended competition, the dispreferred a-structure and f-structure are still residually active, while the link between the preferred a-structure and f-structure gains momentum in the pipeline. This, we represent by placing the preferred structure in boldface (Figure 9).

The next piece is the c-structure that correspond with each a-structure. In a monolingual, a single c-structure maps to each a-structure, as depicted in Figure 5. For a bilingual, the picture is more complicated. Two c-structures compete for each

Arguments	Probability
DoubleObject	.65
Oblique	.35

Figure 8: Bilingual a-structure with probabilities

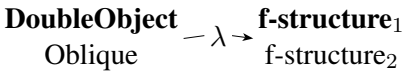

Figure 9: a-to-f-structure mapping

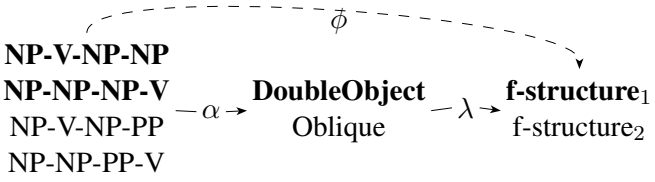

Figure 10: Momentum in the pipeline

a-structure. In the case of the Tamil-English bilingual, an SVOO and an SOOV surface structure compete to map to the double object a-structure, and similarly for the oblique a-structure. Through the interaction between the two languages, and the momentum gained through the linguistic preference(s) for the double object construction, the preference for the respective c-structures in both English and Tamil builds (Figure 10), leading to predictive anticipation.

At the point of mapping c-structure to f-structure, momentum for the preferred construction has built to a point that the likelihood of the competing dispreferred f-, a-, and c-structures has nearly reached zero, but not diminished entirely. In this very late stage of language production, both the English and Tamil word orders are still very much active and must be reconciled. Given the right linguistic circumstances (e.g. a code-switching environment, a certain degree of proficiency in both languages, etc.), the likelihood of neither c-structure can be reduced sufficiently to allow the other to prevail. In such situations, it becomes optimal to blend the two c-structures.

4 Discussion

In light of this body of research on multilingual grammar and cognition, it is no longer possible to look at languages individually, but instead we must integrate both source grammars into a unified model when modeling bilingual production. In this section we highlight how the expanded pipeline we introduce here is an ideal fit for Christiansen & Chater's (2016) *Chunk-and-Pass processing* strategy.

The necessity to integrate various elements of linguistic knowledge simultaneously for the purpose of production and comprehension is an established fact. Christiansen and Chater (2016) correctly point out that "as we hear a sentence unfold, our memory for preceding material is rapidly lost" (p. 95) This leads to their proposal of language processing from a *Now-or-Never* perspective: "if linguistic information is not processed rapidly, that information is lost for good" (p. 95). The most relevant questions that surface in connection with the *Now-or-Never* process-

ing bottleneck concerns ***when*** this information is integrated into a common unit and ***how much*** information can be included in a single unit? From an architectural standpoint Christiansen and Chater (2016) assert that "because memory limitations also apply to recoded representations, the cognitive system further chunks the compressed encodings into *multiple levels of representation* of increasing abstraction" (p. 98). From their perspective, linguistic units which they refer to as ***chunks*** are composed of multiple levels of linguistic representations that are rapidly generated. The process of generating chunks proceeds incrementally, with individual chunks occurring in succession, producing anticipatory processing. This view of language processing results in a model that is heavily dependent on local dependencies, where the learning, production, and comprehension of a grammar takes part at exclusively local intervals.

The rapid integration of multiple cues (Christiansen & Chater, 2016, Chapter 5) in a multi-level architecture of cognition faces working memory constraints due to the *Now-or-Never* nature of language, which means that inevitably some aspects of information will be lost. These representations (i.e., chunks) are, at least to some degree, lossy, in spite of the system's best attempt to be as discrete as possible (Smolensky et al., 2014) and Putnam, Carlson, & Reitter 2017 for a similar position). To ensure that the most important (which is often also the most frequently-produced) information is preserved, information stored within these chunks is compressed, where particular aspects of information are condensed and combined.

Given the rapid demands on successful language production and comprehension, we acknowledge some form of a compression facilitates these processes and reduces algorithmic complexity. The primary function of compression enables the grammar to eliminate informational redundancy whenever possible, thus leading to both more efficient structure building and decoding (see Chater, Clark, Goldsmith, and Perfors (2015, Chapter 2)). Future work on aspects of bilingual grammars from this perspective must revisit and refine the compression algorithm that takes place internally among competing structures of a particular level of representation. Here again is where the notions of overlap and typological proximity may indeed play a decisive role in determining which common elements shared among level-internal candidates may merge (or compress) to become a 'common' or 'shared' structure.

The architectures advanced by Christiansen and Chater (2016) and Putnam, Reitter, and Carlson (2018) are consonant with the neurocognitive research on the bilingual mind reviewed in the previous section, and as we discuss below, can be easily integrated into a parallel architecture such as LFG. As evidenced by the data discussed and analyzed in the previous section, our augmented version of the pipeline architecture can account for the gradient nature of linguistic knowledge without the addition of stipulative theory-internal machinery.

5 Conclusions and directions for future research

In this paper we established the need for the augmented pipeline presented in Figure 2 in order to account for hybrid output representations attested in bilingual grammars (in particular, with reference to code-switching phenomena such as portmanteau constructions). In order to avoid the redundant competition and algorithmic complexity associated with two completely separate grammar systems in conflict with each other, our suggested pipeline introduces an alternative that is consistent with both the parallel correspondence architecture which is commonplace in LFG as well as the literature on the cognitive neuroscience of bilingualism to date.

This proposal raises interesting prospects for the analysis of bilingual data, while at the same time it encounters daunting challenges. To conclude this paper, we allude to three domains of inquiry that emerge as important areas of related research to be pursued in future studies using this model. First, consider the situation when two source grammars possess two contrasting underlying systems for satisfying a shared attribute; how will a compromise be reached? For instance, if two languages mark tense distinctively from one another, how can this best be captured in compressed levels of individual levels of representation (i.e., f-structure)? Second, and related to the first point, how can we best model multiple structures of grammar in an LFG-architecture that are compressed, and as a result, gradient and lossy, as they interact with one another via functional mapping. Initial studies that investigate aspects of syntax from a GSC-perspective have thus far have only involved two levels of grammar (Brehm & Goldrick, 2017; Goldrick et al., 2016a; Putnam & Klosinski, to appear). Future work focusing on linguistic phenomena that involve the role of common information and overlap involving multiple levels of grammar will advance our understanding of the role of contrasting information and compression (e.g., Schwarz, in progress; Schwarz, Brehm, & Putnam, in preparation) and, as a result, may force us to revisit particular architectural assumptions in LFG. This overlapping information, commonly referred to as *mutual information* (Blevins, 2016; Cover & Thomas, 2012), represents an important next step to modeling compression in bi- and multilingual grammars as well as establishing a more detailed description of typological classifications (Brown, Chumakina, & Corbett, 2013). Third, and related to the two previous challenges noted above, in combination with the development of a more detailed compression algorithm, future work must also develop an accessible way and means to establish how elements from other levels of grammar can lead predictive parsing.

In closing, we take solace in the fact that our call for an expanded pipeline is consistent with Christiansen and Chater's 2016 notion of the *Now-or-Never Bottleneck* and other probabilistic models such as Gradient Symbolic Computation (GSC; Goldrick et al. 2016a; 2016b, Smolensky et al. 2014) in our initial attempt to model bilingual grammars in LFG.

References

Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics*, *20*(3), 242–275.

Aguirre Jr., A. (1980). Toward an index of acceptability for code alternation: an experimental analysis. *Aztlan: A Journal of Chicano Studies*, *11*(2), 297–322.

Altenberg, E. P. (1991). Assessing first language vulnerability to attrition. In H. W. Seliger & R. M. Vago (Eds.), *First language attrition* (pp. 189–206). Cambridge University Press.

Asudeh, A., & Toivonen, I. (2015). Lexical-functional grammar. *The Oxford handbook of Linguistic Analysis, 2nd Edition.*, 373–406.

Balukas, C., & Koops, C. (2015). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, *19*(4), 423–443.

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press.

Brehm, L., & Goldrick, M. (2017). Distinguishing discrete and gradient category structure in language: Insights from verb-particle constructions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(10), 1537–1556.

Bresnan, J. (2000). Optimal syntax. *Optimality theory: Phonology, syntax and acquisition.*, 334–385.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston & W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base.* (pp. 75–96). Mouton de Gruyter, Berlin.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In K. I. Bouma Gerlof & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Edita.

Bresnan, J., & Nikitina, T. (2009). The gradience of the dative alternation. *Reality Exploration and Discovery: Pattern Interaction in Language and Life.*, 161–184.

Brown, D., Chumakina, M., & Corbett, G. G. (2013). *Canonical morphology and syntax*. Oxford University Press.

Butt, M., Dalrymple, M., & Frank, A. (1997). An architecture for linking theory in LFG. In *Proceedings of the LFG97 Conference* (pp. 1–16).

Cantone, K. F. (2005). Evidence against a third grammar: Code-switching in Italian–German bilingual children. In *ISB4: Proceedings of the 4th International Symposium on Bilingualism* (pp. 477–496). Cascadilla.

Castro, T., Rothman, J., & Westergaard, M. (2017). On the directionality of cross-linguistic effects in bidialectal bilingualism. *Frontiers in Psychology*, *8*, 1382.

Chan, B. H.-S. (2015). Portmanteau constructions, phrase structure, and lineariza-

tion. *Frontiers in Psychology*, *6*, 1851.

Chater, N., Clark, A., Goldsmith, J. A., & Perfors, A. (2015). *Empiricism and language learnability*. OUP Oxford.

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1283.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Dijkstra, T. (2005). Bilingual visual word recognition and lexical access. In J. F. Kroll & A. M. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches.* (pp. 179–201). Oxford University Press.

Goldrick, M., Putnam, M., & Schwarz, L. (2016a). Coactivation in bilingual grammars: A computational account of code mixing. *Bilingualism: Language and Cognition*, *19*(5), 857–876.

Goldrick, M., Putnam, M., & Schwarz, L. (2016b). The future of code mixing research: Integrating psycholinguistic and formal grammatical theories. *Bilingualism: Language and Cognition*, *19*(5), 903–906.

Gollan, T. H., & Goldrick, M. (2016). Grammatical constraints on language switching: Language control is not just executive control. *Journal of Memory and Language*, *90*, 177–199.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67–81.

Green, D. W., & Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology*, *25*(5), 515–530.

Green, D. W., & Wei, L. (2014). A control process model of code-switching. *Language, Cognition and Neuroscience*, *29*(4), 499–511.

Grohmann, K. K., Kambanaros, M., Leivada, E., & Rowe, C. (2016). A developmental approach to diglossia: Bilectalism on a gradient scale of linguality. *Poznan Studies in Contemporary Linguistics*, *52*(4), 629–662.

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3–15.

Gumperz, J. J., & Hernandez-Chavez, E. (1971). Bilingualism, bidialectalism and classroom interaction. In J. J. Gumperz (Ed.), *Language in social groups.* Stanford, Calif., Stanford University Press.

Gürel, A. (2008). Research on first language attrition of morphosyntax in adult bilinguals. *Second Language Research*, *24*(3), 431–449.

Kootstra, G. J., van Hell, J. G., & Dijkstra, T. (2010). Syntactic alignment and shared word order in code-switched sentence production: Evidence from bilingual monologue and dialogue. *Journal of Memory and Language*, *63*(2), 210 - 231.

Kroll, J. F., Dussias, P. E., Bogulski, C. A., & Valdes Kroff, J. R. (2012). Juggling two languages in one mind: What bilinguals tell us about language

processing and its consequences for cognition. *Psychology of Learning and Motivation-Advances in Research and Theory*, *56*, 229.

Kroll, J. F., & Gollan, T. H. (2014). Speech planning in two languages: What bilinguals tell us about language production. *The Oxford handbook of Language Production.*, 165–181.

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

Lederberg, A. R., & Morales, C. (1985). Code switching by bilinguals: Evidence against a third grammar. *Journal of Psycholinguistic Research*, *14*(2), 113–136.

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

Lipski, J. M. (2015). From 'more' to 'less': Spanish, Palenquero (Afro-Colombian creole) and gender agreement. *Language, Cognition and Neuroscience*, *30*(9), 1144–1155.

Lipski, J. M. (2017). Language switching constraints: More than syntax? Data from Media Lengua. *Bilingualism: Language and Cognition*, *20*(4), 722–746.

MacSwan, J. (2014a). *A minimalist approach to intrasentential code switching*. Routledge.

MacSwan, J. (2014b). Programs and proposals in codeswitching research: Unconstraining theories of bilingual language mixing. In J. MacSwan (Ed.), *Grammatical theory and bilingual codeswitching* (pp. 1–33). MIT Press, Cambridge.

Mahootian, S. (1993). *A null theory of codeswitching* (Unpublished doctoral dissertation). Northwestern University.

Melinger, A., Branigan, H. P., & Pickering, M. J. (2014). Parallel processing in language production. *Language, Cognition and Neuroscience*, *29*(6), 663–683.

Morales, J., Gómez-Ariza, C. J., & Bajo, M. T. (2016). Multi-component perspective of cognitive control in bilingualism. In J. W. Schwieter (Ed.), *Cognitive control and consequences of multilingualism* (Vol. 2, pp. 271–296). John Benjamins Publishing Company.

Morales, L., Paolieri, D., & Bajo, T. (2011). Grammatical gender inhibition in bilinguals. *Frontiers in Psychology*, *2*, 284.

Muysken, P. (2000). *Bilingual speech: A typology of code-mixing* (Vol. 11). Cambridge University Press.

Pfaff, C. W. (1979). Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language*, *55*(2), 291–318.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. *Linguistics*, *18*(7-8), 581–618.

Poplack, S., Wheeler, S., & Westwood, A. (1989). Distinguishing language contact phenomena: Evidence from Finnish-English bilingualism. *World Englishes*, *8*(3), 389–406.

Prince, A., & Smolensky, P. (2008). *Optimality theory: Constraint interaction in generative grammar*. Wiley Online Library.

Putnam, M. T., & Klosinski, R. (to appear). The good, the bad, and the gradient: The role of 'losers' in code-switching. *Linguistic Approaches to Bilingualism*.

Putnam, M. T., Reitter, D., & Carlson, M. (2018). Integrated, not isolated: Defining typological proximity in an integrated multilingual architecture. *Frontiers in Psychology*, *8*, 2212.

Pyers, J. E., & Emmorey, K. (2008). The face of bimodal bilingualism: Grammatical markers in American Sign Language are produced when bilinguals speak to English monolinguals. *Psychological Science*, *19*(6), 531–535.

Sankoff, D., Poplack, S., & Vanniarajan, S. (1990). The case of the nonce loan in Tamil. *Language Variation and Change*, *2*(1), 71–101.

Schwarz, L. (in progress). *Typological proximity and language attrition: Morphological restructuring in Heritage German and Icelandic* (Unpublished doctoral dissertation). The Pennsylvania State University.

Schwarz, L., Brehm, L., & Putnam, M. T. (in preparation). *When grammars overlap: Amelioration effects in hybrid outputs.*

Schwieter, J. W., & Ferreira, A. (2016). Effects of cognitive control, lexical robustness, and frequency of codeswitching on language switching. *Cognitive Control and Consequences of Multilingualism*, *2*, 193–216.

Sells, P. (2001a). *Formal and empirical issues in optimality theoretic syntax*. CSLI Publications.

Sells, P. (2001b). *Structure, alignment and optimality in Swedish*. CSLI Publication.

Smolensky, P., Goldrick, M., & Mathis, D. (2014). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science*, *38*(6), 1102–1138.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism*, *1*(1), 1–33.

Starreveld, P. A., de Groot, A. M., Rossmark, B. M., & Van Hell, J. G. (2014). Parallel language activation during word processing in bilinguals: Evidence from word production in sentence context. *Bilingualism: Language and Cognition*, *17*(2), 258–276.

Sundaresan, S. (2006). The argument structure of verbal alternations in Tamil. In *Proceedings of the 25th West Coast Conference on Formal Linguistics* (pp. 390–398).

Van Heuven, W. J., Schriefers, H., Dijkstra, T., & Hagoort, P. (2008). Language conflict in the bilingual brain. *Cerebral Cortex*, *18*(11), 2706–2716.