A bilingual’s two languages can interact in their mind, but the mechanism of this interaction is still open to debate. In this article we employ a variant of gradient symbolic computation (GSC; Smolensky et al. 2014) to model the code-switched utterances of unbalanced Dutch-English bilinguals. We aimed to evaluate GSC as an appropriate architecture to model bilingual code-switching grammars, and to explore the extent of variability within and across individual bilingual speakers. The results indicate that the structure of individual grammars can vary widely from the structure of the grammar that emerges when the population is studied as a whole. We interpret these results as evidence that individual variation characterizes not only language processing (e.g. Fricke et al. 2019, Kidd et al. 2018), but also the structure of bilingual grammar itself.*

Keywords: bilingualism, code-switching, gradient symbolic computation, computational modeling, individual variation

1. INTRODUCTION. Grosjean (1989) famously suggested that bilinguals cannot be studied as if they had two monolingual grammars in one mind. The following decades of research have borne out his claims. Bilinguals coactivate their languages during both production and comprehension, resulting in phenomena that reveal the intimate interconnections between the structures in their minds and the processes by which they access those structures (for overviews, see Costa 2005, Dijkstra 2005, Kroll & Gollan 2014). Further, the degree of language interaction covaries with individual properties of the speaker-listener, such as proficiency and language dominance (Basnight-Brown & Altarriba 2007, Flege et al. 2003, Kootstra et al. 2012), executive function (Festman et al. 2010, Michael & Gollan 2005), surrounding language context (Elston-Güttler et al. 2005), and a variety of other factors (Tanner et al. 2014); for a review see van Hell & Tanner 2012.

How should we understand language interactions? Models like shared syntax (Hartsuiker & Bernolet 2017, Hartsuiker et al. 2004), the speech learning model (SLM; Flege 1995, 2007), and the competition model (e.g. Hernandez et al. 2005, MacWhinney 2005, 2008) draw on the idea that grammatical structures of the bilinguals’ languages are linked somehow, either through connections between mental representations or through shared memory space, as in SLM. The strength of the links between grammatical structures depends on a variety of influences. The structures’ similarity across the two languages is one such influence; another is the relative activation of the two languages, which in turn depends on factors such as the discourse context and the speaker’s relative proficiency in each language. Interactions are therefore the

* This work was supported by funding from the National Science Foundation (NSF) grant OISE-0968369, and from the Engineering and Physical Sciences Research Council (EPSRC), grants EP/L00058X/1 and EP/M01326X/1. Earlier versions of this work were presented at the 2017 conference on Architectures and Mechanisms of Language Processing (AMLaP) in Lancaster, UK; and the Second International Symposium on Bilingual Language Processing in Children and Adults (ISBPAC-TU) in Braunschweig, Germany.

result of bilingual language processing, affected by an interrelated network, varying widely as each node in the network shifts.

Other strands of work, however, have pursued the idea that interactions arise not solely from processing; rather, there are fundamental grammatical principles and structures that underlie at least some types of language interactions, such as code-switching. These approaches offer both formal, theoretical models (MacSwan 2009, 2014, Muysken 2000, Myers-Scotton & Jake 2009, Poplack 1980, Toribio 2001, 2004) and empirical evidence from experiments (e.g. Fairchild & van Hell 2017, Gollan & Goldrick 2016, 2018, Kootstra et al. 2010) and corpora (Parafita Couto & Gullberg 2019). Although these approaches have not met with universal acceptance (Gardner-Chloros & Edwards 2004), the existence of both processing-based models and grammar-based models highlights two ends of a spectrum of approaches to understanding phenomena like code-switching. In processing-based models, patterns of code-switching derive from individual variation in other cognitive resources, such as working-memory capacity or inhibitory control. In grammatical-based models, patterns of code-switching are governed by the structure of the bilingual grammar itself.¹

Processing-based models are a satisfying way of capturing variability in code-switching across individuals, because we already know that individuals vary widely in processing resources. But the line between grammatically determined phenomena and processing-based phenomena is not straightforward—or, in some views, real. Usage-based approaches to linguistics (Bybee 2006, Bybee & McClelland 2005, McClelland & Bybee 2007, Walsh et al. 2010) propose that grammatical structures are simply generalizations that emerge from repeated exposure to different patterns of language. These patterns must be processed and stored, and since processing and storage depend entirely on cognitive capacities, it is impossible to understand grammatical structures as distinct from cognitive processing.

The goal of this article is to explore the extent of individual variability in these code-switching patterns of language interaction. Specifically, we explore the following hypothesis: individual variation in code-switching patterns may reflect not only differences in processing resources, but also differences in the stored grammatical structures. This hypothesis is compatible both with usage-based models that do not distinguish between grammar and processing and with more traditional approaches that do. For a more traditional approach, our hypothesis means that we are proposing that individuals may have differences that go beyond cognitive capacities, differences that reach to the core grammars they use to generate their speech. For a usage-based approach, this means that individuals may have arrived at distinct emergent generalizations—which we call grammars—even if they are exposed to equivalent input.

We test this hypothesis by employing GRADIENT SYMBOLIC COMPUTATION (GSC; Smolensky et al. 2014) to model the structure of bilingual grammars. GSC is particularly well suited for this investigation because it offers a mechanism to model both language-specific grammars and the interactions between them that compose the bilingual wholes. As we show below, the GSC architecture can accurately capture empirical patterns of Dutch-English bilinguals’ code-switching and can offer insights into the degree of vari-

¹ In this article, we focus on the structure of bilingual grammars by modeling the code-switched utterances of unbalanced bilinguals using gradient symbolic computation. For perspectives that focus on social factors that drive bilinguals’ code-switching, see, for example, Gardner-Chloros 2009, Myers-Scotton 1993, and Torres Cacoullas & Travis 2018.
ation possible across individual grammars. With this tool, we explore the extent to which the grammar of an individual speaker may differ from the emergent grammar that characterizes the broader variety of usage patterns of a group composed of such individuals.

1.1. Principles of gradient symbolic representation. Bilinguals are entirely capable of producing, and indeed usually do produce, output that is almost indistinguishable from the output of a monolingual grammar. This ability is strong evidence that they have separable components to their bilingual grammars. For the sake of simplicity, we consider these components to correspond to separate ‘source’ grammars of the two languages they speak. Yet as we have seen, these grammars interact, allowing bilinguals to employ hybrid representations—that is, those inter- and intrasentential blends and mixes that contain elements from both source grammars. GSC models exactly how the distinct elements from both source grammars are coactivated and integrated to produce these hybrid representations.

GSC is similar in architectural design to harmonic grammar (Legendre et al. 1990) and optimality theory (OT; Prince & Smolensky 2004). Speaker intentions serve as inputs to the model, and each input can be realized as one of a number of potential utterances, or outputs. GSC then evaluates the well-formedness of each potential output on the basis of how many constraints it violates, and how badly it violates them. There are two categories of constraints that a particular output might violate: faithfulness and markedness constraints. Faithfulness constraints are violated when outputs omit or change particular attributes present in the input—for example, by deleting a word or changing a word order. Markedness constraints, by contrast, are violated when outputs contain attributes that are universally disfavored, even if those attributes were present in the input. An example of a markedness constraint violation might be the production of a sentence with OVS or OSV word order, an order that—despite being present in the input—is the most typologically marked (Dryer 2013). The output candidates that best satisfy these two sets of constraints are the most well formed, and are thus the most likely to be produced by speakers.

Importantly, less well-formed candidates may also be produced by a speaker. In this model, ill-formedness does not entirely disqualify a candidate from usage; it simply makes its usage less probable. GSC therefore models the well-formedness of all utterances—including hybrids—at the level of the grammar itself. Utterances that do not follow common patterns of code-switching grammar are not treated as outright errors, but rather as structures with an extremely low probability of occurrence.

The strength of the constraints in shaping the well-formedness of candidates—and hence their usage probability—is determined by a set of weights. These weights adjust either up or down the severity of each candidate’s violation of that constraint. The weighted violations of each constraint are then summed to yield the absolute harmony of a given candidate. From the set of harmony values across each candidate we can generate a probability distribution, which captures the usage probability of each output. (See e.g. Pater 2009 for a detailed discussion of how weighted constraints can successfully be integrated into a research program on linguistic representations.) In what follows, we use the term ‘penalize’ to refer to any part of the model that has the result of lowering a candidate’s harmony, and hence rendering its usage less probable.

GSC models the two sets of grammars that bilinguals draw on as a combination of language-specific constraints and language-general constraints. Language-

---

2 It should be noted that this sets aside the diachronic development of mixed varieties that are beyond the scope of this article; see, for example, DeGraff 1999.
specific constraints can impose violations only on the utterances (or bits of utterances, in the case of code-switching) produced in that particular language. However, language-general constraints can impose further violations on all utterances, regardless of which language they are produced in. These language-general constraints are weighted by combining their language-specific counterparts’ weights in proportion to the speaker’s relative proficiency or dominance in each language. We explain the details of this process in depth in the online supplemental materials; the key point here is that the architecture of bilingual grammars is made of more than the two source grammars (i.e. more than just the two sets of language-specific constraints and weights); they also contain the structured combination of those grammars into the language-general component. Bilingual grammars are thus not two monolingual grammars in one mind (Grosjean 1989).

The combination of the two source grammars into the language-general component highlights a key strength of GSC in its application to code-switching: it has a principled way of capturing the effects of unequal proficiency in the bilingual’s two languages. Depending on the current discourse context, or the bilingual’s proficiency in each of the languages, the source grammars may be active to different degrees. Previous work with this model (e.g. Goldrick et al. 2016, Putnam et al. 2018, van Hell et al. 2016) modeled blends in bilingual corpora, providing a working analysis of how various degrees of language proficiency can be incorporated into a model of bilingual language production. For the sake of simplicity, however, they mostly assumed that bilingual languages were activated equally, and thus contributed equally to the language-general component of the bilingual grammar. In the current article, we build on that implementation by allowing the two languages to contribute unequally, changing dynamically across discourse contexts. We thus extend the reach of the model to more varieties of bilingual language use.

1.2. The current approach. The primary goal of this article is to demonstrate the applicability of GSC to a new type of data—code-switched sentences—in order to illustrate its flexibility and use in exploring new questions of bilingual language use, such as individual variability in grammar. To that end, we first examine how well our implementation of the model can fit new data gathered from empirical experiments of bilingual code-switching, and then use our model to ask a novel question about the degree of variability in bilingual grammars. Our current approach differs from previous works (Goldrick et al. 2016, Putnam et al. 2018, van Hell et al. 2016) in two ways: the empirical nature of the data we use to test our model, and the implementation of the model-fitting algorithm to allow for unequal language proficiency and dynamic changes in language coactivation across discourse contexts. The model we develop here finds support in concurrent research on single-language contexts, dual-language contexts, and dense code-switching contexts (Green & Wei 2014) and in models such as the competition model (MacWhinney 2005) and recent work by Hartsuiker and Bernolet (2017).

Empirical data. In contrast to work to date on hybrid representations in GSC, our primary data set does not originate from a corpus, but rather from a set of psycholinguistic experiments. While corpora provide a wide variety of constructions and a large amount of data, they cannot always illustrate how individuals will respond across a full range of theoretically possible constructions. Such constructions, while permitted by the grammar, may appear too infrequently in a corpus to permit quantitative analysis.

3 The supplemental materials are available at http://muse.jhu.edu/resolve/137.
Experiments, by contrast, can manipulate prompts and utterances in such a way as to provide a balanced set of language data, eliciting responses to a full range of combinatorially possible prompts.

The data source for the analysis presented in this article is Kootstra et al. 2010, in which four experiments were conducted in order to examine how individuals balance conflicting word-order demands when they code-switch between two languages, and whether their word-order choices and switch locations can be primed by an interlocutor. The basic procedure of each experiment was the same: participants were presented with an image—for example, of a girl chasing a donkey—and asked to describe the image using one of three provided lead-in fragments or preambles. In English, the picture description itself would always have SVO word order, but in Dutch, the word order can be SVO, SOV, or VSO, depending on the preamble. Examples 1–3 illustrate the options (from Kootstra et al. 2010).4

(1) English SVO (a), Dutch SVO (b)
   a. A funny picture, because [the girl$_S$ chases$_V$ the donkey$_O$].
   b. Een grappig plaatje, want [het meisje$_S$ achtervolgt$_V$ de ezel$_O$].

(2) English SVO (a), Dutch SOV (b)
   a. A funny picture, in which [the girl$_S$ chases$_V$ the donkey$_O$].
   b. Een grappig plaatje, waarop [het meisje$_S$ de ezel$_O$ achtervolgt$_V$].

(3) English SVO (a), Dutch VSO (b)
   a. On this picture [the girl$_S$ chases$_V$ the donkey$_O$].
   b. Op dit plaatje [achtervolgt$_V$ het meisje$_S$ de ezel$_O$].

The background color of the picture cued participants to complete the description using either at least one word in the same language as the preamble, or at least one word in the language not used in the preamble, thereby creating an environment where the participants were encouraged but not required to code-switch. In experiments 1 and 2, the ‘monologue’ experiments, participants completed this task alone; in experiments 3 and 4, the ‘dialogue’ experiments, they alternated with a partner, who was an experimental confederate.

In the dialogue task, confederates performed the same task as subjects, modeling particular word orders and code-switches to see if they could prime subjects to reproduce the same patterns. Confederates and subjects would always use the same preamble in critical trials. Thus, a turn might begin with the confederate producing an utterance like 4, describing a picture of a girl chasing a donkey with SOV order and a switch from Dutch into English after the second constituent of the subordinate clause.

(4) Een grappig plaatje, waarop [het meisje$_{S,Du}$ de ezel$_{O,Du}$ chases$_{V,En}$].
   a. A funny picture on which [the girl$_S$ the donkey$_O$ chases$_V$].

If the subject repeated the primed utterance, they would in their turn describe a picture of a wizard calling a moose as in 5, using the same SOV order, with the same switch from Dutch into English after the second constituent.

4 In all glossed examples, the subscripts $V$, $S$, and $O$ are used to represent verb, subject, and direct object, respectively. Subscripts $En$ and $Du$ are used to indicate whether a given constituent is produced in English or Dutch, respectively.
Individual variation in the structure of bilingual grammars

(5) Een grappig plaatje, waarop [de tovenaar\textsubscript{S,Du} de eland\textsubscript{O,Du} calls\textsubscript{V,En}].
   a funny picture on which [the wizard\textsubscript{S} the moose\textsubscript{O} calls\textsubscript{V}]

However, subjects were also free to produce other combinations of word order and code-switches, as in 6, which employs SVO word order and switches into English after the first constituent.

(6) Een grappig plaatje, waarop [de tovenaar\textsubscript{S,Du} calls\textsubscript{V,En} the moose\textsubscript{O,En}].
   a funny picture on which [the wizard\textsubscript{S} calls\textsubscript{V} the moose\textsubscript{O}]

Such departures from the primed structure usually occurred when the confederate produced an infelicitous or unattested code-switch. These types of primes were included in the experiment to ensure that the prompts represented the full combinatorial possibilities of word order and switch location. We return to the question of the consequences of this decision for the data in §4.1.

The results of this original experiment showed three primary patterns. First, participants preferred the shared word order (SVO) when switching between Dutch and English. Second, code-switched responses with the Dutch word orders SOV or VSO occurred only when the switch was from English into Dutch, as in 7.

(7) a. A funny picture, on which [de tovenaar\textsubscript{S,Du} de eland\textsubscript{O,Du} roept\textsubscript{V,Du}].
   [the wizard\textsubscript{S} the moose\textsubscript{O} calls\textsubscript{V}]

b. On this picture [roept\textsubscript{V,Du} de tovenaar\textsubscript{S,Du} de eland\textsubscript{O,Du}].
   [calls\textsubscript{V} the wizard\textsubscript{S} the moose\textsubscript{O}]

Finally, participants tended to align word-order choices and language switches with the confederate (except in cases of infelicitous primes, as mentioned above).

This data set therefore allows us to examine how the GSC architecture can handle the full range of possible word orders and switch locations in a carefully controlled set of data that provides a representative set of outputs to a complete set of possible inputs.

Model-fitting algorithm. In order to fit the model to data, it is necessary to determine the appropriate weights for the language-specific constraints.\footnote{Since the language-general constraints are a weighted combination of the language-specific constraint weights, knowing the latter provides the former without requiring additional fitting.} Previous work by Goldrick et al. (2016) and Putnam et al. (2018) assumed that bilinguals are relatively balanced, and that the weights for each language-specific constraint could be determined by analyzing a monolingual set of data and determining the relative constraint weights for each component language before combining them in the bilingual model. This approach is amenable to the MaxEnt grammar tool (Hayes 2009), an implementation of a MaxEnt learning algorithm designed to determine optimum constraint weights in constraint-based grammars (Goldwater & Johnson 2003). The drawback to this approach, however, is the assumption that the weights that fit monolingual data can be straightforwardly transferred into a bilingual model. We know that bilingual grammars are not simply combinations of monolingual grammars (Grosjean 1989), and so we cannot determine a bilingual’s grammatical parameters by simply combining the parameters from two monolingual grammars. The optimal weights for each language-specific constraint must be determined for the full bilingual architecture.

As it turned out, however, the MaxEnt Grammar Tool cannot determine the best-fitting constraint weights if the constraints are divided into multiple components—such as language-specific and language-general constraints—that each can have different weights. Further, it cannot determine constraint weights if constraint violations are frac-
tional, which means that it is not well suited to handling the way we chose to model dy-
namic shifts in language activation across different discourse contexts. Since we needed
to model the priming effect of the interlocutor’s location in the dialogue experiments,
we wanted to be able to vary the strength of violations for omitting constituents in one
language or another as a function of the degree to which that language was activated. In
other words, we wanted the model to be able to recognize that omitting a word in En-
glish is not as severe a violation if the participant was primed to produce an utterance
primarily in Dutch as it is if the participant was primed to produce an utterance mostly
in English. We explain the details of our implementation in the next section, but for now
the key point is that we could not use the MaxEnt Grammar Tool. Rather, we fit the con-
straint weights using a different parameter-optimization algorithm, which had, as we
will see, its own strengths and weaknesses.

The logic of our analysis is as follows. First, in §2, we walk through the details of
how we customized the GSC model to our data. We lay out which parameters we used
(§2.1) and how we defined our constraints (§2.2). By comparing the probability distrib-
ution across all possible utterances predicted by the model to the actual usage frequen-
cies across all possible utterances produced by the participants, we were able to
determine whether a given set of parameter values produced a good fit to the empirical
data. In §2.3 we show how a parameter-optimization algorithm iteratively changed pa-
rameter values and recalculated the probability distribution to arrive at the best possible
fit of the model to the data.6

In §3 we then examine the properties of the fitted model, with special attention to
how well the model actually could fit the data, and what we can infer about the structure
of bilingual grammars by examining the sets of best-fitting constraint weights. We con-
ducted three different analyses. In §3.1 we consider the entire set of data as a whole,
using our model to capture the patterns that emerge from an entire population of speak-
ers. In §3.2 we separate the monologue experimental data from the dialogue experimen-
tal data, in order to see whether the model should in fact be trying to model two separate
types of discourse situations with two separate grammars. Finally, in §3.3 we fit the
model to the data produced by each individual participant, to see whether the grammars
that structure individual utterances have the same properties as the grammars that
emerge from those individuals when they are collected into a whole speech community.
We discuss the implications of our findings in §4 before concluding in §5.

2. Analysis.

2.1. Model parameters. Since the GSC architecture is derived from the principles
of OT, it works by taking an input and generating a set of outputs, which are evaluated
as more or less suitable realizations of the input. For the current analysis we defined the
input as the set of experimental conditions that prompted an utterance. Consider, for ex-
ample, a participant who saw a picture of a girl chasing a donkey, with a code-switch
prompted by the background color and the preamble given in Dutch. In this case, the
input to the model was the set of English and Dutch lexical items for ‘girl’, ‘chase’, and

6 These features are laid out in broad outlines here. In the online supplemental materials (available at
http://muse.jhu.edu/resolve/137) we have provided a much more extensive explanation where the interested
reader can see, step by step, how the model works, from the assignment of raw constraint violations through
to the optimization algorithm. We also include in the OSF archive our model input data, as well as our
parameter-optimization script, so that our analysis can be replicated (https://osf.io/ybz6w/?view_only
=87e100242bbe47019f92fae06c5dbc80).
‘donkey’, combined with the word order required by the structure of the preamble and the prompt to code-switch from Dutch into English. The outputs to be evaluated were then the set of all possible realizations expressing the concept of a girl chasing a donkey, created by crossing all possible word orders with all possible language combinations and switch locations (including failing to switch).\(^7\)

To fit the GSC model to the data, we wanted to create a model that would take the input for each trial and generate a probability distribution over (relevant) possible output utterances. If the model provided a good fit to the data, the probability distribution it generated should approximate the actual usage proportions of these output utterances employed by the participants in the experiment. We built our model to include three sets of parameters in order to capture three domains in which the grammars can vary: language proficiency, discourse context, and constraint weights.

**Proficiency.** Individual bilinguals do not all speak their two languages with the same balance of proficiency. We included this source of variability in the model through the inclusion of two language proficiency parameters, one for Dutch and one for English. These were constrained so that the bilingual’s two proficiency parameters must sum to 1. A monolingual Dutch speaker would have a Dutch proficiency parameter of 1 and an English proficiency parameter of 0. A fully balanced bilingual would have proficiency parameters of 0.5 for both Dutch and English.

Kootstra et al. (2010) collected two measures of language proficiency: (i) self-ratings on a seven-point scale for reading, writing, speaking, and listening in English, and (ii) performance out of 100 on an English lexical-decision task. By dividing the average English self-ratings by 14, we arrived at an estimated English proficiency weight. An average self-rating of 7, indicating top proficiency across all four categories, divided by 14 would yield a weight of 0.5 for English, consistent with a 50/50 split for balanced bilinguals. Across all participants, the average English/Dutch weighting split by this method stood at 0.396 for English and 0.694 for Dutch, with a minimum English weight of 0.26 and a maximum of 0.49.

We also converted English lexical-decision task results by assuming that a perfect score of 100 would indicate peak English proficiency—that is, a 50/50 balance between English and Dutch proficiencies. We therefore divided every lexical-decision score by 200 to arrive at a value for the English proficiency, and took the complement of that value for the Dutch proficiency. In this way, a perfect score of 100 on the lexical-decision task would convert to a 0.5 English/0.5 Dutch split, a score of 90 would convert to a 0.45 English/0.55 Dutch split, and so on. This conversion yielded an average split of 0.393 English and 0.61 Dutch proficiencies, with a minimum English weight of 0.29 and a maximum of 0.49. These figures were very similar to the results from self-ratings. For each analysis that follows, we used the English/Dutch proficiency split derived from lexical-decision tasks, on the logic that these measures were derived from an objective test rather than a self-evaluation. For the group-level analyses we averaged together the proficiencies for all individuals belonging to the group, and for the individual-level analyses we used the proficiencies derived from each individual’s lexical-decision task results.

---

\(^7\) Strictly speaking, the principles of OT require all conceivable utterances to be generated and evaluated against the constraints, where ‘all conceivable utterances’ is construed very broadly. For the sake of simplicity, however, we do not consider utterances that depart too greatly from the input. Utterances such as blue rhinoceros triangle, for example, would violate too many faithfulness constraints to have any reasonable nonzero usage probability.
Discourse context and language activation. The second set of parameters were designed to handle dynamic variation in language activation across discourse contexts. Even if a participant is a fully balanced bilingual, the speaking situation can affect whether they are drawing on only one of their languages ('monolingual speech mode'; Grosjean 1989) or both; and the properties of the individual words that they are drawing on, such as repetition, cognate status, and semantic constraint, can affect how much the two languages interact (Dijkstra et al. 2015, Kootstra et al. 2012, van Hell et al. 2016). In the experimental data we used here, different experimental trials prompted different levels of activation with the two languages. ‘No-switch’ trials did not prompt speakers to use both languages, while ‘switch’ trials did prompt them to include at least one word from both of their languages. Further, in the dialogue experiments, interlocutors could prime an early or a late switch location, thus priming minimal insertion of another language (such as the final word of the utterance only) or relatively equal activation of both languages (such as switching right in the middle of an utterance).

We captured this utterance-specific variation in language activation with two language-activation parameters. The first, $\varepsilon$, represents the degree to which English was activated in a particular utterance, and $\delta$ represents the degree to which Dutch was activated. Like the language proficiency parameters, $\varepsilon$ and $\delta$ were constrained to sum to 1. The logic behind these settings runs as follows: when participants are prompted to switch, the target language becomes activated, and the individual lexical items that will make up the utterance become more activated in the target language. The earlier in an utterance a participant is primed to switch, the more that language will be activated. For example, if a participant is primed to switch from Dutch into English only at the very last word in the utterance, the $\delta$ value will be quite high, because most of the utterance will consist of Dutch lexical items; the $\varepsilon$ value will be lower, because only one of the constituents in the sentence will be in English. If the switch occurs earlier in the sentence, more of the utterance will be in English and correspondingly less will be in Dutch; as a result, the $\delta$ value will drop, and the $\varepsilon$ value will rise. These parameters therefore enable us to represent fluctuations in temporary language activation—which we term ‘language activation’—across different conversational situations, while leaving stable the more long-term differences in language activation that derive from factors such as proficiency. Table 1 summarizes how we set these parameters across the different trial types that appeared in the experiments.

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>$\delta$</th>
<th>Promoted Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>Three Dutch constituents, zero English constituents: Dutch preamble with no switch prompted, but minimal lingering English activation due to bilingual nature of experiment</td>
</tr>
<tr>
<td>0.33</td>
<td>0.67</td>
<td>Three Dutch constituents, zero English constituents: English preamble with a switch to Dutch prompted before first constituent, but lingering English activation due to English preamble</td>
</tr>
<tr>
<td>0.45</td>
<td>0.55</td>
<td>Two Dutch constituents, one English constituent (resulting from English preamble with switch to Dutch prompted after first constituent, or Dutch preamble with switch to English prompted after second constituent)</td>
</tr>
<tr>
<td>0.55</td>
<td>0.45</td>
<td>Two English constituents, one Dutch constituent (resulting from English preamble with switch to Dutch prompted after second constituent, or Dutch preamble with switch to English prompted after first constituent)</td>
</tr>
<tr>
<td>0.67</td>
<td>0.33</td>
<td>Three English constituents, zero Dutch constituents: Dutch preamble with switch to English prompted before first constituent, but lingering Dutch activation due to Dutch preamble</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>Three English constituents, zero Dutch constituents: English preamble with no switch prompted, but minimal lingering Dutch activation due to bilingual nature of the experiment</td>
</tr>
</tbody>
</table>

Table 1. Summary of situations leading to different values of $\varepsilon$ and $\delta$. 
Constraint weights. The third set of parameters, and the set of primary interest in this investigation, comprised the actual weighted constraints that the model uses to generate its probability distribution across possible utterance candidates. For the purposes of this article, we focus on the constraint weights that generate the best fit between the model’s probability distribution across utterances and the actual proportions of those utterances that the speakers produced. In what follows, when we describe similarities and differences in bilingual grammar structures, we refer primarily to similarities and differences in the sets of best-fitting constraint weights.

2.2. The constraints. Consistent with the classical OT architecture, we draw a distinction between markedness constraints and faithfulness constraints. The former penalize utterances that deviate from an unmarked surface form, independently of any properties of the input, while the latter penalize outputs that deviate from the input, regardless of the final surface form. In this analysis, we have both such constraints. The markedness constraints are a set of three linearization constraints, which govern linear word order of the output candidates, regardless of any properties of the input. The faithfulness constraints relate to the experimental conditions. We consider these experimental constraints to be faithfulness constraints because changing the experimental conditions—such as instructions about whether to switch or the preamble of the target sentence—changes how optimal an output utterance may be for a given input prompt.

Each constraint has three components: two language-specific components, and one language-general component. English-specific components penalize violations in the parts of the utterance that are in English, Dutch-specific components penalize violations in the Dutch portions, and language-general components penalize all violations, regardless of which language is used to produce them. In the online supplemental materials we explain further how these components combine to produce a harmony score, and from there a usage probability. Here it is sufficient to walk through the nature of the constraints we used, and to lay out how violations are assigned. For the purposes of this analysis, the key fact to remember is that violations of constraints with higher weights result in a lower harmony, and lower harmony scores make utterances less probable.

Linearization constraints (markedness). The linearization constraints are based on a basic version of X’ theory mediated through the use of violable constraints following Grimshaw (1997, 2002) and Zepter (2003). We constructed them so that they could assign penalties to all combinatorially possible word orders—not only the SVO, SOV, and VSO at issue in this experiment, but also VOS, OVS, and OSV. This was necessary because we wanted the grammar to be able to rule out unattested word orders by assigning those outputs near-zero usage probabilities, as well as to accurately predict the usage rates of the possible word orders. Although two linearization constraints could, in principle, uniquely discriminate the six word orders, we chose to use three constraints in order to provide a more fine-grained violation profile. In the event that one of the constraints proves redundant, the model can prevent it from playing an active role by assigning it a weight near zero.8 As we shall see, this did not turn out to be the case with any of the linearization constraints.

SpecLeft (SL): The specifier should be in the leftmost position of the phrase. At the clause level under analysis, the specifier corresponds to the subject NP. Each interven-

8 Indeed, in OT and its descendants, this is the standard assumption for the constraints that are accepted as present in universal grammar, but are inactive in a given language.
ing constituent between the specifier and the left edge of the phrase incurs one violation. SOV and SVO word orders incur no violations, VSO and OSV orders incur one violation, and VOS and OVS incur two violations.

**HeadRight (HR):** The head (inflected verb) should be in the rightmost position of the phrase. Each intervening constituent between the head and the right edge of the phrase incurs one violation. SOV and OSV incur no violations, SVO and OVS incur one violation, and VSO and VOS incur two violations.

**HeadLeft (HL):** The head (inflected verb) should be in the leftmost position of the phrase. Each intervening constituent between the head and the left edge of the phrase incurs one violation. VSO and VOS incur no violations, SVO and OVS incur one violation, and SOV and OSV incur two violations.

**Experimental constraints (faithfulness).** The experimental constraints govern how well a particular utterance fits the requirements of the experimental conditions. They are as follows.

**MatchPrompt:** If the output word order matches the order required by the preamble, assign no violations. If it does not match, then the number of violations is the number of position changes, or ‘swaps’, required to turn the output word order into the prompted word order. For example, if the prompted word order were SVO, then SOV would incur one violation, because the V must move only one position to the right, swapping with O, to arrive at the prompted SVO order. (Put another way, the O must move only one position to the left, swapping with V.) OSV would incur two violations, because the O would need to move two positions to the right, swapping first with S (producing an intermediate SOV) and then again with V, to produce SVO. OVS would incur three violations, because the S would need to move two positions to the left, swapping with both V and O (producing intermediate SOV), and then V would need to move one position to the left, swapping with O to arrive at SVO.

In the event that a response contains both English and Dutch components, then the full number of violations is multiplied by the proportions of the utterance in English and in Dutch to arrive at the English-specific and Dutch-specific violations, respectively. See the supplemental materials for more detail.

**MatchSwitch:** Assign one violation if the output contains an unprompted code-switch, or does not contain a prompted code-switch.

**Max:** Assign one violation for every constituent present in the input that is not included in the output. Note that this constraint is impossible to avoid violating entirely, because when a bilingual produces a three-constituent utterance, six constituents are activated, three in each language. To handle this fact, assuming no doubling constructions, we implement this constraint in conjunction with the language-activation parameters, ε and δ. To calculate the violations of Max, we multiply the number of omitted English elements by ε, and the number of omitted Dutch elements by δ. For example, suppose a participant was primed to switch from Dutch into English after the first word of the utterance. The primed utterance would contain in total one Dutch con-

---

9 We modeled the preamble to prompt the same word order in both English and Dutch, to allow for the possibility that an English preamble might prompt word order appropriate to its Dutch translation equivalent in a code-switched utterance.

10 The model presented here does not explicitly rule out doubling constructions, but we assume that the inclusion of a judiciously highly weighted constraint like Quantization (Goldrick et al. 2016) will always ensure that the constituents produced are represented by only one spell-out.
stituent and two English constituents, and hence omit two Dutch constituents and one English constituent. For this type of utterance, the $\epsilon$ value is 0.55, while the $\delta$ value is lower, at 0.45 (see Table 1).

Suppose, now, that the participant who was primed to switch after the first word of the utterance in fact switched after the second word, producing two Dutch constituents (omitting one) and one English constituent (omitting two). This utterance would garner an unweighted Dutch Max violation of 0.45 (one omitted Dutch constituent $\times 0.45 = 0.45$), and an unweighted English violation of 1.1 (two omitted English constituents $\times 0.55 = 1.1$), for a total of 1.55 language-specific violations. If, however, they switched at the position their interlocutor primed—after the first constituent—they would garner an unweighted Dutch Max violation of 0.9 (two omitted Dutch constituents $\times 0.45 = 0.9$) and an unweighted English Max violation of 0.55 (one omitted English constituent $\times 0.55 = 0.55$), for a total of 1.45 language-specific violations. In this way, the use of the $\epsilon$ and $\delta$ parameters allows us to penalize late switches through Max, while the existence of MatchSwitch allows us to penalize immediate switches into the language with the highest $\epsilon$ or $\delta$ parameter.

2.3. Determining weights.

Building input data sets. To evaluate the model’s ability to capture the results of Kootstra et al. (2010), we built a data set of all possible prompts. These could vary according to three fully crossed factors: the language of the sentence preamble (English or Dutch), the order of the sentence completion signaled by the preamble (SOV, SVO, VSO), and where the participant was primed to switch (NS: no switch, 0: switch immediately after the preamble, 1: switch after first constituent, 2: switch after second constituent). This treatment created twenty-four (= $2 \times 3 \times 4$) different prompts.

Finally, an additional factor to consider was utterance type, which coded whether the participant was participating in a monologue or dialogue experiment. In the monologue experiments, participants were simply instructed to switch (coded as a prompt switch location of 0) or not switch (coded as a prompt switch location of NS); in the dialogue experiments participants were primed to switch by their partner’s choice of switch location in the previous trial.11 These prompts were kept separate, which means that each of the twelve prompts with a switch location of ‘0’ or ‘NS’ was represented twice, once from the monologue and once from the dialogue experiments. As a result, this created not twenty-four possible prompts, but thirty-six, of which twelve were effectively duplicates.

For each prompt, we then assembled the set of all theoretically possible responses, which was the set of crossed combinations of response order (SOV, SVO, VSO, and the combinatorially possible but theoretically impossible (or irrelevant in the experiment) VOS, OSV, and OVS) and switch locations (NS, 0, 1, 2, and the multiswitch 0-1, 0-2, 1-2, 0-1-2), leading to forty-eight ($6 \times 8$) possible responses to each prompt. The final data set therefore contained 1,728 lines (36 prompts $\times$ 48 responses). For each combination of prompt and response, we recorded the number of times a participant actually gave that response to that prompt, using the data from Kootstra et al. (2010). Finally, we assigned language-specific violations to each prompt-response pair, as outlined in the supplemental materials.

Parameter-optimization algorithm. In order to calculate the language-general violations, and from there combine the language-specific and language-general viola-

11 They were also primed with the confederate’s word order, but since the confederate’s utterance always aligned with the word order signaled by the preamble, we did not separate these two influences on word order when coding the input.
tions to determine the harmony—and hence the probability—of each response to each prompt as predicted by GSC, it was necessary to determine the relative weights of each of our language-specific constraints. The previous work by Goldrick et al. (2016) determined these weights by means of the MaxEnt Grammar Tool (Hayes 2009) and mostly assumed an ideal, balanced bilingual grammar. However, this algorithm in its current form is not well suited to the sort of bilingual dual-constraint model we are employing here, where not only are language proficiencies imbalanced, but also $\varepsilon$ and $\delta$ values are in flux. The MaxEnt Grammar Tool takes a set of violations and determines the best-fitting constraint weights that allow those weighted violations to generate a harmony profile which, when transformed into a probability distribution, best fits the actual usage proportions across all output candidates in the data. This works well if only one set of constraint weights is needed—and, indeed, Goldrick et al. (2016) used the MaxEnt Grammar Tool to determine the appropriate constraint weights independently for the two languages they examine.

With bilingual data, however, the algorithm cannot be applied, because it cannot determine the appropriate weight for two sets of language-specific constraints. If a constraint such as HeadLeft must have a weight of 9, how should it be divided between the language-specific components? Dividing it equally between them would result in identical weightings for two separate languages—effectively implying that the two languages have identical grammars, which is objectively not true. But if one language must weight HeadLeft higher than the other, then how much higher should it be weighted—and what happens if the bilingual is not equally proficient in both languages? Further, the MaxEnt Grammar Tool can only handle whole-number violations, while our use of $\varepsilon$ and $\delta$ parameters generated decimal violations.

For these reasons, we needed a different algorithm to determine the best constraint weights. One possibility was to create a grid of possible constraint values and compare the fit of the model at each point in the grid. However, since that would have required testing every point in a twelve-dimensional parameter space, it quickly became much too computationally intensive to examine a reasonably large set of possible parameter values. We therefore chose a different parameter-optimization method, in which we started with a set of constraint weights randomly selected from the uniform distribution ranging from 0 to 10, and then iteratively cycled through the set of constraints, adjusting each one in turn, retaining the same order of update throughout each iteration. Each adjustment to a constraint weight was sampled from a uniform distribution ranging from $-2$ to 2, after which the algorithm recalculated the fit between the predicted probabilities and the observed proportions of each response. After each recalculation, the algorithm retained adjusted values that produced a better fit, and undid adjustments that produced a worse fit. A full optimization run tested 10,000 adjustments to each of the twelve language-specific constraints, for a total of 120,000 steps along the path to a final fit.

We calculated the fit between observed proportions and predicted probabilities as kullback-leibler divergence (KLD; Kullback & Leibler 1951). The supplemental materials provide more details on the optimization algorithm. For the current purposes, it is sufficient to know that a lower KLD indicates a better fit.

We ran this optimization algorithm on our data in three different ways. The first analysis considered the entire data set, aggregated across all ninety-three participants and all four experiments. This was designed to examine the parameters that best capture the behavior of a population as a whole—a population-level analysis. The second analysis split up the data into the monologue experiments—those in which participants did not alter-
nate utterances with a confederate—and the dialogue experiments, in which participants alternated with an interlocutor. Weights were optimized for each experiment type independently. This experiment-level analysis allowed us to explore the possibility that the structure of a grammar can in fact vary according to external, situational factors. The final analysis considered each of the ninety-three participants separately—a speaker-level analysis—allowing us to examine the relationship between the parameters favored by an individual and the parameters that emerge in the population as a whole.

For each analysis, we ran our optimization algorithm multiple times. We did this for several reasons. First, we did not know the shape of the optimal parameter space. Conceivably there could be sets of weights that generate better-fitted predictions than surrounding values, and hence are locally optimal, but that do not provide the globally optimal fit of the model to the data. Since the optimization algorithm has a substantial random component to the iterative adjustment of constraint weights, different optimization runs can result in different paths to the optimal fit. By running multiple optimizations, we gave the model the best chance to arrive at the truly globally optimal fit, whereby the final weights were able to generate a predicted probability distribution that most closely resembled the empirical data.

The second reason to run multiple optimizations has a deeper motivation, springing from our core question about the extent of variability in grammar. If multiple different sets of weights turn out to predict the probability of various responses equally well, then there is no reason to suppose that identical output must be generated by identical underlying grammars. Two individuals exposed to identical input may infer two very distinct underlying grammatical architectures. It is for this reason that we used this iterative algorithm with its random component rather than employing a more stable algorithm, such as a variant on MaxEnt customized to fit our model architecture. By allowing the optimization to explore varying paths to its final output, we were able to explore how much variability was possible across equally well-fitting results.

Each optimization run took between one and two hours. All analyses were conducted on a high-performance computing cluster with thirty-two cores running in parallel. For the first two analyses (the first on the full group and the second on the monologue and dialogue experiments), we conducted 1,000 optimizations. For the third, speaker-level, analysis, we ran fifty optimizations per speaker.12

3. Results.

3.1. Population-level analysis. First, to confirm that the model is capable of generating reasonably well-fitted approximations, we compared the final KLDs of our 1,000 optimizations to the KLDs generated from 1,000 randomly chosen sets of weights drawn from the uniform distribution between 0 and 10. Figure 1 shows the distribution of final KLDs from the optimized and random sets of weights. For an intuitive reference point, we also calculated the KLD for a uniform distribution, in which each response was equally likely, and marked it in Fig. 1 with a dotted line.

As Fig. 1 shows, the optimized weights generate probability distributions with substantially lower KLDs than the random weights. In other words, the optimized weights can better predict the empirical proportions of responses than randomly generated weights can. Of the optimized weights, 595 sets of weights out of the total 1,000 produced predicted probabilities with a KLD below 1. This means that they could represent

12 Completing fifty optimizations for each of the ninety-three speakers took nearly three weeks. Running 1,000 optimizations per speaker was not computationally feasible.
the empirical proportions of responses with less than one bit of information loss. Further, these 595 optimizations all converged on a very similar final fit, with a mean KLD of 0.89 and a standard deviation of just 0.06. In the analysis of the actual weight values, we focus on these 595 sets of optimized weights, which we consider to have ‘successfully’ fitted the model predictions to the empirical data.

**Predictive accuracy of successful optimizations.** Table 2 shows the tableau for one representative ‘successful’ optimization, with weights rounded to the nearest hundredth. Candidates are ordered according to the predicted usage probability. For the sake of space and simplicity, we restrict the tableau to show only the top fifteen of forty-eight possible responses to one possible input—an English preamble whose Dutch counterpart would condition SOV word order, with instructions to the participant to include at least one English word in the response. It comes from a dialogue experiment, which means that the confederate had previously responded to a prompt with the same structure and switched after the first constituent.

The weighted violations shown in this tableau were calculated as described in the supplemental materials. Thus, for example, candidates 8 and 9 each have one violation of HeadRight and HeadLeft, since their V was produced medially, one position away from both the left and right edges of the clause. However, in candidate 8, the V was produced in English, because there was no code-switch separating it from the English preamble, while in candidate 9, the V was produced in Dutch, following a code-switch immediately after the preamble. For both candidates, the English-specific weights (9.08 for HeadRight and 12.08 for HeadLeft) are higher than the Dutch-specific weights (6.04 and 7.38, respectively), while the Dutch proficiency of 0.61 is greater than the English proficiency of 0.39. As a result, the weighted violations are very close to each other. For HeadRight, the difference in proficiencies undoes the difference in language-specific weights, so that the final weighted violation of 10.91 is greater for candidate 9, with its Dutch V, than the weighted violation of 10.77 for candidate 8, with its English V. For HeadLeft, however, the difference in language-specific weights is greater, and so the difference in proficiencies can minimize, but not undo, the fact that violating that constraint with an English constituent, as candidate 8 does, will incur a larger weighted violation (13.93) than violating it with a Dutch constituent, as candidate 9 does, incurring a weighted violation of 13.72.
probability vs. actual proportions. The table is sorted by decreasing probability (bolded), as predicted by the
supplemental materials, and especially §A.5, for more details on the mathematics underlying this tendency.

divided among more possible candidates, thus leading to an underprediction of the most probable ones. See the
sort on a limited data set. With only seventy-five responses and forty-eight possible outputs, it is only to be ex-
spected that the SVO order that would be correct in English or the Dutch-appropriate SOV

termination of uses than any other response, and then a variety of other responses with much
Table 2: they usually favored one particular response with a substantially higher propor-
tion of uses, although its proportion of uses, at 0.493, was larger than its pre-
ferred probability of 0.31. The runner-up responses that were selected less often are
most often selected, although its proportion of uses, at 0.493, was larger than its pre-
vided by 75. As this tableau shows, the most probable predicted response—candidate 1,
actual proportions represent the number of times a given candidate was produced, di-
tern (English preamble prompting SOV order, with instructions to switch), and so all
prompt, Max 

<table>
<thead>
<tr>
<th>Head</th>
<th>Head</th>
<th>Match</th>
<th>Match</th>
<th>Pred.</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpecLeft</td>
<td>Head</td>
<td>Head</td>
<td>Left</td>
<td>Prompt</td>
<td>Max</td>
</tr>
<tr>
<td>English (0.39)</td>
<td>5.37</td>
<td>9.08</td>
<td>12.08</td>
<td>-1.99</td>
<td>4.23</td>
</tr>
<tr>
<td>Dutch (0.61)</td>
<td>3.19</td>
<td>6.04</td>
<td>7.38</td>
<td>5.04</td>
<td>3.88</td>
</tr>
</tbody>
</table>

table 2. Selected outputs to a possible input, along with weighted violations, harmony, and predicted
probability vs. actual proportions. The table is sorted by decreasing probability (bolded), as predicted by the
model. The input ‘en-sov-1’ indicates a prompt with an English preamble that takes SOV word order, and a
switch into Dutch primed to take place after the first constituent. (Since these input-output pairs relate to a
dialogue experiment, the switch location is in principle distinguished from another location.) The responses
indicate the produced word order, with switch locations indicated by a vertical line.

There were seventy-five observations in the experiment with this particular input pattern (English preamble prompting SOV order, with instructions to switch), and so all
actual proportions represent the number of times a given candidate was produced, di-
vided by 75. As this tableau shows, the most probable predicted response—candidate 1,
with an immediate code-switch into Dutch and SOV word order—is in fact the response
most often selected, although its proportion of uses, at 0.493, was larger than its pre-
dicted probability of 0.31. The runner-up responses that were selected less often are
mostly correctly predicted to be chosen between 0 and 10% of the time, and all have either
the SVO order that would be correct in English or the Dutch-appropriate SOV
order, with a variety of switch patterns possible.

In general, the response patterns across all prompts were similar to those shown in
Table 2: they usually favored one particular response with a substantially higher propor-
tion of uses than any other response, and then a variety of other responses with much
lower proportions of uses. The model regularly identified the most common response
and distinguished it from the cluster of less common responses—although, consistent
with the trends shown in Table 2, it tended to underpredict the most frequently used re-
sponses and overpredict the less common ones.14

13 The harmony weights for each candidate \( H(c) \) are converted to predicted probabilities according to the
formula \( e^{H(c)} \). See the supplemental materials, §A.5, for more details.

14 These deviations from the actual usage patterns are an expected consequence of employing a model of this
sort on a limited data set. With only seventy-five responses and forty-eight possible outputs, it is only to be ex-
pected that most of the outputs with low predicted probability would not be used at all, hence leading to an ap-
pearance of overprediction. Furthermore, since we forced the model to output a nonzero probability for all
responses—even those that are theoretically unattested (e.g. OVS word order)—the probability mass must be
divided among more possible candidates, thus leading to an underprediction of the most probable ones. See the
supplemental materials, and especially §A.5, for more details on the mathematics underlying this tendency.
Commonalities across optimizations. Now that we have seen that the model can produce a reasonably strong fit between predicted probability and actual usage proportions across utterances, we can examine the underlying parameter values that allow the model to fit the data so well. Figure 2 shows the distribution of language-specific constraint weights across the 595 best-fitting optimizations.

These weightings can allow us to draw some inferences about the internal grammars that bilinguals may be using to produce their responses in code-switched utterances. First, with respect to the linearization constraints SpecLeft, HeadLeft, and HeadRight, the English-specific weights are generally higher than the Dutch-specific weights. This is entirely consistent with what we know about constraints on word order in the two languages: English allows only SVO word order, while Dutch allows a more flexible choice between SVO, VSO, and SOV orders. As a result, violations of SVO order will incur higher penalties in English than in Dutch. Nevertheless, the relative weights of the three linearization constraints within each of the two languages are the same: HeadLeft and SpecLeft are both higher than HeadRight. This is consistent with the shared preference for SVO orders across the two languages, as well as a shared dispreference for the three improbable or illegal orders of OSV, OVS, and VOS.15

The fact that Dutch word order is constrained by the word order prompted by sentence context is captured by the vastly different weights allotted to MatchPrompt. Dutch weights for MatchPrompt are enormously higher than English weights, which in fact consistently stabilize at negative values. This reflects the fact that Dutch word order, while also SVO, is less strictly so than English and thus is governed by the prompt, while English word order is so much more strictly SVO that an utterance will in fact be rewarded for violating the prompted word order (i.e. with negatively weighted violations) if in so doing it adheres to SVO order.

Variations across optimizations. Although the pattern of results shown in Fig. 2 makes intuitive sense, the distribution of successful constraint weights shows a degree

---

15 Nevertheless, it is not clear why a preference for SVO order, and a subject-initial order making up two of the three possible Dutch word orders, would not lead to SpecLeft being weighted highest. Perhaps HeadLeft is doing more work in ruling out OSV than SpecLeft is doing in rewarding SVO, or perhaps HeadLeft is fighting against HeadRight.
of overlap between the English and Dutch language-specific constraints. In other words, the general trend observed by aggregating all of the weights may not necessarily obtain in each particular optimization. Is it the case that, for example, the English and Dutch Max constraints are usually about the same, or could it be that sometimes the English weight is higher than the Dutch, and sometimes vice versa? And is it always the case that English linearization constraints are weighted higher than Dutch ones? Answering this question allows us to consider whether there is really only one configuration of constraint weights that can accurately capture the patterns in the observed data, or whether multiple grammars could model it equally well.

Figure 3 shows the weights corresponding to the 100 optimization runs that have the lowest KLD—that is, the best-fitting results. In this case, each run resulted in a final KLD below 0.80. Each dot represents the final weight of an optimization. The weights for the English-specific and Dutch-specific versions of each constraint that resulted from the same optimization are connected by lines. To the extent that the lines connecting the dots for each constraint are parallel, the different optimizations produced roughly the same pattern of constraint weights. To the extent that they cross—and especially to the extent that the English weight is higher than the Dutch weight in some cases, and lower in others—distinct parameter settings are capable of modeling equivalent outputs.

Several patterns emerge from Fig. 3. First, the pattern of English-specific linearization constraints being higher than their Dutch-specific counterparts generally holds. Of the full 595 optimizations with a KLD below 1, none had an English HeadLeft lower than the Dutch HeadLeft; only one (0.17%) had an English SpecLeft lower than Dutch; and only twenty-six (4.4%) had a lower English HeadRight than Dutch. The differences for HeadRight are not negligible, but they are also not enough to obscure a decided pattern of higher weights for English-specific linearization constraints than for Dutch.

However, not all optimizations agree with each other. This is most strikingly apparent with MatchSwitch, which shows a wide variability in which language has the higher

---

16 Nevertheless, the patterns illustrated in Fig. 3 are present throughout the full set of 595 optimizations. We omit the remaining 495 optimizations from the figure merely to reduce visual clutter.
weight. Yet this variability is simply part of a slightly more complex pattern in which the Dutch and English weights appear to trade off. Indeed, a simple linear regression modeling the Dutch weight against the English weight returns an intercept of 0.60 ($SE = 0.008$, $t = 75.24$, $p < 0.001$) and slope coefficient of $-0.64$ ($SE = 0.003$, $t = -240.9$, $p < 0.001$), with this one variable alone producing an $R^2$ of 0.99. In other words, we can account for almost all of the variability in the Dutch MatchSwitch weight simply by saying that its value is about negative two thirds the weight of its English counterpart, plus 0.6.

To determine whether this degree of explanatory power holds among the relationships between other constraint weights, we built a set of six linear regression models to determine what proportion of variance in one language-specific constraint weight—in this case, Dutch—could be explained by its counterpart in the other language-specific constraint (English). The results are summarized in Table 3. Three constraints—HeadLeft, HeadRight, and MatchSwitch—show very strong relationships, in which the weight for the English constraint explains over 90% of the weight for the Dutch constraint. For SpecLeft, the relationship is also strong, with nearly 80% of the variation explained. With Max and MatchPrompt, however, the relationship was weaker. The English weight for Max explained only about 6% of the variation in the Dutch weight, and for MatchPrompt only about 4%.

**Table 3. Linear model estimates predicting the Dutch weights for each constraint as a function of the corresponding English weights. Standard errors are in parentheses below each coefficient estimate.***

<table>
<thead>
<tr>
<th>PREDICTOR:</th>
<th>SpecLeft</th>
<th>HeadLeft</th>
<th>HeadRight</th>
<th>MatchSwitch</th>
<th>MatchPrompt</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG. WEIGHT</td>
<td>0.717 ***</td>
<td>0.650 ***</td>
<td>0.645 ***</td>
<td>$-0.641$ ***</td>
<td>$-0.188$ ***</td>
<td>0.129 ***</td>
</tr>
<tr>
<td>(0.015)</td>
<td>(0.003)</td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.039)</td>
<td>(0.020)</td>
<td></td>
</tr>
<tr>
<td>(intercept)</td>
<td>$-0.644$ ***</td>
<td>$-0.541$ ***</td>
<td>0.146 ***</td>
<td>0.598 ***</td>
<td>4.804 ***</td>
<td>2.806 ***</td>
</tr>
<tr>
<td>(0.097)</td>
<td>(0.21)</td>
<td>(0.010)</td>
<td>(0.008)</td>
<td>(0.065)</td>
<td>(0.071)</td>
<td></td>
</tr>
<tr>
<td>OBSERVATIONS</td>
<td>595</td>
<td>595</td>
<td>595</td>
<td>595</td>
<td>595</td>
<td>595</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.795</td>
<td>0.990</td>
<td>0.993</td>
<td>0.990</td>
<td>0.038</td>
<td>0.062</td>
</tr>
<tr>
<td>$F(1,593)$</td>
<td>2.297 ***</td>
<td>60,900 ***</td>
<td>84,080 ***</td>
<td>58,030 ***</td>
<td>23.45 ***</td>
<td>39.35 ***</td>
</tr>
</tbody>
</table>

Even with the lack of explanatory power in the linear regression model for MatchPrompt, it is clear that there is at least one systematic property that governs the weights: the Dutch value must be higher than the English value. The amount by which it must be higher is quite variable—hence the low $R^2$ of the linear model. But as long as it is higher by a certain amount, the optimization will produce a good fit to the data. Figure 4 illustrates this pattern. The solid line shows the poor-fitting regression line. In other words, it represents an attempt at modeling the relation between the Dutch and English weights as an equation. The dashed line shows what can be gained by modeling the relationship instead as the inequality $y = 2 - x$. As long as the Dutch weight is greater than the value dictated by the inequality—as long as it remains (roughly) above and to the right of the dashed line—its exact value is unimportant.

**Potential limitations on the number of distinct grammars.** The patterns summarized in the previous section show that systematic relationships between English and Dutch constraints can be observed for most domains of optimized grammar. Not considered, however, were potential relationships in the patterns across constraints. Consider, for example, MatchSwitch. Although the trading relationship between the
English and Dutch weights can be well described by the linear regression in Table 3, the fact remains that the specific values of these weights can vary dramatically. In a little under half of the optimizations (282, 47.4%), the English weight was higher for MatchSwitch than its Dutch counterpart. With Max, a similar pattern obtains in reverse: just under half (254, 42.7%) of runs have a lower weight for English Max than for Dutch.

Are these similar rates of exceptions a coincidence, or do they reveal some stronger relationship across different constraints? Is it the case, for example, that those 282 optimizations where English MatchSwitch was higher than Dutch MatchSwitch contain those same 254 optimizations where English Max was lower than Dutch Max? If so, then it may well be that there are indeed multiple patterns of constraint weights that can yield a successful optimization, but that the number of such patterns is limited. In other words, perhaps the same speaker behavior can be modeled by a handful—but only a handful—of distinct grammars.

Figure 5 illustrates that this possible limitation on the number of distinct grammars does not seem to apply for Max and MatchSwitch. The figure shows the pattern of final weights for the set of 100 best-fitting optimizations, divided up according to the pattern in Max. The top row shows the optimizations that arrived at a Max weight that was higher in Dutch than in English, and the bottom row shows those optimizations for which the Dutch Max ended up lower than the English version. Even in this sample of 100 optimizations, it is clear that there is no relationship between the relative weights for MatchSwitch and Max: in both the top and the bottom panels, the full range of possible values for Dutch and English MatchSwitch is represented.

In summary, the analysis of the full set of data aggregated across all subjects and all experiments reveals that, with the exception of Max, it is possible to describe the parameters of the grammar with a limited number of equations and inequalities, which explain a very large proportion of the variation across optimized constraint weights.
Although the absolute values of parameter weights may vary widely, they obey the same constraints on their relationship to each other and can equally well fit the empirical data on which the model was optimized.

Nevertheless, as Table 2 shows, the fit of the grammar to the data is not perfect. One source of this imperfect fit could be that by aggregating across multiple sources of data—different experiments, different individuals—we are in fact asking the grammar to model something that is a compromise of multiple generative systems, beyond the two languages we have built into this implementation. In the remaining sections, we explore this possibility further, starting with the difference between monologue and dialogue experiments in §3.2, and then moving on to individual variation in §3.3.

3.2. Experiment-level analysis: monologue vs. dialogue. To further probe the possibility that our fully aggregated analysis was reflecting a compromise between different possible grammars, we decided to ask whether there was evidence that different grammars were active in different speaking situations. To that end, we split up the set of data according to the experimental design. We then ran our parameter-optimization algorithm to generate separate sets of constraint weights for the data produced in monologue experiments and dialogue experiments, and asked how the resulting patterns of best-fitting weights—the grammars—differed according to the speaking situation.

The monologue and dialogue experiments differed in two key features that may have implications for the types of grammars that speakers might be drawing upon. First, in the monologue experiments the word order was prompted solely by the structure of the preamble, while in dialogue experiments the word order was also primed by the confederate’s response in the previous trial, which always used the word order consistent with the beginning of the sentence. This meant that the word-order prompt in the dialogue experiments was reinforced. Given this situational influence, one difference in the grammars
Individual variation in the structure of bilingual grammars across the monologue and dialogue experiments may emerge in MatchPrompt, the constraint that is responsible for matching word order to the preamble. If so, we would expect MatchPrompt to have a higher weight in the dialogue experiments—at least for Dutch—than in the monologue experiments.

The second key difference between the two experiment types had to do with the location of the switch. In the dialogue experiments, the switch location was primed by the confederate’s utterance in the previous trial, while it was unconstrained in the monologue experiments. This is relevant for the language-activation parameters $\varepsilon$ and $\delta$, which we designed specifically for the dialogue experiments as a mechanism to capture the alignment in switch location that emerged between the participant and the confederate. For monologue experiments, in which the switch location was unconstrained, we fixed the parameter values at one of two settings. If the preamble was in Dutch and a switch prompted, we set $\varepsilon = 0.9$ and $\delta = 0.1$; if no switch was prompted, we set $\varepsilon = 0.1$ and $\delta = 0.9$ (and vice versa, respectively, if the preamble was in English). Effectively, we treated the language-activation parameters as a binary variable, coding the presence or absence of a prompted switch. On the one hand, this simplification could mean that the model might not capture behavior in the monologue experiments as well as in the dialogue experiments. On the other hand, if the $\varepsilon$ and $\delta$ parameters were not a good way of capturing the alignment in switch location in the dialogue experiments, then this could mean that the model would yield a worse fit to the dialogue experiment data than to the monologue experiment data. Furthermore, even if the parameters are well suited to capture both monologue and dialogue switch locations, participants themselves might show qualitatively distinct behavior in the two experimental conditions, resulting in equally good model fits, but distinct patterns of constraint weights.

To explore how these differences in experimental design—and hence speaking situation—affected the structure of the grammar, we divided up the data into two sets, one produced in monologue experiments and one produced in dialogue experiments. We ran 1,000 optimizations on each data set. Of these, 768 optimizations for monologue experiments, and 924 optimizations for dialogue experiments, reached a final KLD below 1. As Figure 6 shows, this difference across experiment type reflects a pattern by which the monologue experiments had a better mean fit (mean KLD 0.716) than the dialogue experiments (mean KLD 0.879), but a broader variation in fits across the 1,000 optimizations (monologue standard deviation: 0.117; dialogue standard deviation: 0.054). Further, comparing Fig. 6 to Fig. 1 also suggests that the results of the dialogue-only analysis are quite similar to the full aggregated analysis in both mean KLD (full mean: 0.890; dialogue mean: 0.879) and standard deviation (full $SD$: 0.061; dialogue $SD$: 0.054).

Figure 7 shows the top 100 best-fitting optimizations for the two experiment types. In the monologue experiments, this corresponded to a KLD range between 0.516 and 0.584, while in the dialogue experiments the range was between 0.755 and 0.794. Of particular interest in Fig. 7 is the fact that the same patterns of variability in constraint weights that we saw in the full population analysis are also repeated here. In both monologue and dialogue data, for example, the linearization constraints are higher in English than in Dutch. With SpecLeft and HeadRight there were a handful of exceptions, with higher weights in Dutch than English (SpecLeft: 35/924 optimizations, or 3.8%, in dialogue experiments, and 15/768, or 2.0%, in monologue; for HeadRight: 53/924, or 5.7%, in dialogue, and 132/768, or 17.2%, in monologue). For HeadLeft, there were no exceptions in the fully aggregated analysis, and only two exceptions in this analysis, both in monologue experiments (0.3%). The same trade-off pattern in
These patterns in both monologue and dialogue data match the patterns in the population-level data closely enough to suggest that the grammars employed in both types of speaking situation do not differ dramatically. Yet at the same time, the higher minimum KLD reached in the dialogue data compared to the monologue data does suggest that the model is not capturing the behavior of speakers in the dialogue experiments as well.
as it could. Something about the language-activation parameters, then, is not working as well as it could be. We return to this issue in §4.

Overall, the population-level and experiment-level analyses show that the aggregate sets of code-switching patterns produced by a population of individuals can be captured reasonably well by grammatical parameter settings that can vary a bit as long as they adhere to a set of patterns regarding their relative values—for example, linearization constraints must have higher English weights than Dutch weights, while MatchSwitch must obey a trading relationship between the English and Dutch counterparts.

To what extent are these patterns detectable in the behavior of individuals? In other words, is it the case that individuals learn this same grammar that emerges from population-level behavior, or might they arrive at different underlying parameter settings, which generate their own idiolects? If the latter, how widely might the grammars vary across individuals, while still in the aggregate yielding the data patterns that emerged in the population-level analysis? To answer this question, we ran a final set of optimizations in order to find the optimal constraint rankings for each of the ninety-three individuals who participated in the experiments.

3.3. Individual-level analysis. Individual participants did not see the full range of possible prompts. In monologue experiments, individuals saw only one of six possible prompts (three word orders crossed with instructions to switch or not to switch), and in dialogue experiments, individuals saw only one of twelve possible prompts (three word orders crossed with four possible switch behaviors). Accordingly, the range of data that was available for the individual-level analysis to model was restricted, relative to the group-level or experiment-level analyses, and especially restricted for the monologue experiments. In light of this fact, we conducted two analyses. First, we ran the model on the experimental data collected from all participants. Then, we ran the model on simulated data, generated by the model from a set of known parameter values, to determine how well the model could recover input weights.

Analysis of experimental data. Because it was not feasible to run 1,000 optimizations for each of the ninety-three participants, we ran fifty optimizations per participant. Figure 8 shows the distribution of KLDs across the 4,650 optimizations (93 individuals × 50 optimizations each), separated by experiment type. Perhaps because we only had one twentieth the number of optimizations, or a reduced amount of data to model for each individual, the final KLDs in the individual-level analysis were substantially higher than in the group-level or experiment-level analyses, indicating overall poorer fit of the model. Further, the difference in fits across dialogue and monologue experiments did not match the results for the aggregated experiment-level analysis. Although a two-tailed t-test confirms that there was a significant difference in average KLD for individuals in the two experiment types, the difference went in the opposite direction from the aggregated analysis: dialogue experiments produced a lower average KLD (1.45) than monologue experiments (1.65; t(4636.1) = −5.94, p < 0.001).

This difference is illustrated in Fig. 8. Although the peaks of the two distributions look as if they are pointed at similar KLDs, the monologue experiments have the same pronounced rightward skew that could be seen in Fig. 6. This skew is responsible for dragging the mean KLD of the monologue experiments upward.

The variation in fits shown in Fig. 8 largely reflects between-subject variability, rather than within-subject variability. Although mean KLDs for subjects could be as low as 0.14 (subject pp_exp1_7) or as high as 7.73 (pp_exp4_42), the standard deviations of final KLDs across each participant’s fifty optimizations were, on average, ap-
approximately 0.41. This is substantially lower than the standard deviation across all by-subject analyses, whose monologue standard deviation was 1.07 and dialogue standard deviation 1.25. In other words, although the individual-level analysis could vary widely in how well the model was able to predict a particular subject’s response distributions, with each individual the model usually converged on a similar degree of fit over the fifty optimization runs. This degree of variability was, nevertheless, higher than the variability in fit across the 1,000 optimizations of the aggregated monologue and dialogue experiments, whose standard deviations were 0.117 and 0.054, respectively. In other words, the larger amount of data used in the aggregated analyses resulted in a more consistent fit of the model, compared to the smaller amount of data used to fit the model in the by-subject analyses.

For the remainder of the by-subject analysis, we retained only those optimizations that resulted in a KLD of 2.5 or less. This cutoff is higher than the cutoff of 1 used in the previous aggregated analyses, but a more lenient criterion was necessary to reduce the data loss in the by-subject analysis. Retaining the cutoff of 1 would have meant discarding the data from thirty-four subjects—over a third of the data. Loosening the cutoff to 2.5 lost only one subject whose results could not be fit by the model, and further allows us to examine the degree of intersubject variability more completely.

Figure 9 shows the patterns of the weights for a sample of four semi-randomly chosen participants, one from each experiment. The colors represent the final KLD for each run. Note that the uniformity of color for each participant’s weights reflects the pattern of substantially smaller within-participant variation in model fit compared to between-subject variation.

The immediate impression that jumps out of Fig. 9 is the decided variability in constraint weights both within and across participants. Although the aggregated analyses presented in Figs. 3 and 7 did show variability, the patterns were nevertheless clearly visible. By contrast, the four participants pictured in Fig. 9 show some evidence of some of these patterns, but they also show evidence of contradictory patterns that are directly opposite the findings of the aggregated analyses.

Consider first participant pp_exp1_7, in the top row. This participant’s optimizations returned a consistently strong fit between model and data, with a mean KLD of 0.145.

\[\text{Experiment type} \quad \begin{array}{ll}
\text{dialogue} & \text{monologue} \\
\end{array}\]

![Figure 8. Distribution of model fits across 4,650 by-subject optimization runs. Despite the overlap, the mean KLD for monologue experiments was significantly higher than the mean KLD for dialogue experiments.](image)

\[\text{Figure 8. Distribution of model fits across 4,650 by-subject optimization runs. Despite the overlap, the mean KLD for monologue experiments was significantly higher than the mean KLD for dialogue experiments.}\]
Their MatchPrompt results show no consistent preference for a higher Dutch weight, and contrary to the aggregated analysis patterns. Further, there is little evidence of preferences for higher Dutch weights—a pattern that was entirely absent in the aggregated analysis.

Participant pp_exp3_15, in the third row, shows more uniformity in preferring higher English weights than Dutch for HeadLeft and HeadRight, consistent with the aggregated analysis. Their preference for a higher English weight is especially strong for Max, however, in decided contradiction to the inconsistent results of the aggregated analysis; and if anything their preference for SpecLeft goes toward a higher Dutch weight than English, also contrary to the aggregated analysis patterns. Further, there is little evidence of the trade-off for MatchSwitch or of the preference for a MatchPrompt higher weighted for Dutch than for English, which were so striking in the aggregated analyses.

In the second row, participant pp_exp2_218 also shows the preference for English Max to outweigh Dutch, but there is very little consistency at all in their linearization constraints. Their MatchPrompt results show no consistent preference for a higher Dutch
weight over English, and their MatchSwitch is characterized more by a tendency for zero-weighting in both languages than any sort of trade-off pattern. Finally, pp_exp4_49, in the bottom row, has a preference for higher English than Dutch weights in HeadLeft, but there is no pattern for any of the other constraints that resembles either the other participants’ individual results or the patterns that emerged in the aggregated analyses.

In summary, then, the by-subject analysis reveals less consistently good fits of the model to the data than the aggregated analyses do. Further, within each subject the patterns of constraint weights can be extremely variable, but they can also reveal consistent patterns that contradict the patterns emerging in the aggregated analyses. This suggests that the grammars governing each individual’s utterances may reflect some of the tendencies that emerge in the aggregated analysis, but the individual grammars can also show quite decided tendencies that do not at all reflect the emergent grammar of the population as a whole.

Analysis of simulated data. To determine how likely it was that the by-subject analysis could recover underlying parameter weights, we also simulated some data and analyzed it in the same way as the individual-level analysis. First, we extracted the parameter weights from one optimization run for a ‘donor’ subject. In this case, the donor was subject pp_exp1_223, from experiment 1—that is, a subject in the monologue condition with Dutch as the frame language. These weights were run through the model to generate a set of probabilities for the different prompt/response pairings.

To simulate the experimental data, we sampled from the set of possible responses for each prompt, using the probabilities to weight how likely each response was to be selected. We repeated this sampling process 100 times, thus creating 100 sets of simulated experimental results, each generated from the same underlying parameter weights. This data was then run through our optimization algorithm, generating fifty sets of optimized weights for each simulated data set. After trimming out optimizations that had KLDs above 2.5, we examined the output weights for each simulation to see if our process could recover the original parameter weights.

Figure 10 shows in gray the full set of optimized weights for the donor subject produced during the individual-level analysis (top panel), with the set of weights used to generate the simulated data in black.

The bottom four panels show the optimized weights produced for four separate sets of simulated data. The four sets of data presented here were selected because they represent the outputs for simulations that had reasonably good fits, ranging from a KLD of 1.61 (simulation 60) to 1.8 (simulation 14). However, the patterns that follow emerged in the other simulations as well. The first result that jumps out is the failure of our simulated optimizations to recover the input values for SpecLeft and MatchPrompt. Instead, those constraint weights ended up hovering around 0, indicating that violations of those constraints—in either English or Dutch—did not seem to penalize any of the candidates in the simulated data sets.

For the other four constraints, however, the optimized weights of the simulated data did replicate the patterns of the input weights. For HeadLeft and HeadRight, the optimized weights in all four simulations correctly reproduced a pattern of higher English than Dutch weights, and for Max the optimized weights in all four simulations correctly reproduced the pattern of higher Dutch than English weights. Most interestingly, the optimized values for MatchSwitch, although generally higher than the donor values, still in aggregate show signs of the trade-off pattern between Dutch and English weights that was present not only in the full set of optimized weights for the donor participant, but indeed in every analysis of all subsets and supersets of the experimental data.
In summary, the analysis of simulated data reminds us to be cautious in interpreting the weights emerging from the individual-level analysis. Although the model can recapture the underlying parameter weights for some constraints, it fails to recapture the underlying weights for others—in this case, SpecLeft and MatchPrompt. It was beyond the scope of this analysis to run multiple such simulations, to see whether some constraints were systematically more vulnerable than others. Nevertheless, the reemergence of familiar patterns, such as the trade-off between Dutch and English MatchSwitch, even when the input parameter values were fixed, provides further evidence that a variety of parameter values can generate equivalent outputs.

3.4. Consistency in results across analyses. The focus of this article was to explore the degree of variability that was possible within the range of possible grammars our GSC architecture could generate in modeling our empirical data. Yet it is equally important not to overlook the repeating patterns that emerged across the 7,650 optimizations we ran in our three levels of analysis.

First, the GSC architecture could describe the distribution of participant responses with a distribution that lost less than one bit of information. Yet despite this ability to fit the data closely, it nevertheless emerged that the grammars for monologue experiments produced more variation in their ability to fit the data than grammars for dialogue experiments, although in aggregated analyses this variation did not obscure an ability to fit the data more closely. Figure 11 illustrates these tendencies. In the experiment-level analysis, the pattern of better fits to monologue data is extremely clear. In the subject-level analysis, the wider variability in fit for monologue data is also clear.

The second pattern to consider is the tendency for higher weights for English-specific linearization constraints, and higher weights for Dutch-specific MatchPrompt.

Figure 10. Optimized weights for subject pp_exp1_223 (gray, top panel), with donor set of weights in black. The bottom panels show optimized weights (gray) produced for four representative simulations of the experimental results, generated using the donor set of weights (repeated in black for each simulation). To the extent that the gray weights match the black weights in the bottom four panels, the model has recovered the underlying parameter values used to generate the data.
Figure 12 illustrates how often this pattern held across the three levels of analysis and two experiment types. It shows that, although the majority of optimizations adhere to the patterns named in each panel, they do not adhere to the same degree across different analyses. English linearization constraints outweighed Dutch linearization constraints in almost all optimizations when the data was aggregated across the whole group (black bars) or across experiment types (middle bars in each panel); and the same is true of the pattern by which Dutch MatchPrompt outweighed its English counterpart. Yet when the data was analyzed across individuals (rightmost bars in each panel), a smaller proportion of optimizations show each pattern. They still form a majority, but not as decided a majority. Although distinct patterns of constraint weights emerged repeatedly, they by no means constitute the only possible way to generate the response patterns in the experimental data.

Figure 11. Distribution of KLDs across all 7,650 optimizations, broken up by experiment type. On average, dialogue experiments had higher (worse-fitting) KLDs than monologue experiments.

Figure 12. Patterns of relative constraint weights across the three levels of analysis: aggregated across the whole population (black bars on left of each panel; §3.1), aggregated by experiment type (middle set of bars in each panel; §3.2), and separated by individual participant (right-hand set of bars in each panel; §3.3). Dark gray bars show data from dialogue experiments, and light gray bars data from monologue experiments. Dashed line indicates 50% adherence to the pattern in the named panel.
4. Discussion. Using existing experimental data on Dutch-English code-switching (Kootstra et al. 2010), we tested whether the architecture of GSC (Goldrick et al. 2016, Smolensky et al. 2014) could model the production of code-switched sentences produced by bilinguals, who must balance two interacting grammars. Further, we wanted to see whether this model could provide insight into the following three areas of inquiry: (i) the structure of a population grammar governing the emergent linguistic patterns to be found across an entire speaker group, (ii) the variability of the population grammar across different speech situations, and (iii) the relation between the population grammar and the grammar of the individuals who make up that population. Is there only one unique grammar that can capture the patterns of variability over utterances within a group of speakers? Does this grammar shift around depending on specific external speech contexts? Do individuals all draw on the same grammar when they speak the same language? The results of this analysis suggest that the answers to these questions are, respectively, ‘mostly’, ‘probably’, and ‘definitely not’.

The first takeaway from our analysis is a proof of concept that GSC can be productively applied to model bilingual language data. Van Hell et al. (2016) showed that the GSC architecture is amenable to expansion to describe characteristics of bilingual grammar that were not discussed in Goldrick et al. 2016—in their case, lexically specific properties such as cognate status—but they were not able to test whether that expansion accurately captures patterns in the data. The current work expanded the model further, adjusting the architecture to capture code-switching behavior across unbalanced bilinguals, in different speaking contexts that produce unequal lexical activation, and then tested that expansion on actual speaker productions. After fitting the model to the data, the model generated probability distributions that could capture the distribution of responses with less than one bit of information lost. In other words, the work presented here shows that the GSC architecture can be successfully expanded and adjusted for different types of data and speaker groups, and still captures empirical patterns reasonably accurately.

It should not be overlooked that the data analyzed here is an artificially simplified subset of the variety of language that speakers would actually produce outside of the lab, and the model structure we present is a simplified version of what speakers would actually draw upon to generate their utterances. Nevertheless, this successful expansion of the applications of GSC from its earlier applications (Goldrick et al. 2016, Smolensky et al. 2014, van Hell et al. 2016) shows the potential of the model to be developed until it is ready to tackle a wider variety of naturalistic speech. Further, the successes of the model even in the current experimental context allow us to explore the types of insight that might emerge from further research. In what follows, we discuss the potential of this model in analyzing dialogue (§4.1) and variation in individual grammars (§4.2), and in generalizing to new domains (§4.3) such as language attrition and language acquisition, and we even speculate on how the term ‘grammar’ should best be defined (§4.4).

4.1. Monologues, dialogues, and interactive alignment. The comparison between the population-level and experiment-level analyses reveals certain shortcomings in the current implementation. Specifically, the higher KLDs that emerged in the data from the dialogue experiments suggest that the model could not capture the patterns in dialogue situations as well as it could in monologue situations. This pattern emerged especially clearly when the dialogue data was modeled as a whole, although it did not apply to the by-subject-level analysis. We can think of several accounts for this shortcoming, any or all of which may be at play.
A prosaic explanation could be that we employed certain mistaken assumptions in applying the GSC architecture to this new type of data. Consider first the language-activation parameters $\varepsilon$ and $\delta$. We included them as a mechanism to generate the switch-location priming effect in the dialogue experiments. In the original analysis, Kootstra et al. (2010) observed that if the experimental confederate switched in a particular location, participants were more likely to switch in that location too. By modeling switch location through dynamic changes in language activation, we could penalize the omission of more active items compared to less active items, and thus penalize too-early or too-late switch locations. In this way, we thought that our model could capture the alignment of switch locations between participants and confederates. Yet by penalizing the inclusion of less active constituents in a target sentence, the model rewarded not only switches in the prompted locations, but also any other set of switch locations that yielded the same proportion of English and Dutch constituents.

We implemented the language-activation parameters as a way of exploring whether activation of lexical items might be the mechanism by which switch locations could be primed. Yet the poorer fit of the model to the data in the dialogue experiments, where exactly such a mechanism should have been most active, suggests that these language-activation parameters may not be the mechanism by which switch locations actually are primed.

Another explanation for the difference in model fits relates to the details of the original experiment from Kootstra et al. 2010. In the dialogue experiments, certain conditions required the confederate to model unattested or infelicitous code-switching patterns. Consider 8, for example, with its combination of a Dutch word order (SOV) and an English verb.

\[(8)\] Een grappig plaatje, waarop [het meisje$_{S,Du}$ de ezel$_{O,Du}$ chases$_{V,En}$].

\[a\] funny\ [\[the\ girl\ the\ donkey\ \[chases\]]\]

According to Myers-Scotton’s (2002) matrix language frame (MLF) model, certain types of grammatical elements should come from one language only (i.e. the matrix language). According to the uniform structure principle, which is part of the MLF model, the structure of a sentence should be grammatically uniform, and in bilingual sentences, structures of the matrix language are always preferred (see Deuchar 2020 for an overview of the different perspectives on how to define the matrix language). This means that, according to the MLF model, not only certain grammatical morphemes but also the structure of a code-switched sentence should ideally come from one language only. Thus, sentences like 8, with word order from one language and verb conjugation from the other, are not consistent with the MLF model.

Although confederates did produce these types of switches (see §1.2), participants were reluctant to reproduce them. Instead, they would change either the switch location or the word order. For example, if they maintained the SOV word order, they produced the English component of the sentence early—for example, at the subject—and then switched back to Dutch for the object and verb: *Een grappig plaatje, waarop the wizard$_{S,En}$ de eiland$_{O,Du}$ roept$_{V,Du}$. Another solution was to switch the word order to SVO, which is compatible with both languages: *Een grappig plaatje, waarop de tovenaar$_{S,Du}$ calls$_{V,En}$ the moose$_{O,En}$. In general, then, some conditions were more suited to trigger alignment between confederate and participant than others, while our GSC implementation assumed that each experimental condition was much the same as the other. This assumption seems to have worked perfectly well in the monologue experiments, with
Individual variation in the structure of bilingual grammars

no confederate present to prime infelicitous switch patterns, but it could have been responsible for the poorer fit of the model to the dialogue data.

A third explanation for the difference in model fits between monologue and dialogue experiments is that the language-activation parameters are capturing the mechanism by which switch-location alignment occurs—that the alignment does occur when an interlocutor’s behavior activates similar behavior in the speaker—but that the alignment takes time to reach completion. If the mechanism for switch-location priming were truly captured by differences in the activation levels of the particular constituents, then the participant should, in principle, match the switch location of the confederate as closely on the first trial as on the last, because the confederate’s switch location would trigger the same pattern of lexical-item activation in the first trial as in the last. Yet structural priming effects are cumulative (Kaschak et al. 2014) and can accrue across languages in bilinguals (Kootstra & Doedens 2016), suggesting that language users continuously adapt their linguistic choices to the ongoing linguistic environment. This idea of continuous adjustment of language structures encompasses phenomena well beyond syntactic choices. The Five Graces Group et al. (2009), for example, argue that language grammars are part of a complex adaptive system, emerging from individuals’ linguistic experience of usage patterns and changing dynamically as individuals’ experiences change. Kleinschmidt and Jaeger (2015) propose a similar view on the phonetic level, employing a Bayesian belief-update model to capture how new input can change listeners’ internal models for interpreting acoustic input. To account for the observed pattern of switch-location alignment, Kootstra et al. (2010) drew on a third, related model: Pickering and Garrod’s (2004) interactive alignment model.

Interactive alignment employs an activation-based mechanism to account for how interlocutors come to employ the same linguistic behavior when they are in dialogue. A speaker’s use of some linguistic structure activates it in the listener, who as a result employs it in turn, thus reinforcing that heightened activation throughout the conversation. Thus far, the interactive alignment model is similar to our computational implementation of the language-activation parameters $\varepsilon$ and $\delta$. But interactive alignment is more than that: it applies at all levels of linguistic behavior, from phonetic realizations all the way up to the speaker’s internal situation model of the discourse. Although such alignment can begin to take place quickly, it cannot fully ripen with only one or two turns of a conversation.

Perhaps, then, the poor fit of the model to the data in the dialogue conditions is not due to a fault in the model architecture, but rather reflects the fact that the model was trying to capture a moving target. The same fixed set of possible $\varepsilon$ and $\delta$ values were employed for the entire dialogue data set, both when aggregated into one set and also when analyzed by individuals. Yet if the interactive alignment between confederate and participant took place gradually, then the set of $\varepsilon$ and $\delta$ values, which capture the priming effect of the confederate’s switch location, might also have shifted gradually over the course of the experiment. If this is the case, then it is to be expected that the model will always fit dialogue data less well than monologue data, because a fixed set of $\varepsilon$ and $\delta$ parameters like the ones we employed can never fit a moving target as well as they will a fixed target.

But even if the alignment does take place quickly, and the $\varepsilon$ and $\delta$ parameters do rapidly arrive at a set of values that capture the speaker’s internal grammar state, this particular experimental setup enforced a degree of speaker variability in that grammar state. Because the dialogue experiment systematically varied word orders and switch
locations in the confederates, there was not any stable behavioral pattern for the speakers to align to. As a result, the $\epsilon$ and $\delta$ parameters, even if they could capture gradual change as the speaker aligns to the confederate in stable conditions, would have much more difficulty capturing constantly shifting speaker alignment in the variable experimental conditions.

**4.2. Group vs. Individual Grammars.** Although the degraded fit of the model to the data in the dialogue condition does require further consideration, the results of the group-level and experiment-level analyses suggest that the GSC architecture can model bilingual grammars reasonably well. With this tool at our disposal, we were able to probe other questions of bilingual language use. One result that emerged from the population-level analysis was the repeated emergence of consistent patterns of constraint weights. Although these weights were not equivalent in absolute values, they consistently adhered to systematic relationships, such as the trade-off pattern for MatchSwitch or the inequality illustrated in Fig. 4 for MatchPrompt. In other words, there are a variety of parameter values that can describe extensionally equivalent grammars, but those values are governed by consistent, higher-order relationships.

We wanted to know whether individual behavior also showed these same relationships. Were individuals drawing on the same sets of grammars that emerged from aggregated group-level linguistic behavior? As Fig. 9 showed, they were not. Although some patterns observable in the population-level analysis were replicated in some individuals, they by no means applied to all individuals, as summarized in Figs. 11–12. Indeed, some individuals seemed to draw upon grammar structures that were quite distinct from the structures characterizing the group behavior aggregated across individuals—as seen, for example, in the seeming preference for Dutch SpecLeft to outweigh its English counterparts in pp_exp3_15 (third row of Fig. 9) or in the consistent preference for English Max to outweigh Dutch Max (top three rows of Fig. 9).

This finding is made more tenuous by the fact that each individual responded to only a subset of the full prompts. What is more, our analysis ran only fifty optimizations per individual, rather than 1,000, and as the simulated data showed, analyzing experimental data at this scale is not guaranteed to recover the underlying parameter values. It may well be the case that a similar analysis would converge on a much more consistent and accurate set of grammatical structures within each individual if the models had a larger variety of data to use during optimization. Yet it is also the case that individuals can never produce as much quantity or variety of output in their own utterances as a full speech community can generate in the aggregate. Our conclusions, then, rest on the assumption that the reduced variety of utterance types in our individual-level analysis is a reasonable reflection of the reduced variety of utterance types produced by an individual speaker relative to the larger speech community.

To the extent that this assumption is valid, then the patterns from the by-subject analysis indicate that the set of internal grammars which individuals draw upon to produce their utterance are not necessarily governed by the same parameters that best describe the full set of utterances to be observed across a population.

**4.3. Generalizing the Model.** Although we focused primarily on the pattern of constraint weights in this analysis, our model architecture includes multiple degrees of freedom in both the parameters and the fitting architecture, which we believe can be explored for further insights into the behavior of human languages. The language-activation parameters $\epsilon$ and $\delta$, for example, as currently defined in Table 1, cannot be used outside of the context of this experiment. Yet they have the potential to distinguish between all shades of monolingual and bilingual modes a speaker might encounter (Gros-
Individual variation in the structure of bilingual grammars (Jean 1989). As another example, consider the six constraints that we modeled in this work: of course, no one pretends that six constraints are sufficient to model any natural human language. One particular strength of this model architecture is its ability to be generalized and adapted to provide insights into new linguistic phenomena. In the following sections, we consider how the insights from this analysis alone can shed light on two other phenomena that are particularly relevant to the study of bilingualism: language attrition and language acquisition.

Proficiency effects and language attrition. Our language proficiency parameters, which we determined as a function of lexical-decision scores and self-rating, were developed to reflect the fact that the participants were more proficient in Dutch than in English. The action of these parameters is to adjust the language-specific constraint weightings; and since English was less activated than Dutch, an English-specific constraint will need to have correspondingly higher weights than the Dutch-specific constraints in order to produce the equivalent set of violations. With this in mind, we can see that the pattern of higher weights for the linearization constraints SpecLeft and HeadLeft in English than in Dutch may not actually reflect the fact that English grammar restricts word order more than Dutch grammar does. Rather, they may indicate instead that the linearization constraints in English and Dutch operate similarly, differing only in that they must be weighted more heavily in English to compensate for the lower overall activation of English constraints. This would mean that the word-order differences between English and Dutch are governed almost entirely by the MatchPrompt constraint, which is weighted much more heavily in Dutch than in English. Indeed, this disparity would be even more exaggerated if English and Dutch had equivalent activations.

This understanding of affairs—that linearization in English and Dutch works similarly and differs only to the extent that Dutch subordinate clauses respect the word order dictated by their syntactic context—captures the pattern observed by Kootstra et al. (2010) in their original analysis. They observed that Dutch does indeed have an SVO preference, just like English, on which is overlaid a respect for the word order required by the matrix clause. This observation would align well with a model in which linearization constraints have similar weights for English and Dutch, and the primary difference in word order is governed by MatchPrompt.

This interplay between language proficiency, captured by the language proficiency parameters, and grammatical structure, captured by constraint weights, suggests that bilinguals have a variety of options in their grammatical structures, which can change as the balance of proficiency in their languages changes. An unbalanced bilingual may be able to generate proficient-seeming utterances—that is, utterances matching the patterns in the population—simply by increasing the constraint weightings in their less proficient language to compensate for the reduced activation. Yet as their proficiency increases, these higher constraint weightings would result in utterances that diverge from the population norms, unless they downweight the constraints to compensate for the increased proficiency. Similarly, an originally balanced bilingual who begins to undergo language attrition as one language is used less and less frequently may be able to continue producing proficient-seeming utterances by increasing the constraint weights of their attriting language to counteract the decrease in baseline language proficiency. This trade-off between constraint weights and proficiency could be another internal source of variation in the progression of language acquisition and attrition across individuals.

These roles of constraint weights and proficiency are also at the core of the competition model (e.g. MacWhinney 2005, 2008), which is often used to account for cross-language interactions in sentence processing of bilinguals and L2 learners. The
competition model proposes that sentence processing in bilinguals is based on interaction between competing linguistic cues from both languages, whose weights are, among other things, based on relative language proficiency. Thus, the roles of constraint weights and proficiency that we discuss here can be viewed as a computational implementation of some of the model’s core ideas.

**Parameter optimization and language acquisition.** We originally developed our parameter-optimization algorithm as a workaround, since the MaxEnt Grammar Tool (Hayes 2009) could not operate on our bilingual architecture with varying proficiencies and fractional violations. Yet our algorithm has its own shortcomings, not least of which is the computational capacity required to run a sufficient number of optimizations to ensure that the patterns are reproducible. Yet we find the random component and concomitant variability in our optimization to be intuitively satisfying for other reasons.

We have already discussed how the GSC architecture might offer insights into how grammars change over time, as proficiencies shift. If it is also used to model language acquisition, then our optimization algorithm provides an intuitive functional parallel to how learners might acquire a GSC-like language-production system. During acquisition, learners do not follow monotonically increasing developmental trajectories. They might regress, or hit plateaus, or oscillate around a certain level of attainment; and, crucially, these learning patterns can vary across individuals, both children acquiring L1 (van Geert & van Dijk 2002) and adults acquiring an L2 (de Bot et al. 2007). Oscillations in improvement and regression observed in learners are functionally similar to the oscillations in correlation between the model predictions and the actual data that result from the model walking through parameter space in an attempt to discover better-fitting weights at each iteration. Plateaus in learning mirror the tendency of the model to get stuck at local maximum correlations before finding a set of parameter values that allows the model to jump out of that parameter space. It would be interesting to see whether the types of distributions generated by the optimization algorithm bear any relation to the types of utterances learners make across their acquisition trajectories.

**4.4. What is grammar?** Throughout this article we have been allowing for the possibility that a connection can be drawn between abstract grammars, which emerge in an individual as they generalize patterns from raw input, and speech communities, as individuals converge on mutually comprehensible patterns of output. We have further allowed for the possibility that an abstract grammar can be distinguished from more biological cognitive processes, such as executive function and working-memory capacity, and also more individual experiences, such as language dominance and proficiency. In our model’s architecture, these components of linguistic knowledge can be modeled separately. Language proficiency is modeled in our language proficiency parameters; situational variation in language activation is modeled in our \( \varepsilon \) and \( \delta \) language-activation parameters. For those who distinguish between processing and grammar, it is only the constraint weights that constitute the grammar in this model.

This sharp distinction between processing and grammar, however, is by no means clear-cut. Recent work has emphasized that the architecture of the language-processing system can be better understood by focusing on the role of individual differences (Kidd et al. 2018). Fricke et al. (2019) argue that individual differences are particularly likely to shape language processing in bilinguals, so that research in bilingualism is funda-
mentally important to understanding how language works. These components—abstract structures, linguistic experience, and cognitive processes—are not fully independent, and in light of this fact, it may well be useful to consider the full ‘grammar’ to consist of all the free parameters in our model—not just the constraint weights themselves, but also the representations of language proficiency and dominance.

We have, then, two ways of defining a grammar in this model. The abstractionist perspective would distinguish constraint weights from other components of this model, while the more global perspective would include the associated processes that affect how any abstract structures are used. Under either definition, however, the work presented here has shown that the ‘grammars’ emerging from our analysis are characterized by considerable variability—both across and within speakers. Both definitions include at a minimum the set of constraint weights that we generated with our optimization process, and it is exactly these constraint weights that show this variability.

The cross-speaker variability would be interesting enough. It is the source of our claim that the grammar the individual acquires does not necessarily match the grammar characterizing the input they use to infer it. More interesting, though, is the within-group and within-speaker variability, which reveals how multiple parameter settings can equally well model the data that we fed into the model. Under a traditional understanding of how these models operate, each set of parameter values constitutes a distinct grammar. It is exactly the different weights of the Dutch-specific and the English-specific constraints here, or of the English-specific and Tamil-specific weights in Goldrick et al. 2016, that constitute the differences between English, Dutch, and Tamil grammars. But if multiple sets of weights—multiple grammars—can generate the same data, then can we know which of these extensionally equivalent grammars listeners employ? We see three ways to answer this question.

Accept ignorance. One possibility is to resign ourselves to ignorance. We can only infer grammars from output, and choose between potential candidates on the basis of other desirable principles—parsimony, generalizability, and so forth—that we want to hold for a given grammar (or grammatical theory). In this case, the various different sets of constraint weights are all equally economical and generalizable as grammars. If they do not predict different behavior, we cannot know which grammar the listeners are employing to produce the behavior that we observe.

Analyze more data. The second solution is a more prosaic appeal to ecological validity. We have seen that the smaller variety of utterances produced by individuals results in more variability than the larger variety of utterances that were analyzed for the aggregated group-level analysis. We have also seen from the simulations that the amount of data available for the by-subject analysis may not be sufficient to recover all underlying parameter values. Yet even the group-level analysis was based on a constrained set of data, collected in an experimental setting that enforced a type of code-switching that participants may not often employ in daily life. This source of data was convenient, because its restricted variety allowed us to model it with a minimal set of constraints. But of course it does not represent the full set of possible utterances that the speakers are capable of producing, and even with this constrained data set, we did see that certain patterns of constraint weights were more common than others. Perhaps with more variety of utterances to analyze, our model would be able to converge on the underlying parameter values that are responsible for generating that data.

19 Their self-rated code-switching ability averaged 2.9 on a five-point scale.
Redefine ‘grammar’. The third solution is to redefine our understanding of ‘grammar’ under this model, in a more radical way than simply including processing abilities under the umbrella. Rather than saying that a single grammar is defined by a single set of parameters, we instead associate distinct grammars with the set of extensionally equivalent parameter values (or higher-order patterns of parameter values) that describe the data we want to model. Some patterns of parameter values may be more commonly used than others (as we saw in Fig. 12), but they are not the only options. A given individual may draw on only one set of parameter values throughout their life, or they may draw on many, changing dynamically across situations or over time. The trade-off between proficiency and language-specific weights we discussed in §4.3 is one example of how this shifting internal grammatical structure may play out.

Under this understanding, the claim that individuals have different grammars for the same language is akin to saying that different people have different pizza preferences. There is a defined space of variability—the language, the concept of pizza—and within the space there are a variety of options—constraint weights, pizza toppings—they can employ to realize the concept. Speakers can vary in which of those options they prefer, resulting in largely but not entirely overlapping sets of preferences, unique to each individual. Which grammar realization is selected to generate any one utterance can then vary dynamically across situations or lifespans.

5. Conclusion. The analysis presented here suggests that the gradient symbolic computation architecture can be fruitfully applied to bilingual language data, and offers insights into the underlying structures of bilingual grammars. By applying this architecture to the speech produced by specific individuals, we were able to show that the individual grammars need not share the same properties that characterize the speech of a larger population—although this property may also reflect differences in the amount of data each individual can furnish. Further, the variability of extensionally equivalent grammars that emerged within individuals and aggregated analyses suggests that the term ‘grammar’ might need redefinition. Rather than denoting a distinct, unique generative structure, ‘grammar’ might more profitably be conceptualized as a range of possible forms of a generative structure, which can vary continuously across individuals and contexts.

Our approach was not without its complexities and challenges. We needed to introduce three levels of parameters—language activation and language proficiency, as well as constraint weights—to capture the multifaceted sources of variation in bilingual competence. This added complexity paid off, by providing a mechanism for an individual’s bilingual grammar system to change dynamically over time. In the short term, adjustments in language activation across different speaking situations can in principle model the unfolding interactive alignment between interlocutors; and in the long term, trade-offs between constraint weightings and language proficiencies can describe individual variation in language acquisition and attrition. Yet these different parameter types had drawbacks: the poorer fit to the dialogue experiments compared to monologue experiments suggests that our language-activation parameters were not adequately capturing the dynamic changes in alignment between interlocutors; and the complex model structure necessitated the use of a computationally slow optimization algorithm to find the optimal set of constraint weights. Nevertheless, we are encouraged by the success of this analysis, and we look forward to seeing further research investigating whether the predictions of this model accurately capture patterns of bilingual language variation across contexts, individuals, and time.
REFERENCES


