Maria Polinsky*

Structure vs. Use in Heritage Language

DOI 10.1515/lingvan-2015-0036
Received May 1, 2016; accepted June 9, 2016

Abstract: This paper provides an overview of the phenomenon of heritage language and offers evidence in support of representational differences between (baseline) native grammars and heritage grammars, arguing that such differences cannot be reduced entirely to the effects of processing constraints or memory limitations. Data from heritage Spanish number/gender agreement and from Russian ellipsis indicate that baseline native grammars and heritage grammars may have a fundamentally different organization of certain categories.

Keywords: Heritage language, bilingualism, agreement, ellipsis

1 Setting the Stage

In his classical (if somewhat forgotten) paper “Literate and illiterate speech” (1927), Leonard Bloomfield describes several speakers of the Algonquian language Menomini (Menominee), whom he worked with in Wisconsin:

Red-Cloud-Woman, a woman in the sixties, speaks a beautiful and highly idiomatic Menomini. She knows only a few words of English, but speaks Ojibwa and Potawatomi fluently, and... a little Winnebago.... Linguistically, she would correspond to a highly educated American woman who spoke, say, French and Italian in addition to the very best type of... English.

<...> Stands-Close, a man in the fifties, speaks only Menomini. His speech, though less supple and perfect than Red-Cloud-Woman’s, is well up to standard. It is interlaced with words and constructions that are felt to be archaic...

<...> White-Thunder, a man round forty, speaks less English than Menomini, and that is a strong indictment, for his Menomini is atrocious. His vocabulary is small; his inflections are often barbarous; he constructs sentences on a few threadbare models. He may be said to speak no language tolerably. His case is not uncommon among younger men, even when they speak but little English.

Little-Doctor who died recently in his sixties, spoke English with some Menomini faults, but with a huge vocabulary and a passion for piling up synonyms. In Menomini, too, his vocabulary was vast; often he would explain rare words to his fellow-speakers. In both languages his love of words would sometimes upset his syntax...

Little-Jerome now in his fifties, is a true bilingual. He speaks both English... and Menomini with racy idiom, which he does not lose even when translating in either direction. He contrasts strikingly with men (usually somewhat younger) who speak little English and yet bad Menomini. (Bloomfield 1927: 437)

Bloomfield’s description is essentially a snapshot of a speech community on the wane, where fluent monolingual speakers co-exist with bilinguals who show significant variance in competence. Little-Doctor has some deficiencies; White-Thunder is at the low extreme of bilingualism; Little-Jerome’s “striking” status as “a true bilingual” reflects the tacit assumption that bilinguals tend to be unbalanced (Grosjean 2010). Both Little-Doctor and White-Thunder would today be called “heritage speakers”: unbalanced, possibly recessive, bilinguals. Their Menominee is less fluent than that of monolingual speakers, and their respective paths to language acquisition may have differed significantly. White-Thunder in particular, the youngest in his cohort, may instantiate the growing trend away from the ancestral language and toward a greater use of English, however “atrocious.”

Today, the Ethnologue lists Menominee as moribund, with only a couple dozen fluent speakers, all elderly. In fact, with speakers like Red-Cloud-Woman and Stands-Close long gone, one wonders what the criteria for “fluency” are in the 2000s. The trajectory of Menominee’s loss is typical: once “the conquering language” – the dominant language of the society – takes over, it is a matter of only a generation or two

*Corresponding author: Maria Polinsky, Department of Linguistics, University of Maryland, 1401 Marie Mount Hall, College Park, MD 20742–7505, USA, E-mail: polinsky@umd.edu
before the original language disappears. Heritage language (HL) is often a step along this path: it is the language of a generation of speakers who grew up hearing their ancestral language but are more comfortable with the dominant societal tongue.

Students of bilingualism acknowledge a continuum of proficiency, with speakers like Little-Jerome close to monolingual, speakers like White-Thunder barely functional, and speakers like Little-Doctor somewhere in the middle. In (1) below, the “acrolectal” speakers on the left edge may differ from monolingual speakers only in subtle details of register variation, literacy, or rare constructions; conversely, “basilectal” speakers often possess only a marginal ability to understand speech in a given language, and are entirely unable to produce it (see also Polinsky and Kagan 2007).\footnote{Of course, while continua allow us to depict dazzling variance efficiently, they can only represent differences, without explaining them.}

(1) Bilingual continuum

Acrolectal speakers > Mesolectal speakers > Basilectal speakers

Where does HL fit into this continuum? Rothman (2009: 156) provides a helpful discussion:

A language qualifies as a heritage language if it is a language spoken at home or otherwise readily available to young children, and crucially this language is not a dominant language of the larger (national) society. Like the acquisition of a primary language in monolingual situations... the heritage language is acquired on the basis of an interaction with naturalistic input and whatever in-born linguistic mechanisms are at play in any instance of child language acquisition. Differently, however, there is the possibility that quantitative and qualitative differences in heritage language input and the introduction, influence of the societal majority language, and differences in literacy and formal education can result in what on the surface seems to be arrested development of the heritage language or attrition in adult bilingual knowledge... An individual qualifies as a heritage speaker if and only if he or she has some command of the heritage language acquired naturalistically..., although it is equally expected that such competence will differ from that of native monolinguals of comparable age.

Several crucial points can be extracted from this definition. First, heritage speakers (henceforth, HSs) are bilinguals, and their L1, i.e., the language they learned in childhood, is not their dominant language; instead, they are dominant in L2, the language of the society they live in. Second, the temporal relationship between L1 and L2 is not defined; HSs may be simultaneous or sequential bilinguals, or perhaps even early L2 learners (speakers who acquire a second language roughly between ages five and twelve) (Bialystok and Hakuta 1999; McLaughlin 1984, 1985). Third, acquisition of an HL is (to a point) similar to acquisition of a monolingual L1: it proceeds naturalistically based on input from a baseline and learning hypotheses rooted in universal grammar (Rothman’s “in-born linguistic mechanisms”). This acquisition route separates HSs (narrowly defined) from the broader population of people whose family heritage includes some connection to language X but who did not grow up with it (Polinsky and Kagan 2007; Rothman 2009).

Naturalistic learning by monolinguals and HSs differs, however, in at least two respects: amount of input and degree of mastery. Consider: a monolingual learner spends all of her awake time hearing one language, while an HS divides the same time among two or more languages; sheer volume of L1 input to the monolingual learner is therefore necessarily higher. The number of speakers providing input to a monolingual learner is also greater: input to an HS may come only from a few family members. Furthermore, the exposure of heritage learners to the home language diminishes as these speakers get older and socialize in the dominant societal language (see He 2014, 2016; De Houwer 2007; Birdsong 2016; Unsworth 2016; Treffers-Daller 2016, for discussion). As for range of attainment, monolinguals are expected to master all (or the majority of) the registers of their language. Among HSs, on the other hand, “the possibility [of] quantitative and qualitative differences in HL input, the influence of the societal majority language, and differences in literacy and formal education can result in what on the surface seems to be arrested development of the HL or attrition in adult bilingual knowledge” (Rothman 2009: 156).
The question of HSs’ attainment has evoked heated debate in the literature, much of which surrounds the following questions:

(2) a. are there observable differences between monolingual and heritage grammars?
   b. if yes, are these differences in linguistic knowledge or differences in linguistic behavior?

The initial answer to question (2a) is a qualified ‘yes’. The evidence amassed suggests that, despite varying demographics and proficiency levels, HSs are remarkably consistent across a number of parameters that set them apart from both baseline native speakers (NSs) and L2 speakers (L2ers).

In the realm of phonology, HSs clearly diverge from NSs in terms of pronunciation and prosody (Godson 2003, Barlow 2014, Chang 2016), while Yang (2015) shows that HSs’ perception and production of tone falls between that of L2ers and NSs. In general, HSs consistently outperform L2ers on perception tasks, but do not always demonstrate native competence: HSs pattern with NSs on a number of straightforward perception tasks (cf. Lee-Ellis 2012 on Korean; Lukyanchenko and Gor 2011 on Russian), but well below NSs on certain more complex tasks (Lee-Ellis 2012). Within inflectional morphology, HSs are known to eliminate irregular forms (Benmamoun et al. 2013a, 2013b) and struggle with agreement and case marking (Polinsky 2006; Scontras et al. submitted). Syntactically, HSs tend to impose rigid word order where NSs allow for flexibility (Isurin and Ivanova-Sullivan 2008; Ivanova-Sullivan 2014); they also unnecessarily limit their inventory of syntactic dependencies (Polinsky 2011). Finally, evidence from Chinese and English suggests that heritage grammars lack quantifier scope ambiguities (Scontras et al. 2016). Although the full range of differences between native language and HL is still being evaluated, three main categories of difference can be identified: transfer (from the dominant grammar), divergent attainment (Rothman’s “arrested development of the heritage language”), and attrition over the lifespan (see Scontras et al. 2015 for a detailed discussion of these categories).

Many studies have examined the processing limitations faced by HSs due to lack of experience and interference from the dominant language – differences that are likely to manifest themselves in production errors (e.g., Tsimpili et al. 2004). The present paper asks a deeper question: are there actual differences in the grammatical knowledge of HSs versus monolingual NSs? In other words, per (2b) above: is the variance in heritage speech purely behavioral, or do monolingual-heritage differences reflect fundamental differences in linguistic representation? In what follows, I present emerging evidence in favor of a difference in underlying grammar between HSs and NSs.

2 In support of differential representations

2.1 Initial hypotheses and ways of testing them

HSs and NSs do indeed show differences in their language performance, as outlined above. In considering the possible roots of such differences, let us take it as our base hypothesis that HSs possess the same knowledge as NSs: a heritage and a monolingual-native speaker of Mandarin both “know” the same Mandarin. If this null hypothesis stands, then any differences between these speakers must reflect the use rather than the structure of their language. In this case, observed differences are likely to be

2 A number of researchers argue that HSs are in fact NSs in their own right (Rothman and Treffers-Daller 2014; Otheguy 2016; Scontras et al. submitted). While this is a reasonable terminological argument, there remains a need to contrast (bilingual) HSs with speakers of the (monolingual) baseline. In this paper, I refer to the latter type of speakers as NSs. New terminology may be developed in the future to better describe both groups.

3 For studies comparing the phonetics in HSs to that in L2ers, see Tees and Werker (1984); Oh et al. (2003); Au et al. (2002).
symptomatic of a shortage of online resources. Processing difficulties may become apparent both in language production and comprehension, where HSs may be impeded by memory limitations, task costs associated with language control (Van Hell and De Groot 1998), or lack of confidence in their judgments (Orfitelli and Polinsky 2016).

An alternative hypothesis would hold that at least some differences between NSs and HSs are symptomatic of deeper, structural differences in the heritage grammar. In other words, the two groups of speakers may have different representations of their language. To test this hypothesis, data on HS comprehension are needed. As mentioned earlier, production is a resource-intensive task that may be subject to extralinguistic limitations; as such, production tests introduce processing-based complications into studies of language behavior in HSs. At the same time, production allows speakers, be they native or near-native, to avoid areas of difficulty (cf. Kleinmann 1977, for L2ers, and Valdés 2005, for HSs). Finally, HSs at the low end of the proficiency spectrum may lack significant production ability, leaving comprehension testing as the only option for these speakers (e. g. Sherkina-Lieber et al. 2011, on Inuit HSs).

Of course, comprehension testing offers an indirect window on language knowledge, and the success of the enterprise depends largely on the nature of the task. Orfitelli and Polinsky (2016) argue at length about the dangers of grammaticality judgement tasks, but what about other types of testing: how well do we know their limitations? Assembling more data on different tasks is a necessary first step in addressing this question. At this juncture, the tasks that seem most appropriate include visual world paradigm tasks (Sekerina and Trueswell 2011), self-paced listening (Katsika and Allen 2013), picture-matching (Montrul 2015: 188–207), combined Likert-scale rating/response-time recording (Scontras et al. 2016; Scontras et al. submitted), and truth-value judgment tasks (Montrul and Ionin 2012; Scontras et al. 2016).

### 2.2 Existing evidence in support of representational differences

Several studies have amassed intriguing evidence that HSs and NSs indeed differ in their underlying representations. Scontras et al. (2016) tested possible differences between NSs and HSs of Central American Spanish with respect to agreement in number and gender. The experimental paradigm included stimuli such as (3), where the predicate of the small clause escrit- agrees with the small-clause subject el libro (head noun) in gender and number:

\[
\text{Juan consideraba el libro en la mesa}
\]

Juan considered the book on the table

exceintemente escrito.

excellently written.

‘Juan considered the book on the table well-written.’

When the agreeing predicate and the agreed-with DP are separated by a local noun inside a prepositional phrase, agreement attraction effects can occur. Such effects are observed in English cases such as (4), where the predicate incorrectly agrees with the local noun cabinets rather than the head noun key (Bock and Miller 1991; Bock et al. 2001; den Dikken 2011; Franck et al. 2006, a.o.).

(4) The key to the cabinets are lost.

---

4 Self-paced reading is much less appropriate for HSs, who usually have little or no experience with literacy in the home language.

5 See Contreras (1987); Jiménez-Fernández and Spyropoulos (2013) for the syntactic analysis of these small clauses.
In Scontras et al.’s study, participants were asked to listen to sentences such as (5) (in which the local noun either matched or did not match the head noun in gender and/or number) and to rate them on a Likert scale; their response times were also recorded.

(5) Juan consideraba el libro en las mesas
Juan considered the book on the tables
HEAD NOUN: excelente
LOCAL NOUN: escrito/*escritos/*escrita/*escritas.

NSs showed a clear grammaticality effect for both number and gender. They also showed attraction effects when the local noun was plural, much like in English.

NSs rated errors in number and gender as less acceptable than errors in just number (where presumably attraction masks the offense). Number violations alone and gender violations alone were rated higher than the violations in both features together, suggesting that NSs recognize each of these agreement features separately in their linguistic representation (Fuchs et al. 2015). In other words, they seem to entertain a model with two separate syntactic heads, number and gender, and one feature per head:

(6) \([\text{NumP}\ [\text{GenP}]]\) split representation

In contrast, HSs did not differentiate between errors in one feature and errors in two features when rating attraction conditions: errors in number alone were rated equally as (un)grammatical as errors in both number and gender. This finding cannot be attributed to loss of gender in the heritage grammar, because the same speakers rated feminine head-noun attraction conditions higher than masculine head-noun attraction conditions, indicating their awareness of and sensitivity to gender. Scontras et al. (2016) take this finding as evidence that Spanish HSs bundle number and gender features in their grammar, a departure from the native baseline of feature splitting.

What could motivate this change in the structure of Spanish agreement? Economy is one possibility: representational economy dictates a preference for structures with fewer structural heads. However, eliminating a head also leads to greater interpretive opacity for its associated features, which in turn leads to impoverished agreement—a property commonly observed in languages under contact in general (McMahon 1994: Ch. 4). The overall result is a snowball effect, where minimal adjustments, arguably driven by economy considerations, lead to a large-scale realignment of the internal grammar.

The commitment of Spanish HSs to representational economy is fascinating in light of the various online limitations of HSs, which are normally attributed to processing economy instead (Tsimpli et al. 2004). The next section explores the competition between representational and processing economy under ellipsis and presents yet another argument for representational differences between monolingual and heritage grammars.

2.3 Ellipsis

A pronoun can be connected to its antecedent in at least two ways (Heim 1982; Grodzinsky and Reinhart 1993; Reuland 2011; Koornneef et al. 2011): it can be represented as a variable, bound by an antecedent, or it can be associated with an argument in the preceding discourse through coreference. In simple cases, the two configurations yield the same results, but under ellipsis, differences can emerge. Consider the following ambiguous example (Koornneef 2008; Koornneef et al. 2011):
The acrobat likes his jokes and the clown does too.

The sloppy reading, in which the acrobat likes the acrobat’s jokes and the clown likes the clown’s jokes (9), represents the bound-variable strategy of pronoun resolution:

(9)  
  a. The acrobat, likes his jokes and the clown, likes his jokes too. SLOPPY
  b. The acrobat (λx (x likes x’s jokes)) & the clown (λx (x likes x’s jokes))

The strict reading, in which both the acrobat and the clown like the acrobat’s jokes, represents a simple coreference whereby the pronoun picks up the acrobat as its antecedent and this referent is carried over into the unpronounced VP:

(10)  
  a. The acrobat, likes his jokes and the clown, likes the acrobat,’s jokes too. STRICT
  b. The acrobat (λx (x likes a’s jokes)) & the clown (λx (x likes a’s jokes))

Numerous processing studies indicate that the bound-variable dependency (sloppy reading) is easier to process, both offline and online (see Koornneef et al. 2011 for an overview of the existing work). Assuming that heritage grammars favor processing economy, all factors being equal, HSs should eliminate the coreferential dependency altogether or at least show a stronger preference for the bound-variable dependency than NSs.

To test this prediction, a study was conducted comparing Russian HSs and NSs. Russian has a number of clause-level constituent ellipses, two of which are relevant here. Polarity ellipsis targets the complement of a polarity head Σ above T (11); verb-stranding verb-phrase ellipsis (VVPE) elides a vP-sized constituent and causes the verb to move to an Asp head just below T (12).6

(11) Maša ego videla, a ja net.  
    Masha._nom him._acc saw but I._nom not
    ... [... NegP ne [AspP videla [VP ego [VP t]]]

    ‘Masha saw him, but I didn’t.’

Pronouns in such elliptical constructions can be part of a bound-variable dependency or a coreferential dependency, yielding a sloppy or strict interpretation, respectively.

(13) VVPE

    Direktor už e pokazal mal’čikam ix mesta,
    principal._nom already showed boys._dat their seats._acc
    a zavuč devočkam poka ne pokazala ix-mesta.
    but vice-principal._nom girls._dat yet not showed
    (i) ‘The principal already showed the boys their seats, but the vice principal did not yet show the girls’ seats.’ SLOPPY
    (ii) ‘The principal already showed the boys their seats, but the vice principal did not yet show the girls the boys’ seats.’ STRICT

6 See Gribanova (2013) for a detailed syntactic analysis.
(14) Polarity Ellipsis

Direktor už pekažal mal’čikam ix mesta,
principal.NOM already showed boys.DAT their seats.ACC

a zavuč devočkam poka net pokazala ix mesta.
vice-principal.NOM girls.DAT yet not

But vice-principal.NOM girls.DAT yet not 
(i) ‘The principal already showed the boys their seats, but the vice principal did not yet show the girls their seats.’
SLOPPY
(ii) ‘The principal already showed the boys their seats, but the vice principal did not yet show the girls the boys’ seats.’
STRICT

We conducted an auditory experiment on these stimuli with two groups of subjects: Russian NSs in Moscow (N = 23, avg. age 42;2) and Russian HSs dominant in American English (N = 27, avg. age 22;3). On the assumption that bound-variable dependencies require less processing effort, we predicted that all speakers should prefer reading (i) over (ii); furthermore, for HSs, this reading should be either unique or strongly preferred.

The stimuli included sentences such as VVPE and Polarity Ellipsis, with ditransitive and benefactive verb frames. For the purposes of this article, we discuss only the stimuli containing coordinate structures, which were of two types:

(15) a. assertion in the antecedent – negation in the ellipsis site
b. negation in the antecedent – assertion in the ellipsis site

The stimuli manipulated the VVPE and polarity ellipsis contrasts, with eight lexical verbs in each condition. Stimuli were normed for naturalness and plausibility with five NSs, who also evaluated each sentence for strict/sloppy bias on a 5-point scale. Only those stimuli that were rated as unbiased (3 and up for each reading) were used. Fillers included coordinated sentences with two numerical phrases; see (17) below.

Stimuli were presented auditorily; subjects were asked to rate the plausibility of a single answer, formulated as an inference of the original critical sentence.

(16) a. SLOPPY INFERENCE
Zavuč dolžna pokazat’ mesta devoček.
vice-principal.NOM must show seats.ACC girls.GEN
‘The vice principal must show the girls’ seats.’

b. STRICT INFERENCE
Zavuč dolžna pokazat’ mesta mal’čikov.
vice-principal.NOM must show seats.ACC boys.GEN
‘The vice principal must show the boys’ seats.’

Note that standard experimental design would offer both inferences to participants after each scenario (cf. Koornneef et al. 2011); the current design broke with this configuration for two reasons. First, the auditory nature of the experiment (chosen to accommodate HSs’ generally low literacy levels) made it harder for participants to track the data. Second, HSs are known to do poorly when given too many options (Montrul 2015: 256–258; Polinsky 2016; Orfitelli and Polinsky 2016); we hoped to reduce confusion by limiting their

---

7 The age discrepancy between the two groups is intentional; the input received by HSs is the language spoken by their parents or even grandparents, not by their peers, so we deliberately compared younger HSs to NSs of the preceding generation.
8 Using ditransitive frames (where the indirect object serves as the antecedent of the possessive pronoun in direct-object position) allowed us to sidestep possible complications concerning the use of possessive pronouns in elliptical contexts (see Bailyn 2014 for a discussion).
options to a simple plausibility rating. In order to avoid the repetition of identical sentences, we kept the verb constant but changed the lexical items in subject and object position in each pair.

The fillers were followed by inferences evaluating the numbers stated in the antecedent. For example:

(17) a. **Antecedent sentence (filler)**

Na kuxne užě rabotalo šest’ p caravan, a v restorane ešče sidelo tol’ko piját’ posetitelej.

‘Six cooks were already working in the kitchen, but there were only five customers in the restaurant.’

b. **Inference, correct**

Povarov bylo bol’se, čem posetitelej.

‘There were more cooks than customers.’

c. **Inference, false**

Posetitelej bylo stol’ko že, skol’ko povarov.

‘There were as many cooks as visitors.’

Based on responses to the fillers, two HSs were removed from analysis. All other participants responded adequately to the stimuli and fillers. Responses were consistent across stimuli with assertions and negations in the antecedent; below, I discuss the results together.

In all the stimuli, the native controls (NSs) showed a clear preference for the sloppy interpretation, as confirmed by a planned paired-sample t-test (Table 1). These results enrich the empirical evidence supporting a preference for sloppy readings, originally reported based on English and Dutch ellipsis (cf. Koornneef et al. 2011). Although these results were obtained for a new language using a different methodology, they are nevertheless fully compatible with existing generalizations.

<table>
<thead>
<tr>
<th>Sloppy interpretation</th>
<th>Strict interpretation</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>WVPE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.89</td>
<td>2.75</td>
<td>0.027</td>
</tr>
<tr>
<td>Polarity ellipsis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.55</td>
<td>2.82</td>
<td>0.074</td>
</tr>
</tbody>
</table>

Note: 1: inference not plausible at all, 5: inference very plausible.

Table 1: NSs’ rating of inferences in Russian elliptical constructions.

Note that the NSs rated implausible inferences on the fillers much lower than strict inferences on the elliptical sentences (Likert rating 1.8; SD 1.15), indicating that the strict reading on ellipsis is not impossible, but dispreferred.

The picture from the HSs’ ratings is quite different (Table 2):

<table>
<thead>
<tr>
<th>Sloppy interpretation</th>
<th>Strict interpretation</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>WVPE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.08</td>
<td>4.32</td>
<td>0.08</td>
</tr>
<tr>
<td>Polarity ellipsis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.21</td>
<td>2.34</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Note: 1: inference not plausible at all, 5: inference very plausible.

Table 2: HSs’ rating of inferences in Russian elliptical constructions.
The HSs rated implausible fillers at 2.15, in line with the general HL reluctance to produce low ratings on acceptability tasks even for ungrammatical items (the so-called ‘yes-bias’). Crucially, the HSs rated the inferences under ellipsis equally implausible as the (genuinely implausible) filler ratings. In other words, HS-assigned scores in the 2 area of a 1–5 scale should be considered maximally low: a rejection of a particular reading.

Heritage Russian speakers were also tested on English ellipsis, with the object condition and coordination only, in order to assess their knowledge of ellipsis in their dominant language. All the stimuli had third-person subjects, as in (18). There were 18 sentences in total, each followed by the sloppy or strict inference interpretation. Again, the fillers consisted of sentences with numerical expressions.

(18) The Petersons like their neighbors and the Woolards do too.
   i. The Woolards like their neighbors. SLOPPY INFERENCE
   ii. The Woolards like the Petersons’ neighbors. STRICT INFERENCE

The results from the English test support the generalization that sloppy inferences are preferred under ellipsis. HSs’ ratings on the less-preferred strict interpretation were significantly higher than their ratings on the implausible filler inferences (Wilcoxon two-tailed test, \(p = 0.032\)), cf. Table 3.

Table 3: Rating of inferences in English elliptical constructions by HSs.

<table>
<thead>
<tr>
<th></th>
<th>Sloppy interpretation</th>
<th>Strict interpretation</th>
<th>Implausible inference on fillers</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>3.85</td>
<td>2.44</td>
<td>1.80</td>
</tr>
</tbody>
</table>

Note: 1: inference not plausible at all, 5: inference very plausible.

It seems from these results that HSs allow only one type of inference under (each type of) ellipsis: the strict interpretation under VVPE, (13), and the sloppy one under polarity ellipsis, (14). The end result is an elimination of ambiguity (a common property of HLs generally; see Kondo-Brown 2005; Benmamoun et al. 2013a; Scontras et al. 2015) and a striking representational difference between the HL and the baseline. The behavior of heritage Russian is particularly intriguing because both the Russian and English baselines favor the sloppy reading in VVPE – precisely where heritage Russian eliminates it.

A lack of sloppy readings under VVPE has also been observed in child language. The explanation offered in that case appeals to early L1-learners’ general lack of variable-binding (Ruigendijk et al. 2011; Baauw et al. 2011). Clearly, such an explanation cannot be evoked for heritage Russian: adult HSs unequivocally have the ability to construct variable-binding dependencies, under polarity ellipsis, and in their dominant language. Interference from the dominant language also lacks explanatory power, given that the heritage Russian pattern is the exact opposite of what we find in English.

What, then, leads to this unexpected change in the representation of ellipsis in heritage Russian grammar? Why does the elimination of ambiguity proceed by associating VVPE with the strict reading and polarity ellipsis with the sloppy interpretation?

I contend that this elimination of ambiguity arises incidentally, as part of an independently-motivated structural reorganization of heritage Russian syntax. Several factors are at play. The first is the restructuring of aspect in heritage Russian. The distinction between perfective and imperfective verbs is clear and central to the syntax of baseline Russian; in particular, the verb in Russian moves to the aspectual head rather than T (Bailyn 2012; Kallestinova 2007; Gribanova 2013). However, aspectual specification is either obscured or lost entirely in heritage Russian (Polinsky 2006, 2008; Laleko 2010, 2011; Mikhailova 2012). As a result, AspP is lost in heritage Russian. Researchers disagree over whether the heritage Russian aspectual head

---

9 22 Russian HSs were tested on English. Of these, six were tested only in English, and 16 were part of the cohort tested in Russian as well. Among the 16, ten subjects were first tested in English and returned a week later for the Russian test; six subjects were tested in Russian first.
disappears completely or, as in the case of Spanish gender, become bundled with some other head; at this point, I remain agnostic on this issue. Either way, without an independent AspP, V-to-Asp movement becomes unavailable. We now see again the “snowball effect” I mentioned in the discussion of Spanish above: once verb movement is lost, verb-stranding – an indispensable component of Russian VVPE – is no longer available. Accordingly, the VVPE structure is reanalyzed as argument deletion:

(19) a. baseline VVPE 
    b. heritage N(P)-ellipsis

It is still, in principle, possible to interpret the missing DP object as a pronoun bound by an antecedent, as in (9). Now, however, another property of HLs in general and heritage Russian in particular comes into play: the restricted range of null pronouns. The loss or degradation of null arguments is one of the most commonly observed properties across HLs (Sorace and Serratrice 2009; Sorace et al. 2009; Ivanova-Sullivan 2014; Montrul 2008, 2015; Keating et al. 2011; Keating et al. 2016; Tsimpi 2004, a.o). Since all the stimuli in the current experiment involved object ellipsis, the absence of null pronominal objects in heritage Russian eliminates the possibility of a bound variable interpretation in a. baseline VVPE b. heritage N(P)-ellipsis. Instead, argument drop or N-deletion emerges as a last resort, yielding the strict reading that is so prominent in the pattern of results.

Expletive subjects and clausal anaphors are among the few null elements typically preserved in HLs and L2 (Ivanova-Sullivan 2014; Keating et al. 2016 for HL; Rothman and Iverson 2007 for L2). The null clausal anaphor is still available and is found under polarity ellipsis; that’s the structure where only the sloppy reading is possible. In principle the strict reading could be expected here as well, but since the strict reading is reserved for VVPE, it is possible that paradigmatic considerations make such a reading less salient or impossible under polarity ellipsis.

(20) Polarity ellipsis

Altogether, the changes observed in the grammar of heritage Russian ellipsis cannot reasonably be reduced to performance considerations or transfer. Instead, they follow from a conspiracy of changes independently motivated for heritage Russian, (21a), and other HLs, (21b):

(21) a. loss of aspeclual specifications 
    b. loss of null definite pronouns

\[10\] Indeed, cross-linguistically, the presence of null definite/referential subjects entails the presence of null expletives, but not vice versa (Nicolis 2005; Camacho 2013).
Once we recognize the effect of these changes, we can see that the fate of heritage Russian elliptical structures is not random but arises through a systematic reorganization of grammar. The observed lack of ambiguity under ellipsis is just a side effect.

The Russian data explored in this section show even more clearly than the Spanish agreement data that not all variation in HSts’ linguistic performance can be attributed to processing economy. Sometimes, the internal heritage grammar may undergo a profound reorganization unrelated to pressures from the baseline language, the dominant language, or processing economy considerations.

3 Conclusions

This paper has provided a general overview of heritage language and offered nascent evidence in support of representational differences between baseline native and heritage grammars – differences that cannot be reduced to the effects of online processing constraints or memory limitations. Data from heritage Spanish number/gender agreement and from Russian ellipsis indicate that heritage grammars may differ fundamentally from their baseline counterparts in the organization of certain categories. In particular, heritage grammars tend to eliminate structural heads; this process yields a more economic structural representation, but may also lead to greater interpretive opacity in the remaining structural heads. Like in a Jenga game, the grammatical blocks removed from the heritage grammar must be extracted with care... otherwise, the entire tower may fall.

As a final caveat, I wish to stress that the existence of structural differences between baseline and heritage grammars does not deny or diminish the (well-documented and significant) effect of performance pressures on HL speech. However, as we learn more about Hls, we must recognize that online limitations cannot account for all the differences observed.

Acknowledgment: I am grateful to David Adger, Vera Gribanova, Jason Merchant, Eric Potsdam, Greg Scontras, and an anonymous reviewer for helpful comments. All errors are my own.

Note: The title of this paper echoes work presented in Adger (2015), and our conclusions are in fact similar. This work was supported in part by NSF (grants BCS-1144223 and BCS-1619857), the Davis Center and the Instituto Cervantes Observatory at Harvard University. Abbreviations follow the Leipzig glossing rules.

References


Lee- Ellis, S.-Y. 2012. Looking into bilingualism through the heritage speaker’s mind. College Park, MD: University of Maryland, PhD dissertation.


Otheguy, R. 2016. The linguistic competence of second-generation bilinguals: A critique of “incomplete acquisition”.


