Interpreting (un)interpretability

George Walkden (University of Konstanz) & Anne Breitbarth (Ghent University)

NB: *This is a pre-print version, not the publisher's PDF (for which see https://doi.org/10.1515/tl-2019-0022). This article appeared in* Theoretical Linguistics *45(3–4), 309–317; when citing, please use the page numbers given there. The publisher should be contacted for permission to re-use or reprint the material in any form.*

Keywords: sociolinguistic typology, syntactic change, L2 acquisition, simplification, Interpretability Hypothesis

1. Introduction

Our target article (henceforth W&B) proposed a diachronic connection between a structural property of grammars and particular sociohistorical situations: all else being equal, we predict that in sociohistorical situations in which adult L2 learners are particularly dominant quantitatively or qualitatively, uninterpretable features will typically be lost. W&B outlines a research programme rather than a full-fledged knockdown argument, and we thank BIBERAUER, VAN GELDEREN and YANOVICH for reading it in the spirit it was intended and raising important issues in their commentaries. For space reasons we cannot respond to all of these, of course. In section 2 we address questions of innateness; in 3 the characterization of (un)interpretability; in 4 specific issues to do with our analysis of negation; and in 5 the contact scenarios involved. Section 6 concludes.

2. What's the (fundamental) difference?

BIBERAUER takes issue with W&B's adoption of the Fundamental Difference Hypothesis (FDH), which she understands as the idea that Universal Grammar (UG) is not available to post-critical-period learners. Pointing to Bley-Vroman's (2009) revisiting of the hypothesis, she observes that its content is less clear now than it was in 1989. A minimal or empty UG – in particular one which does not prespecify a set of features is not taken to be a primitive – has the logical consequence that the difference between UG availability and unavailability is itself minimal or even non-existent.

Bley-Vroman (2009: 176) states that the FDH can be understood narrowly or broadly, and it is the broad sense that was intended in W&B (though we did not make this explicit): unlike child language learning, post-critical-period learning is not reliable and not convergent. This difference may, as in much work on second language acquisition from a generative perspective following Bley-Vroman (1989), be explained by limited or partial access to UG – but crucially it need not. In theory, any of Chomsky's (2005) three factors (innate endowment, experience, or language-independent principles)

could give rise to the indisputable discrepancy between first and second language acquisition.

Biberauer takes W&B to assume a rich, innately given set of features, in contrast to more radically minimalist proposals. In fact, we did not (and do not) take a stance on this issue, as it is not required for our hypothesis. The exact aetiology of our 'substantial inventory of language-specific features' – whether they are acquired or innate – is not important for us: all that matters is that, once cognitively present, they are either interpretable or uninterpretable (on which see section 3). Thus, our proposal is fully compatible with not only the traditional UG-based view but also the emergentist view that Biberauer (2017), among others, has influentially espoused.

We ought to say more than this about second language acquisition of features, of course, and about how it plays out at the population level (see also Parodi 2016). YANOVICH outlines two conceivable scenarios. In the first, the L2 learner is simply blind to the effects of uninterpretable features, and is incapable of producing clauses that reflect their use. Yanovich considers this implausible, as do we. Yanovich's alternative is that such features are difficult to acquire, but not impossible. One variant on this that we would like to explore is that, in the absence of uninterpretable features, learners draw on general strategies of non-linguistic cognition and pattern matching in order to 'patch' their grammar (the term is from Morgan 1972, applied to L2 learning by Bley-Vroman 2009). 'Patches' are characterized by uncertainty, and – we might hypothesize – incur a higher processing and production cost than the fully internalized operations of the L1; see for instance Grüter et al. (2012), who show that even extremely proficient L2 learners of Spanish (who perform at ceiling in offline comprehension tasks) display errors in elicited production of grammatical gender agreement. This cost affects the output of L2 users, and thus also the input to future acquirers. We consider this scenario to be a plausible mechanism for the diachronic consequences of the unavailability of uninterpretable features in late L2 acquirers, though it is only one among several possibilities, and the exact nature of grammatical 'patches' remains vague at present.

3. Interpreting interpretability

The difference between interpretable and uninterpretable features is central to W&B, and therefore needs to be watertight. YANOVICH observes that uninterpretable features, as purely formal devices without interface implications, are intuitively a good fit for syntactic complexity. BIBERAUER, however, is sceptical, describing feature interpretability as 'ill-defined' and 'shifty'. We disagree. Historically, at least, the interpretable/uninterpretable distinction is one of the few notions that has survived since Chomsky (1995) into current theorizing, and is almost universally adopted in those current Minimalist analyses that are at all explicit about the features involved. In fact, the distinction can be seen as virtually conceptually necessary: given the Y-model of the architecture of the grammar, any feature present in syntax must either have an interpretation at the LF interface or not (cf. Chomsky 1995: 276–279). As Biberauer

notes, Zeijlstra (2014) rejects the idea that interpretable features are interpretable at LF, instead treating them as purely a subtype of formal feature. This is not the place for a full response to Zeijlstra; however, we note that he only argues that his proposal is a possible (and perhaps superior) alternative to the mainstream view, not that the latter is untenable (see also section 4 on negation below).

Conceptually, then, the interpretable-uninterpretable distinction is not problematic. Biberauer is of course right that empirically it is not always straightforward to establish what uninterpretable features are involved in a given construction or language, and where. A substantial part of pursuing the intuition behind W&B will be to nail these analyses down more precisely. But this is the bread-and-butter of syntactic analysis: the usual combination of empirical evidence and reasoned argumentation can be brought to bear.

A case in point is the analysis of definite articles. VAN GELDEREN tests the predictions of W&B in this domain using typological databases such as eWAVE and APiCS. Her starting point is the idea that definite articles in D bear an uninterpretable nominal feature that needs to enter into an Agree relation with an interpretable counterpart on a noun (cf. van Gelderen 2007, 2011). If this is correct, then definite articles ought to be vulnerable in contact situations involving extensive L2 acquisition. Assuming the languages in eWAVE and APiCS instantiate this kind of scenario (on which see section 5 below), we see the opposite of what we expect: articles are in fact *more* common in these languages than they are in the world's languages overall (as represented by WALS).

It is reasonable to ask at this point what the evidence is that definite articles are bearers of uninterpretable features. Van Gelderen adduces the fact that articles cannot occur without a corresponding noun (e.g. in English **I saw the*), suggesting that the article c-selects a nominal constituent. But this on its own is not compelling: an alternative explanation is simply that definite articles are the spellout of a functional head in the nominal extended projection that does not license ellipsis of its complement. In order to be sure that an uninterpretable feature is present, we'd need sure-fire evidence of an Agree relation. At least for languages like modern English, in which the form of the definite article is completely invariant, there is no such evidence. Preminger (2019) makes a powerful case (based on Person-Case Constraint effects) that "abstract" agreement, with no morphophonological footprint anywhere in the paradigm, does not generally exist; this position is appealing on learnability grounds, since – regardless of whether the features in question are learned or innate – it is unclear how an acquirer could associate them correctly with the right heads. If we take this seriously, then absence of evidence is evidence of absence, in this case: definite articles in languages like modern English do not bear [*u*N] features.[1]

W&B's prediction is then not that definite articles should be vulnerable in extensive L2-acquisition contact scenarios, but rather that definite articles *which Agree* (as reflected

[1] Moreover, definite articles in such languages (or rather the lexical items they spell out) seem to bear at least one *interpretable* feature, namely definiteness itself.

overtly, for instance, in number or gender) should be vulnerable in such scenarios. Research on SLA in Dutch for instance suggests that learners tend to overgeneralize the common gender determiner *de*, leading to the loss of the common/neuter distinction in the determiner system (e.g. Cornips 2008, Orgassa & Weerman 2008, Unsworth 2008, and references cited in those publications). Assuming that common gender in Dutch is absence of gender agreement, based on the fact that (a) it is the default option (80% of Dutch nouns are common gender), and (b) the common determiner *de* is also used in the plural, where gender is neutralized (Orgassa & Weerman 2008: 338), this would seem to confirm W&B's prediction. Unfortunately, this is not a feature tracked by APiCS or eWAVE. Eyeballing of the examples of definite articles provided in these two sources strongly suggests that (where present) they are morphologically invariant, though more research is needed on this point.

4. Issues in the analysis of negation

The main case study in W&B involves the shift from bipartite to single negation. VAN GELDEREN adduces evidence to suggest that we are on the right track in considering negative agreement to be L2-complex: only six of the 76 varieties in APiCS have bipartite negative markers, and at least one of these – Afrikaans – can be set aside (see footnote 3).

BIBERAUER, however, contests W&B's hypothesis that uninterpretable features are L2-complex in particular in connection to this case study. Her criticism consists of three main points:

(i) As a third factor bias, van Gelderen's Feature Economy should also be active in L2 acquirers, creating a problem for them, as they, like L1 speakers, would seek to reduce the use of interpretable features, at the same time as – under the IH – having no access to uninterpretable ones.
(ii) The original negative marker in a language undergoing JC bearing a [*u*NEG] feature is not shared by Zeijlstra (2004, 2008), as uninterpretable features require some form of doubling, like agreement or syntactic movement.
(iii) Under standard Minimalist assumptions, uninterpretable and interpretable features need to occur in tandem to establish an Agree relation, and cannot occur on their own, as W&B's representation of stage I of JC seems to suggest.

We return to the criticism regarding the universality of Feature Economy in section 5 below. But in her first criticism, BIBERAUER also suggests that assuming the crucial property of the original negative marker in a language undergoing Jespersen's Cycle to be that it bears a [*u*NEG] feature, as van Gelderen as well as W&B do, coupled with Feature Economy, would imply assuming that all instances of Jespersen's Cycle should be seen as driven by a compulsion to reducing L2-complexity. As long as it is

not clear that *all* instances of Jespersen's Cycle are contact-induced (which there is reason to doubt, cf. Breitbarth, Lucas & Willis forthc.: ch. 4), this is not a necessary conclusion. Besides, W&B only discuss the transition from stage II to stage III in their case study.

Regarding the second and third criticisms, we would like to point out that we mainly owe to Zeijlstra's work the general acceptance of null [*i*NEG] operators as a way to analyse languages with negative heads and pre-verbal NPI subjects. That is, Zeijlstra himself does not object to there being stage I languages with [*u*NEG] negative heads, where Agree relations are only vacuously expressed (see e.g. Zeijlstra 2004: 249–250 on the stage I language Czech; though, again, stage I was not the focus of W&B's case study). While such a null [*i*NEG] operator is not explicitly included in the crude representation in (5a), it would under standard Minimalist assumptions of course need to be there.

In particular with an eye on the third criticism, the dependency between interpretable and uninterpretable features, under standard Minimalist assumptions (not Zeijlstra's) is asymmetric. While uninterpretable features do need an interpretable one to license them, interpretable ones do not require uninterpretable features to be present, under the assumption that they are indeed interpretable at LF.[2] As W&B state, however, this is not the main point; the main point is that the original negator bears a [*u*NEG] feature.

5. Contact scenarios

W&B focuses on contact-induced simplification, understood as the loss of uninterpretable features. Importantly, however, this is not the only outcome of language contact that Trudgill (2011) envisages. Situations of long-term language contact, involving child multilingualism, are hypothesized by Trudgill to give rise not to simplification but to additive complexification, via transfer (copying) of features and structures from one language to another.[3] Cases he discusses under this rubric include the Balkan Sprachbund and the Amazonian linguistic area (Dixon & Aikhenvald 1999). In our terms, syntactic complexification involves the innovation of new (interpretable or uninterpretable) features. A plausible case is the development of bipartite negation (*ur* … *ša* replacing single *ur*) in northern varieties of Berber,

[2] If, as Biberauer attribues to Zeijlstra and Wallage, negators at stage I of JC are in fact [*i*NEG], one might expect there to be a covert [*u*NEG] element. That interpretable features (in the standard Minimalist sense) do not in fact need uninterpretable features to build representations interpretable at LF is at the heart of Haegeman & Lohndal's (2010) approach to Negative Concord, for instance. Further detail would lead too far afield in the present context.

[3] Adding weight to this distinction, BIBERAUER (note 13) points out – tantalizingly – that the modern Afrikaans varieties in which bipartite negation is robust are best characterized as developing in situations of long-term co-territorial L1-based contact.

which Lucas (2012: 287–290) convincingly argues is a case of syntactic borrowing from Arabic.

VAN GELDEREN refers to 'high-contact' varieties in her commentary. Although all the varieties in APiCS can be said to be high-contact, and 22 of the 76 varieties in eWAVE are explicitly described this way, it is less clear whether all of these varieties emerged via the sociohistorical scenario that Trudgill (2011) predicts to lead to simplification. Irish and Welsh English, for instance, are included in eWAVE as high-contact varieties, but the contact situation (with Celtic) seems to us to be more of the long-term type likely to lead to complexification: language shift is ongoing, but has been ongoing since at least the twelfth century in both cases. Pidgins and creoles too can develop rapidly or less rapidly (see e.g. Baker 1996). It will be of the utmost importance for us to distinguish between the two types of scenario in order to assess the hypothesis in W&B; demographic evidence will be crucial in doing so.[4]

The type of learner involved matters, then. This also helps to stave off a further criticism. BIBERAUER states that researchers have 'adopted diametrically opposed positions on the relative complexity of interpretable and uninterpretable features', contrasting W&B's approach with that of van Gelderen (2011), who argues that interpretable features are computationally costly ('Feature Economy'). The diametric opposition is only apparent, though: van Gelderen argues (2011: 22–26) that her Feature Economy is at play in child language acquisition, whereas the Interpretability Hypothesis as adopted by W&B is intended to characterize adult L2 acquisition. It's possible that both are right. In discussing complexity it's always useful to keep in mind the question: complex for whom?

5. Outlook

W&B argue, following the Interpretability Hypothesis (e.g. Tsimpli & Dimitrakopoulou 2007), that uninterpretable features are a particular sticking point for adult L2 acquirers in syntax, and that in certain sociohistorical scenarios where L2 acquisition plays a central role (Trudgill 2011) this will be reflected in the historical record. In doing so, we argue for a rapprochement and dialogue between subfields that have grown distant from one another: syntactic theory, L2 acquisition studies, sociolinguistics, historical corpus linguistics.

Many things remain to be done, and the commentators raise important questions. For instance, BIBERAUER outlines not one but two alternative ways of thinking about complexity: Kolmogorov complexity as applied to the description length of grammars, and prosodic weakness as L2-difficulty. VAN GELDEREN flags up new empirical arenas

[4] It is worth noting that quantitative data from APiCS and eWAVE should be used with caution for another reason: English is the lexifier for just over a third (26 of 76) of the languages in APiCS, and is in the mix for all 76 of the varieties in eWAVE. The dominance of English here means that we should not treat the varieties as separate, historically independent data points.

in which to put our proposal to the test. And YANOVICH highlights, rightly, the problem of equifinality: for several of the examples that we mention, alternative explanations are available. As he observes, this is not a problem that is unique to our approach: it is endemic to historical linguistics and indeed all of historical science. All we can do is to construct a historical narrative (in the sense of Mayr 2004: 32–33) and then confront it with as many relevant facts as we can muster.

Time will tell if our proposal about the nature of complexity as L2-difficulty and its diachronic consequences is borne out. We will be pursuing these in the ERC-funded project "Sociolinguistic typology and responsive features in syntactic history" (STARFISH), and are keen to discover more.

References

Baker, Philip. 1996. Some developmental inferences from historical studies of pidgins and creoles. In Jacques Arends (ed.), *The early stages of creolization*, 1–24. Amsterdam: John Benjamins.

Biberauer, Theresa. 2017. Factors 2 and 3: a principled approach. *Cambridge Occasional Papers in Linguistics* 10: 38–65.

Bley-Vroman, Robert. 1989. What is the logical problem of foreign language learning? In Susan M. Gass & Albert Schachter (eds.), *Linguistic perspectives on second language acquisition*, 141–168. Cambridge: Cambridge University Press.

Bley-Vroman, Robert. 2009. The evolving context of the Fundamental Difference Hypothesis. *Studies in Second Language Acquisition* 31: 175–198.

Breitbarth, Anne, Christopher Lucas, & David Willis. Forthcoming. *The development of negation in the languages of Europe and the Mediterranean*. Vol. 2: *Patterns and processes*. Oxford: Oxford University Press. [to appear March 2020]

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam. 2005. Three factors in language design. *Linguistic Inquiry* 36: 1–22.

Cornips, Leonie. 2008. Losing grammatical gender in Dutch: The result of bilingual acquisition and/or an act of identity? *International Journal of Bilingualism* 12/1&2: 105–124

Dixon, R. M. W., & Alexandra Y. Aikhenvald. 1999. *The Amazonian languages*. Cambridge: Cambridge University Press.

Grüter, Theres, Casey Lew-Williams & Anne Fernald. 2012. Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research* 28 (2): 191–215.

Haegeman, Liliane & Terje Lohndal. 2010. Negative concord and (multiple) Agree: A case study of West Flemish. *Linguistic Inquiry* 41: 181–211.

Lucas, Christopher. 2012. Contact-induced grammatical change: towards an explicit account. *Diachronica* 29 (3): 275–300.

Mayr, Ernst. 2004. *What makes biology unique? Considerations on the autonomy of a scientific discipline*. Cambridge: Cambridge University Press.

Morgan, J. L. 1972. Verb agreement as a rule of English. *Chicago Linguistics Society* 8, 278–286.

Orgassa, Antje & Fred Weerman. 2008. Dutch gender in specific language impairment and second language acquisition. *Second Language Research* 24/3: 333–364.

Parodi, Teresa. 2016. Formal features and vulnerable domains in L2 acquisition and an outlook on language contact. In Susann Fischer & Mario Navarro (eds.), *Proceedings of the VII Nereus International Workshop: "Clitic Doubling and other issues of the syntax/semantic interface in Romance DPs"* (Arbeitspapier 128), 91–103. Konstanz: Department of Linguistics.

Preminger, Omer. 2019. What the PCC tells us about "abstract" agreement, head movement, and locality. *Glossa* 4 (1): 13.

Trudgill, Peter. 2011. *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford: Oxford University Press.

Tsimpli, Ianthi Maria, & Maria Dimitrakopoulou. 2007. The Interpretability Hypothesis: evidence from *wh*-interrogatives in second language acquisition. *Second Language Research* 23, 215–242.

Unsworth, Sharon. 2008. Age and input in the acquisition of grammatical gender in Dutch. *Second Language Research* 24/3: 365–395.

van Gelderen, Elly. 2007. The definiteness cycle in Germanic. *Journal of Germanic Linguistics* 19: 275–308.

van Gelderen, Elly. 2011. *The linguistic cycle: language change and the language faculty*. Oxford: Oxford University Press.

Zeijlstra, Hedde. 2004. Sentential negation and negative concord. PhD dissertation, Universiteit van Amsterdam.

Zeijlstra, Hedde. 2014. On the uninterpretability of interpretable features. In Peter Kosta, Steven L. Franks, Teodora Radeva-Bork & Lilia Schürcks (eds.), *Minimalism and beyond: radicalizing the interfaces*, 109–129. Amsterdam: John Benjamins.