The Combinatorial Variability Hypothesis in the Second Language

Roumyana Slabakova and Johnathan Gajdos
University of Iowa

1. Introduction

Variability in the production and comprehension of inflectional morphology is widely attested and well documented in L2 acquisition. The sources of this variability, are, however, disputed. Several influential proposals have been advanced in recent years to explain morphological variability. We will discuss some of them in turn.

The Missing Surface Inflection Hypothesis (Prévost and White, 2000) argues that variable production of the morphology may be due to learners' imperfect mapping of specific morphological forms to abstract categories. In such cases (which could be due to processing difficulties and more often attested under communication pressures) learners resort to defaults, forms that are underspecified in some features. The authors contend that L2 learners have acquired the relevant features of the terminal nodes in the syntax (from the L1, or on the basis of L2 input constrained by UG).

The Prosodic Transfer Hypothesis (Goad & White, 2004, 2006, Goad, White and Steele, 2003), on the other hand, shows that the production of functional morphemes is constrained by L1 prosodic representations. The L2 learners, according to this hypothesis, have difficulty in accurately producing inflectional morphology whenever the L2 requires a prosodic representation unavailable from their L1. In dealing with the difficulty, learners sometimes resort to L1 prosodic structures; and sometimes they adapt the native prosodic structures in UG sanctioned ways

Thirdly, the Feature Assembly Hypothesis (Lardiere, 2007, 2008) draws attention to the fact that the ways in which grammatical features are morphologically combined and conditioned present formidable learning problems in L2 acquisition. For example, primitive interpretable features are often clustered differently in different languages and there is a kind of morphological competence that must be acquired by the learner. This morphological competence involves figuring out how to reconfigure features into new or different formal configurations or remap native features onto new functional morphology. The prediction of this hypothesis is that the more re-assembly of features the L2 learner must do, the more difficulty she will face, and such morphemes (and meanings) will take longer to acquire.

A brand-new proposal, the Contextual Complexity Hypothesis (Hawkins and Casillas, 2007) argues that while native speakers have vocabulary entries specified in bundles of features at the point of lexical insertion (the terminal nodes), non-native speakers at early stages of development have vocabulary entries specified for the terminal nodes that co-occur with the exponent of that item. For example, to represent the *–s* of 3rd person singular agreement morphology, native speakers associate the form /s/ with any terminal node having features [V, –past, +singular, 3p], while learners have something like: "insert /s/ in the context of a verb which is in the context of a non-past T, itself in the context of a 3rd person singular N." (p. 8) The more sister terminal nodes specified in a morpheme's environment, the more costly its lexical access will be. Underlying this proposal is the assumption that L2 learners do not always have access to uninterpretable features; specifically, when these same uninterpretable features do not come from their native language.

© 2008 Roumyana Slabakova and Johnathan Gajdos. Selected Proceedings of the 2007 Second Language Research Forum, ed. Melissa Bowles et al., 35-43. Somerville, MA: Cascadilla Proceedings Project.

Finally, the influential emergentist explanation for morphological variability has it that properties of the input encountered by L2 learners play a decisive role in the type of representations they establish. Associative (contingency) learning is mainly based on type and token frequency of forms in the input (Bybee, 2001; Ellis, 2002, 2006a,b). Order and accuracy of acquisition of inflectional morphology is determined by how robust the particular construction is in the ambient input. High frequency forms (including irregular forms) have lexically stronger representations than low frequency forms. In addition, other factors that can determine order and accuracy of inflectional morphology acquisition are: perceptual salience, semantic complexity, morpho-phonological regularity, and syntactic category (Goldschneider and DeKeyser, 2001).

In this paper, we look at the potential of another hypothesis to explain variable performance in this respect, Adger's (2006) Combinatorial Variability Hypothesis (CVH). The CVH explains intra-personal morphosyntactic variation as arising from the combinatorial mechanisms of language itself. It offers an evaluation metric of uninterpretable feature combinations, predicting which forms are going to surface more, and may even become defaults, in the acquisition of inflection.

2. The Combinatorial Variability Hypothesis

2.1. Lexical items and features

Lexical items are built up of combinations of features in unstructured sets. (Bivalent features are assumed by this theory.) How does the learner decide on the features? The learner has access to the Conceptual Structure of human thought, which provides a range of possible features that have to be semantically motivated: e.g., number, participant in the discourse, etc. in pronouns. The particular forms of the lexical items in the input tell the learner which available contrasts are actually marked in her (L1 or L2) language.

2.2. Pronouns

We shall exemplify the use of features with the necessary features in the make-up of personal pronouns, see matrix in (1) below. The feature [± singular] marks the number of the pronoun. The feature [± participant] marks whether the pronoun refers to a participant in the speech act. *I*, *we*, and *you* are [+ participant]. The feature [± author] distinguishes between addressee and speaker. *I* and *we* are [+author]. The Feature Co-occurrence Restriction: A lexical item is specified for [± author] if and only if it is specified for [+ participant]. Thus, third person pronouns do not need to specify [± author].

(1)

[singular: + participant: + author: +]	I	[singular: – participant: + author: +]	we
[singular: + participant: + author: –]	you	[singular: – participant: + author: –]	you
[singular: + participant: –]	he/she/it	[singular: – participant: –]	they

However, this is not all. Certain lexical items carry features that are purely formal in nature. Their job is to establish syntactic dependencies. They are called uninterpretable (Chomsky, 2001) and are marked *u*F. In a sentence like *He is tired,* the person features [±singular], [±participant], [±author] are interpretable on the pronoun. They value and then check the uninterpretable features on the *verb* in an agreement chain (Pesetsky & Torrego, 2001). Thus, we have interpretable person features on the pronoun and their uninterpretable equivalent on the agreeing verb.

2.3. What is an agreement chain?

A pair of lexical items where the uninterpretable features of one item **can be only a subset** of the interpretable features of the other. Full Interpretation requires every uninterpretable feature to be in an agreement chain. (2) presents an agreement chain, where "…" stands for c-command.

(2)

singular: + participant: + author: +	…	*u*singular: + *u*participant: + *u*author: +

In (3) below, we give a schematic overview of the theory. LI stands for lexical item.

(3)

LI_1 {F_1, F_2, F_3} … LI_2 {uF_1} → PF (LI_2) = x

LI_3 {uF_2} → PF (LI_3) = y

LI_4 {uF_3} → PF (LI_4) = z

An LI_1 with three features as exemplified above, is able to combine with a range of other items bearing different subsets of uninterpretable features. It is important to notice that the array of the interpretable features is the same in all three representations, even though their phonological forms are different. "What allows the variability is the possibility that particular lexical items may be underspecified for the uninterpretable agreement features that they contain. This underspecification is irrelevant to the semantic system, since these features are not interpreted." (Adger, 2006: 510).

This system also predicts frequencies: if there are two ways for the grammar to output a specific PF, as is the case for x in the example below, but only one way to output a y, then x will appear more often than y, roughly in proportions 66% vs. 33%.

(4)

LI_1 {F_1, F_2, F_3} … LI_2 {uF_1} → PF (LI_2) = x

LI_3 {uF_2} → PF (LI_3) = x

LI_4 {uF_3} → PF (LI_4) = y

2.4. How does the variability arise: a concrete example

(5)

singular: + participant: + author: +	am	singular: – participant: + author: +	are
singular: + participant: + author: –	are	singular: – participant: + author: –	are
singular: + participant: –	is	singular: – participant: –	are

The matrix in (5), showing the uninterpretable features that can be specified on the verbal form, offers fully specified lexical items. However, as the reader can see, there is a large amount of homonymy and hence, generalizations are missed. For example, all the [author: –] forms get spelled as *are*, and so do all the [singular: –] forms. Applying underspecification of the feature content of the verb, we get:

(6) [*u*author: –] are
[*u*singular: –] are
[*u*singular: +, *u*author: +] am
[*u*singular: +, *u*participant: –] is

The algorithm of generalization is based on the premise that lexical items with fewest features are best. The system also uses some additional filters (Adger, 2006: 518). The following simultaneous constraints work in the lexicon:

A) Reject Optionality (an LI is kept if there is always a matching form)
B) Reject Synonymy (an LI is kept if it does not create synonyms)
C) Minimize the size of the Lexicon

The reader is referred to the original publication for details on how these are implemented.

3. German copula *sein* 'be' in the present tense

Using this evaluation metric and various filters the system uses, we analyzed the uninterpretable features of the German verb *sein* 'be' in the present tense.

(7)	1. sg.	Ich bin	1. pl. wir sind
	2. sg.	du bist	2. pl. ihr seid
	3. sg.	er/sie ist	3. pl. sie sind

To come up with the lexical specifications of features that are maximally generalized, we go through the unary feature specifications, e.g. [*u*singular: +] and find that none of them is sufficient to describe exactly one lexical item. Then we go to all the binary specifications:

(8) a. [*u*singular: +, *u*author: +] bin
b. [*u*singular: +, *u*author: –] bist
c. [*u*singular: +, *u*participant: +] bist, bin REJECT
d. [*u*singular: +, *u*participant: –] ist
e. [*u*singular: –, *u*author: +] sind
f. [*u*singular: –, *u*author: –] seid
g. [*u*singular: –, *u*participant: +] sind, seid REJECT
h. [*u*singular: –, *u*participant: –] sind
i. [*u*participant: +, *u*author: –] seid, bist REJECT
j. [*u*participant: +, *u*author: +] sind, bin REJECT

One form, namely *sind*, is output by two underspecified combinations of features. The rest of the forms are uniquely identified.

(9) [*u*singular: –, *u*author: +] → PF *sind*
[*u*singular: –, *u*participant: –] → PF *sind*

4. Prediction for accuracy in the beginning stages of L2A

In this experiment, we were interested in how beginning learners of German (with English as their L1) calculate the uninterpretable features of agreement, as reflected in the copula verb. Based on our analysis of the forms of *to be* in German using the premises of the Combinatorial Variability Hypothesis, we made the following predictions about the acquisition of those forms by these learners. We hypothesized that the copula form *sind* would account for a larger proportion of the errors in acquisition. If L2 learners are guided by the same universal feature evaluation metric as German acquiring children are, then we expect English learners of German to demonstrate evidence of overusing *sind* in the process of learning the target agreement morphology, and hence, making more errors with it. Next, we discuss an experimental study that investigates this hypothesis.

5. The experimental study

5.1. Participants, testing procedure and materials

Twenty-four beginner and 18 intermediate learners of German took a written test. The proficiency levels were established based on the number of class hours of exposure to German instruction at a US university. At the time of study, the beginners were exposed to roughly 40 hours of German classroom instruction; the intermediate learners—to 140 hours. We also tested 17 native German speakers in Düsseldorf, Germany.

The test was a written pen-and paper task and contained simple sentences with missing subjects. Participants had to choose which subject (out of four options) went well with the provided sentence. They could choose more than one option, which was exemplified in the instructions to the test. The test contained six items for each form of *sein* and 10 fillers with other verbal forms, for a total of 40 items.

(10) _____ bist ein guter Freund.
are a good friend

☐ Moritz
☐ du 'you'
☐ die Schüler 'the students'
☐ er 'he'
☐

5.2. Results

A first, very casual look at the data suggests that our hypothesis is indeed supported. Table 1 tabulates the percentages of errors with *sind* versus all the rest of the forms. As the table illustrates, errors with *sind* are roughly double those with other forms of the copula.

Table 1: Percentage errors combined: *sind* versus all the rest of the forms

Type of error	Beginners		Intermediate	
	non-sind	*sind*	*non-sind*	*sind*
Not choosing correct subject	6.75	18.2	5.83	15.2
Choosing wrong subject	8.05	12.5	3.1	5.1
Combined errors	**7.4**	**15.35**	**4.46**	**10.15**

However, a more careful look reveals another picture. Tables 2 and 3 present the error rates for each individual form of the copula by beginning and intermediate learners of German, respectively. As the reader can ascertain, errors with *ist* and *seid* are not half of the errors with *sind* for the beginning learners, while errors with *ist* are more than those with *sind*, for the intermediate learners.

Table 2: Percentage errors by verb form for beginning learners

Type of error	*bin*	*bist*	*ist*	*seid*	*sind*
Not choosing correct subject	2.1	3.5	8.9	12.5	18.2
Choosing wrong subject	4.82	4.86	9.5	13.05	12.5
Combined errors	**3.46**	**4.18**	**9.2**	**12.75**	**15.35**

Table 3: Percentage errors by verb form for intermediate learners

Type of error	*bin*	*bist*	*ist*	*seid*	*sind*
Not choosing correct subject	1	3.7	16.7	1.9	15.2
Choosing wrong subject	0.6	1.5	5.5	4.8	5.1
Combined errors	**0.8**	**2.6**	**11.1**	**3.35**	**10.15**

However, an even more careful look at the data reveals that the picture changes when we divide the answers according to the types of subjects the learners had to choose. They made a disproportionate amount with errors with full DP subjects, as compared to choosing pronoun subjects, as Table 4 demonstrates.

Table 4: Percentage errors in all forms of *sein* depending on type of subject

Type of error	Beginners	Intermediate
Errors in choosing correct **pronoun** subjects	7.50	4.50
Errors in choosing correct **DP** subjects	20.18	29.80

How can this discrepancy be explained? A consideration of which subjects appear in the input with which copula forms points to an answer. The forms for the 1st and 2nd person singular copula are most often encountered in the input together with equivalent pronouns for *I* and *you*-sg. In a language classroom, they are learned in a paradigm, by rote. That is why, learners create the easiest lexical associations: *I* + *am*, *you* + *are*, etc. On the other hand, the form for 3rd person singular most often appears with overt subject DPs like *the teacher* or *the student.* The probability of *ist* combining with any DP subject instead of *er* or *sie* is much higher (in comparison to 1st and 2nd person forms). What is more, the paradigm learned by rote does not provide the learner with DP subjects for the forms of the copula, the agreement has to be calculated. We believe it was the interpretable features of the subject that our participants had difficulty calculating, which leads us to eliminate them from consideration, since in this paper we are concerned with uninterpretable feature calculation.

Table 5: Percentage errors and standard deviations on **pronoun** subjects

	bin	*bist*	*seid*	*sind*
Beginners	4.82 (0.13)	4.86 (0.16)	7.6 (0.19)	10.4 (0.23)
Intermediates	0.8 (0.09)	2.6 (0.12)	1.9 (0.12)	8.4 (0.23)

The final look at the data in Table 5 abstracts away from DP subjects and only calculates the errors in supplying pronoun subjects. One-way ANOVA reveals a significant difference between accuracy on different forms of *sein*, $F(3, 527) = 3.41$, $p = 0.02$ for the beginners, $F(3, 395) = 2.84$, $p = 0.04$. The Combinatorial variability Hypothesis prediction is partially supported: the beginning learners make roughly twice more errors with *bin*, *bist* and *seid*, compared to *sind*, while the intermediate learners improve accuracy on all other forms but *sind*, thereby increasing the accuracy gap.

6. Discussion

To summarize what we have uncovered in this experimental study, we repeat some interesting observations from the data. First of all, the intermediate learners do not get much better with *sind*, compared to the beginners, after a considerably longer exposure to the language (see Table 5). This is a puzzling result. Secondly, the intermediate learners demonstrate even worse accuracy than the beginners when they have to combine a DP subject with the copula (see Table 4). This low accuracy is despite possible L1 transfer of the person and number calculating mechanism. After all, the learners' native English and their L2 German work in exactly the same way, with respect to these features. Furthermore, we find that on the crucial *sind* forms, the intermediate and the beginning learners are not sufficiently different by two tailed t-test ($P = 0.75$). Finally and most importantly, the error rates with *sind* are roughly double those with *bin* and *bist* when subjects are pronouns.

Next, we consider how the proposals that we briefly discussed in the introduction can explain our findings. To start with, the Prosodic Transfer Hypothesis has not been tested, so we leave it aside**.** The Feature Assembly Hypothesis cannot explain the findings, since the L1 and the L2 in this particular case make use of exactly the same features and they are mapped in exactly the same configuration, so no re-assembly is necessary. Why then the persistent difficulty?

The Contextual Complexity Hypothesis cannot explain the findings, either, since the number of terminal nodes in whose context the different forms of *sein* appear is exactly the same, so we should not expect any difference in error rates. This hypothesis predicts that error rates in acquisition of the copula forms should be exactly the same.

The emergentist explanation is not supported either. The forms of *sein* in German are suppletive, hence irregular and should be highly dependent on the frequency factor. If it is robustness of presence in the input that gives rise to stronger representations, hence less errors, then we would expect the following hypothetical error rates as in Table 6, based solely on frequency. We used *A Frequency Dictionary for German*, Randal Jones and Erwin Tschirner, Routledge, (2006) to get the frequency of the lexical items. The dictionary is based on the Leipzig/BYU Corpus of 4.2 million words of spoken and written German, which contains one million each of spoken language, literature, newspapers, and academic texts, and 200,000 words of instructional language. The frequencies in Table 6 are per one million words. We calculated that if we take *ist* as the starting point with roughly 10 000 occurrences, we expect the least mount of error with it. Next, *sind* is roughly three times less common, hence we expect three times more errors, etc.

Table 6: Frequency of verb forms in German corpus and hypothetical error rates

Form	Frequency	Expected error %
ist	10,229	1-2
sind	3,835	6-7
bin	1,025	20-22
bist	260	76-78
seid	36	< 100

But what if the observed error rates in our experiment are not based on frequency of lexical forms in the German language usage captured in the frequency dictionary, what if errors are based on the actual input the learners are exposed to? We also counted the occurrences of the forms of *sein* in the textbook to which the learners are exposed. Table 7 presents the results for the counts from half of the textbook (six of twelve chapters—odd chapters only).

Table 7: Frequency of verb forms in German textbook and hypothetical error rates

Form	Token frequency	Expected error %
ist	480	1-2
sind	73	6-12
bin	11	48-96
bist	11	48-96
seid	3	< 100

Note that even if we do not abstract away from DP subjects, as we do in Table 5, frequency alone cannot explain the error rates in Tables 3 and 4. In addition, none of the factors put forward by Goldschneider & DeKeyser (2001), namely, perceptual salience, semantic complexity, morpho-phonological regularity, and syntactic category, are relevant for explaining the error rate discrepancies, as these factors do not distinguish between the forms of *sein*.

The Missing Surface Inflection Hypothesis is supported and actually extended by proposing that variation stems from the grammatical representations within the learners' interlanguage grammars, and is not at all random. Recall that the latter hypothesis argues that learners may resort to defaults, that is, forms that are underspecified in some features. But that hypothesis does not really explain how exactly are defaults chosen. In contrast, the Combinatorial Variability Hypothesis explains intra-personal morphosyntactic variation as arising from the combinatorial mechanisms of language itself. It offers an evaluation metric of uninterpretable feature combinations, predicting which forms are going to surface more, and become defaults, in the acquisition of inflection. We argue that the attested error rate discrepancies in this study are best explained by the Combinatorial Variability Hypothesis, which we extended to L2A. Our proposal is in consort with another recent proposal, that of McCarthy (2007).

In conclusion, we argue that there may be different underlying sources for variable L2 morphosyntactic performance: feature re-assembly, perceptual salience, phonological regularity, and semantic complexity may all be relevant. When all of these factors are held constant, though, and errors still persist, only language-internal explanations, such as combinatorial variability of features, are really credible.

References

Adger, D. (2006). Combinatorial Variability. *Journal of Linguistics*, 42, 503-530.

Bybee, J. (2001). *Phonology and Language Use*. CUP.

Chomsky, N. (2001). Derivation by phase. In Kenstowitz, M. (Ed.), Ken Hale: *A Life in Language*, pp. 1-52. Cambridge, MA: MIT Press.

Ellis, N. C. (2002). Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143-188.

Ellis, N. C. (2006a). Language acquisition as rational contingency learning, *Applied Linguistics* 27, 1: 1–24.

Ellis, N. C. (2006b) Selective attention and transfer phenomena in L2 Acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics* 27, 2: 164–194.

Goad, H., and White, L. (2004). Ultimate attainment in L2 inflection: Effects of L1 prosodic structure. In S. Foster-Cohen, M. Sharwood Smith, A. Sorace, & M. Ota, (Eds.), *Eurosla Yearbook 4* (pp. 119-145). Amsterdam: John Benjamins.

Goad, H., and White, L. (2006). Ultimate attainment in interlanguage grammars: A prosodic approach. *Second Language Research*, 22, 3, 243-268.

Goad, H., White, L. and Steele, J. (2003). Missing inflection in L2 acquisition: defective syntax or L1 – constrained prosodic representations? *Canadian Journal of Linguistics* 48, 243-263.

Goldschneider, J. M. and R. DeKeyser. (2001). Explaining the ''natural order of L2 morpheme acquisition'' in English: A meta-analysis of multiple determinants, *Language Learning* 51: 1–50.

Hawkins, R. and Casillas, G. (2007) Explaining frequency of verb morphology in early L2 speech. *Lingua* (article in press, available online from journal web site).

Lardiere, D. (2007). *Ultimate Attainment in Second Language Acquisition. A case Study*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lardiere, D. (2008). Feature assembly in second language acquisition. In Liceras, J., H. Zobl and H. Goodluck (Eds.), *The Role of Features in Second Language Acquisition*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

McCarthy, C. (2007). Morphological variability in second language Spanish. Unpublished PhD dissertation, McGill University.

Pesetsky, D. and Torrego, E. (2001). T-to-C movement: causes and consequences. In Kenstowitz, M. (Ed.), Ken Hale: *A Life in Language*, pp. 355-426. Cambridge, MA: MIT Press.

Prévost, P and White, L. (2000) Missing surface inflection hypothesis or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research*, 16, 103-133.

Smith, J. (2000). *Synchrony and diachrony in the evolution of English: Evidence from Scotland*. PhD dissertation, University of York.

Selected Proceedings of the 2007 Second Language Research Forum

edited by Melissa Bowles, Rebecca Foote, Silvia Perpiñán, and Rakesh Bhatt

Cascadilla Proceedings Project Somerville, MA 2008

Copyright information

ISBN 978-1-57473-425-6 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Slabakova, Roumyana and Johnathan Gajdos. 2008. The Combinatorial Variability Hypothesis in the Second Language. In *Selected Proceedings of the 2007 Second Language Research Forum*, ed. Melissa Bowles, Rebecca Foote, Silvia Perpiñán, and Rakesh Bhatt, 35-43. Somerville, MA: Cascadilla Proceedings Project.

or:

Slabakova, Roumyana and Johnathan Gajdos. 2008. The Combinatorial Variability Hypothesis in the Second Language. In *Selected Proceedings of the 2007 Second Language Research Forum*, ed. Melissa Bowles, Rebecca Foote, Silvia Perpiñán, and Rakesh Bhatt, 35-43. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1733.