

Vyatkina, N. (2013). Analyzing part-of-speech variability in a longitudinal learner corpus and a pedagogic corpus. In S. Granger, G. Gilquin & F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 479-491. Publisher's official version: <http://www.uclouvain.be/en-404131.html>. Open Access version: <http://hdl.handle.net/1808/11215>.

Please share your stories about how [Open Access to this item benefits you](#).

[This document contains the author's accepted manuscript. For the publisher's version, see the link in the header of this document.]

Paper citation:

Vyatkina, N. (2013). Analyzing part-of-speech variability in a longitudinal learner corpus and a pedagogic corpus. In S. Granger, G. Gilquin & F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 479-491.

Keywords:

learner corpus analysis
pedagogic corpus
parts-of-speech
automatic profiling

Abstract:

This study investigates the development of part-of-speech variety in the writing of a cohort of beginning college-level learners of German over three semesters of study in comparison with the pedagogical input they received from their workbook. The study fills existing gaps in Second Language Acquisition research by targeting beginner learners of German as a foreign language, analyzing semi-automatically annotated corpora (a learner corpus and a corresponding workbook corpus), and eliciting learner data over a long period of time at dense time intervals. As a result, it presents a developmental Second Language (L2) profile of the target learner population in terms of verb classes and verb morphology. The study shows how participants gradually enrich their verb form repertoire, both in accordance with and diverging from the pedagogical input they receive.

Text of paper:

Analyzing part-of-speech variability in a longitudinal learner corpus and a pedagogic corpus

Nina Vyatkina

University of Kansas

1. Introduction

An increasing number of Second Language Acquisition (SLA) researchers have become “interested in instabilities, variation, and discovering larger developmental trajectories rather than focusing on the discovery of stable conditions that apply uniformly at a particular time” (Byrnes 2009: 63). This trend has been accompanied by calls to account for degrees of such variation, to gather dense developmental data, and to report on individual differences which may be masked by cross-sectional averages (Ellis & Larsen-Freeman 2006). Larsen-Freeman & Cameron (2008) suggest that collecting and analyzing

Please share your stories about how [Open Access to this item benefits you](#).

longitudinal learner corpora may propel SLA research in these directions. It is a welcome current development that scholars in Learner Corpus Research (LCR), a field that has been long dominated by large-scale cross-sectional studies, started to join in (e.g. Hasko & Meunier 2013). Notably, Meunier (2010) in her recent “checkup” writes a prescription for a healthy development of LCR: collection of smaller-scale, longitudinal, and locally contextualized corpora as well as annotation of these corpora for parts-of-speech (POS) and syntactic categories. The present study responds to this call by analyzing a POS-annotated dense longitudinal learner corpus with the goal of closely tracking the writing development of a cohort of beginning college-level learners of German over three semesters of study in comparison with the pedagogical input they received.

2. Research background

2.1. POS measures as indicators of grammatical complexity

This study investigates the development of L2 complexity in terms of range and variety of specific morphosyntactic forms, an area underexplored in SLA research (Lu 2011; Norris & Ortega 2009). POS measures have been used as surface indicators of underlying grammatical complexity in learner texts (Aarts & Granger 1998; Borin & Prütz 2004; Lu 2011). POS are located at the intersection between the linguistic levels of grammar and lexicon and thus present a promising locus for research on complexity at a higher level of abstraction than afforded by word-based analyses (Meurers & Müller 2009). With advances in corpus linguistics, automatic tagging can considerably speed up POS analyses of large amounts of linguistic data. For native speaker (NS) corpora, automatic taggers achieve a 95-96% accuracy (Schmid 1994). Although they perform worse on learner language (van Rooy & Schäfer 2003), a number of studies have attacked SLA research questions using corpus tagging software.

Granger & Rayson (1998: 119) proposed fully annotating learner corpora for the purpose of “automatic profiling” of learner language by establishing a “unique matrix of frequencies of various linguistic forms”. They compared POS frequencies in two similar-sized POS-annotated corpora of argumentative essays: a corpus collected from advanced L1 French English learners and a comparison NS English corpus. The results showed that learners underuse nouns, conjunctions, and prepositions but overuse pronouns, determiners, and adverbs in comparison to NSs. However, when the authors zoomed in on finer-grained subclasses, they found that only indefinite determiners accounted for learner overuse, whereas definite determiners were, in contrast, underused. Granger & Rayson conclude that learner academic essays showed more features indicative of orality and involvement (indefinite determiners, 1st and 2nd person pronouns, auxiliaries, and infinitives) than of expert academic writing (definite determiners, lexical verbs, and participles). This conclusion parallels findings from a number of other studies investigating intermediate to advanced L2 English writing of learners with various L1 backgrounds: Reid (1992) for cohesive devices, Aarts & Granger (1998) for pronouns, auxiliaries, and adverbs, and Borin & Prütz (2004) for conjunctions, adverbs, and participles. An example of using automatic profiling of annotated learner corpora for defining L2 proficiency measures is a large-scale

Please share your stories about how [Open Access to this item benefits you](#).

project recently taken up by the English Profile Programme (*e.g.* Hawkins & Buttery 2010; Saville 2010). In sum, most significant advances performed on annotated corpora have been achieved in cross-sectional studies.

2.2. Pedagogic corpora

Studies reviewed above involved learners at relatively advanced proficiency levels and compared their L2 written production with baseline corpora of comparable NS writing samples. However, when it comes to learners at lower proficiency levels, it becomes harder to define what baseline their production can be compared to. One of the promising avenues in this respect has emerged in recent studies reporting on creation and analysis of pedagogic corpora. The term “pedagogic corpus” was suggested by Hunston (2002: 16) who defined it as a “corpus consisting of all the language a learner has been exposed to”. However, Meunier & Gouverneur (2009: 186), having reviewed relevant research, posit that this definition is rather unrealistic and suggest a fine-tuned definition of a pedagogic corpus: “a large enough and representative sample of the language, spoken and written, a learner has been or is likely to be exposed to via teaching material, either in the classroom or during self-study activities”. Most pedagogic corpora collected to date represent textbooks.

Studies of textbook corpora have primarily compared them with NS corpora and pointed out notable discrepancies. Some studies have looked at the instructional progression in the presentation of selected structures. Gouverneur (2008) systematically compared pedagogical tasks in an intermediate and advanced English as a Foreign Language (EFL) textbook corpus and found a lack of consistency in the presentation of the lexical material as well as in the task progression. In a rare study that included materials for beginning learners, Römer (2004) analyzed EFL textbooks used in German secondary schools from the beginning through the advanced level and found mismatches in the presentation of *if*-clauses in comparison with the frequency of these structures in NS corpora.

Although textbook corpus studies point to some notable deficiencies in L2 input presented in textbooks, the latter are the primary input source in instructed FL contexts (Römer 2004; Tono 2004). As Tono (2004: 51) argues, “beginning- or intermediate- level texts are designed to contain a level and form of English which can facilitate learning” and therefore, “textbook English is a useful target corpus to use in the study of learner language”. However, to the best of our knowledge, Tono (2004) remains the only study systematically comparing a learner corpus and a textbook corpus. Tono performed a lexical search on ten verbs most frequently appearing in the textbook corpus. The results show that, although learners used the very same ten verbs more frequently, the accuracy of use was not affected by the input frequency. Tono concludes the study with a call for more studies comparing developmental learner corpora and textbook corpora. The present study seeks to address this research gap by comparing a learner corpus and a pedagogic corpus, both POS-annotated.

Please share your stories about how [Open Access to this item benefits you](#).

3. Study design

3.1. Research questions

This study seeks to explore how the word class variety changes in the writing of a cohort of beginning learners of German over time and to answer the following research questions:

- 1) How is the POS profile of writing samples of a learner cohort similar to and different from the POS profile of the specific pedagogical input provided at 14 time intervals over three semesters of collegiate L2 study?
- 2) How do developmental POS profiles of two individual learners compare to each other and to their learner cohort profile?
- 3) What developmental patterns emerge in the learner language and what is a possible explanation for these patterns?

3.2. Participants, tasks, and corpora

The data for the learner corpus were collected from students who enrolled in a beginning German language program at a large public US university over three 16-week-long subsequent semesters. The classes met for 5 weekly contact hours in the first and second semester and for 3 contact hours in the third semester. The participants represented a fairly homogeneous language learner population as all of them had American English as their L1 and most of them grew up in the same region. Furthermore, all participants had no or very little prior knowledge of German or travel experience in German-speaking countries. Finally, their exposure to German was primarily restricted to classroom instruction and teaching materials. The general teaching approach combined communicative activities with focused-based grammar instruction. Writing assignments (essays) supplied data for the learner corpus. Students typed each essay in class, under controlled conditions. They were required to write during whole 50-minute-long class periods and were allowed to use online dictionaries but no other reference materials. The number of samples varied from time point to time point in data collection because not all participants submitted all essays (Table 1).

The data for the pedagogic corpus were taken from Briggs *et al.* (2008), a workbook that included grammar, vocabulary, reading, and writing practice activities. Only the workbook was selected for creating the pedagogic corpus because it was available in electronic format via the Quia© interface. Moreover, the workbook contained all prompts and writing tasks to which learners responded in their essays, so this increases the comparability of the two corpora. The complete text from each workbook chapter was copied and saved as a separate electronic file in the pedagogic corpus database. It must be noted that the workbook text was not differentiated based on pedagogical tasks (*cf.* Gouverneur 2008) and should be considered as a lump pedagogical input to which learners were exposed via self-study during each respective time interval.

Please share your stories about how [Open Access to this item benefits you](#).

Each writing task (1- 14) concluded a respective workbook chapter and reflected the instructional content and focus of the workbook and corresponding textbook chapter. The tasks in chapters 1-5 (1st semester) and chapters 6-10 (2nd semester) requested learners to write personal narratives. Chapters 11-14 (3rd semester) required students to write personal narratives and personal accounts with added explanation elements.

3.3. Method

This is an exploratory empirical study investigating writing complexity in terms of word class variety. For this purpose, the *integrated contrastive model* (Granger 1996) is used; it couples learner language analysis in its own right with contrastive analysis, which compares learner corpora with baseline comparison corpora. First, the method of automatic profiling of corpora (Granger & Rayson 1998) was applied. The focal corpora were automatically tagged for 50 distinct POS using the Tree Tagger for German (Schmid 1994). The tagger accuracy was evaluated by two independent annotators on a sample constituting ca. 9% of the learner corpus. In the tagger output, an automatically assigned tag was evaluated as correct when at least one source of “evidence from distribution, lexis, and morphology” was present, even in cases when the evidence did not “converge on a single POS classification” (Díaz-Negrillo *et al.* 2010: 151) in case of learner errors. For example, in the sentence *Sie haben ihre Wäsche waschen* (‘They have wash their clothes’), the tagger marked the main verb as an infinitive based on its morphological form. This tag was accepted although the student probably intended to use the past participle *gewaschen* (‘washed’). As a result, the tagger accuracy was evaluated at ca. 96% for the learner corpus.¹ In contrast, the tagger output for the workbook corpus revealed a number of systematic errors (such as marking interjections like *hallo* as nouns), and the tagger accuracy varied from 85% to 95% from chapter to chapter. Therefore, a set of rules were formulated, based on which of these systematic errors were manually corrected in the whole workbook corpus. Thus, the workbook corpus was tagged semi-automatically (Garretson & O’Connor 2007).

¹ It should be noted that Tree Tagger most frequently assigns tags based on morphology. It is acceptable for the purposes of the present study (which focuses on L2 complexity but not accuracy). If it were decided to use the distributional information or the “target hypothesis” (Lüdeling *et al.* 2005) as the basis for POS tagging, then the tagger accuracy would have been substantially lower. See Díaz-Negrillo *et al.* 2010 for a discussion of what it means to “accurately” assign POS-tags to learner language.

Please share your stories about how [Open Access to this item benefits you](#).

Morphosyntactic variety was measured as frequency of certain grammatical forms considered to be sophisticated (Ellis & Yuan 2005; Robinson 2007). For this study, four verb forms were selected as grammatically sophisticated for beginning learners of German: separable verb prefixes (SEP), verb past participles (VVPP), reflexive pronouns indexing reflexive verbs (REF), and infinitival constructions (INF)² because they sequentially served as instructional foci at different time points: SEP at T4, VVPP at T7, REF at T8, and INF at T13.

Semester	Time point / Chapter	Learner corpus			Workbook corpus
		Samples	tagged words (total)	tagged words (mean)	tagged words (total)
first	1	28	1914	68	1512
	2	25	2176	87	2294
	3	26	3404	131	1939
	4	27	2864	106	2455
	5	25	2445	98	2360
second	6	40	3854	96	2967
	7	40	4164	104	2239
	8	29	3056	105	2381
	9	38	4083	107	2945
	10	35	4072	116	3123
third	11	30	4148	138	2964
	12	24	3627	151	2533
	13	24	3517	147	3278
	14	20	2758	138	2720

Table 1. Size of corpus subsets

Next, the *WordList* tool of *WordSmith Tools*© (Scott 2008) was run on the tagged corpus to automatically compute frequencies of the four selected POS.³ Frequency data were transferred to Excel, normalized per 100 words, and plotted in stacked column graphs that visualize the repertoire and proportion of the forms used at each point on the time line. Then the data were explored in terms of multidimensional qualitative variability (van Geert & van Dijk 2002). This method helps analyze not only differences in the levels of the measured variables but also appearance and disappearance of certain variables at different time points. Two sets of comparative developmental profiling were performed:

² For the sake of readability, original automatically assigned tags are replaced here with more transparent English acronyms and abbreviations.

³ SEP, VVPP, and REF were automatically recognized by the tagger, and the number of INF was computed by adding the frequencies for the infinitival particle *zu* ('to') and infinitives with *zu* inserted as an interfix.

Please share your stories about how [Open Access to this item benefits you](#).

1) mean frequencies for the learner cohort vs. frequencies for the workbook corpus; and 2) comparisons of individual data for two selected learners. The focal learners (who were given the pseudonyms ‘Braden’ and ‘Cassie’) were selected based on the results of two earlier studies (Vyatkina 2012, 2013). Those studies showed that the writing of both learners was close to the cohort average on a number of general syntactic complexity measures, although they were different both from each other and the cohort average on other, more specific measures. Therefore, it seemed interesting to explore their performance vis-à-vis more specific morphosyntactic features. Finally, a contextual analysis of the target features was performed using the *Concord* tool of *WordSmith Tools*©.

4. Results

4.1. POS profiles of the workbook corpus and the learner corpus

A comparison of the two graphs shows that spikes in learner use of the target verb forms occur as an immediate response to focused instruction in each of these forms. A qualitative analysis showed that these frequencies are especially high at time points when essay tasks contained explicit prompts triggering the use of the target forms. More specifically, the T4 task listed a number of verbs with separable prefixes that the learners were asked to use in their essay to describe their daily routine, and T7 task contained questions in the present perfect tense which logically triggered answers with VVPP forms. This result was desired and expected from the pedagogical perspective because one of the instructional purposes of the essay tasks was practicing target grammatical forms in free but prompted production.

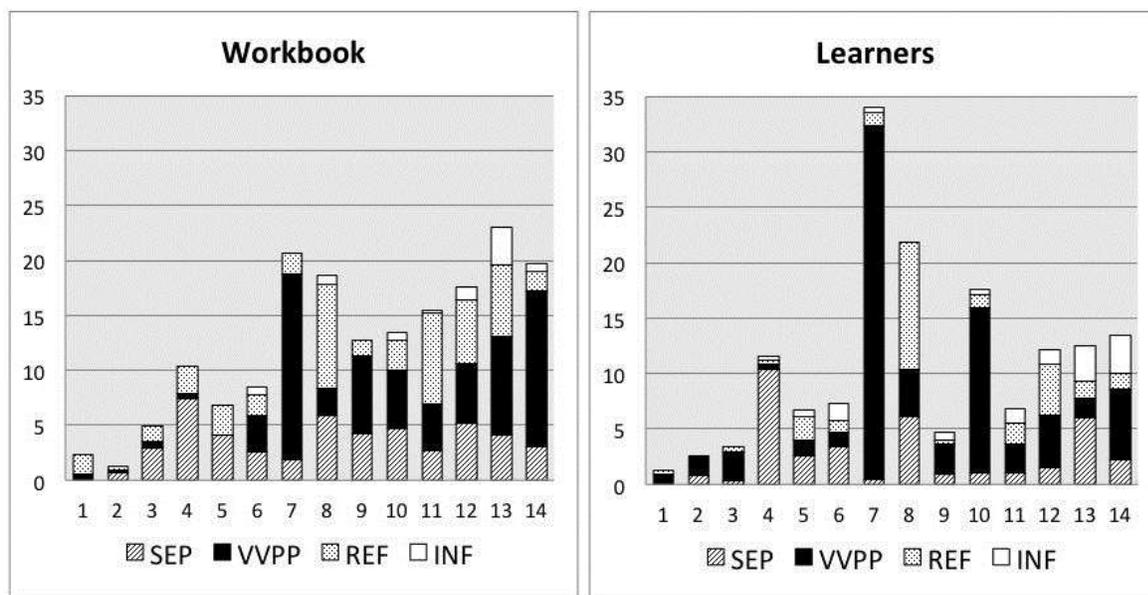


Figure 1. POS frequencies per 100 words in the workbook corpus and the learner corpus

Please share your stories about how [Open Access to this item benefits you](#).

Next, it is apparent from the graph that once picked up, the target POS were never dropped in learner production: although relative frequencies for each form vary from time point to time point, each measurement occasion supplied several of each focal verb form. Moreover, learners occasionally use 'more advanced' forms prior to focused instruction. For example, past participles are sporadically used at each time point from the very beginning before being formally introduced in chapter 7. This fact may be explained by prior exposure to these forms in the case of 'false beginners' or consulting the online dictionary or grammar as well as the teacher during essay writing. Additionally, there is another reason for high frequencies of VVPP at early time points. Although present perfect had not been explicitly taught until T7, the textbook and the workbook repeatedly listed several present perfect forms in the early chapters, particularly the ones used as chunks. These forms apparently served for learners as input and models for their production. In fact, 32 out of 42 occasions of VVPP in learner writing at T1 falls at the expression *Ich bin in... geboren* ('I was born in...') and 24 out of 52 occasions at T2 at the expression *Das Zimmer/die Wohnung/das Haus ist möbliert* ('the room/apartment/house is furnished').

However, the two charts also exhibit marked differences. First, the growth in new verb forms is more consistent in the pedagogic corpus: frequencies of each new form increase linearly, thus gradually expanding the overall repertoire. In contrast, the variability is much higher in the learner corpus as illustrated by sharp spikes and drops. A salient example is again the relative frequency of past participles in the learner corpus at T7, which is twice as high as in the workbook corpus (32:17). This difference can be explained by the fact that the workbook chapter covers additional instructional foci and uses additional verb forms other than the present perfect. On the other hand, learners use the present perfect (the appropriate verb tense for narrating about past events in German) almost exclusively while responding to the topic of this writing assignment (describing their past weekend) as well as while utilizing specific VVPP prompts provided for this task. Another surge in the use of past participles in the learner data occurs at T10, when students were again asked to describe a past event.

4.2. POS profiles of two individual learners

The longitudinal data presented in Figure 2 show that Braden and Cassie contributed to the class average described above in two very different ways. Braden used only two of the focal forms (SEP and VVPP), and only at time points when these forms were the focus of instruction (T4 and T7, respectively). After a long break, he only used SEP again at T13.⁴ Cassie's data present a very different picture. Her use of the focal verb forms is more varied and much more evenly distributed across the timeline than Braden's. First, she starts using all focal POS at the time of explicit instruction or prior to it. In fact, she tries out all four POS already during the first semester (T1-5). From T11 on, she regularly uses 2 to 3 of the four focal verb forms at each measurement occasion until the close of the observation period. She

⁴ All focal POS reappear in Braden's writing in the 4th semester, the data from which are beyond the scope of this paper (see Vyatkina 2013).

Please share your stories about how [Open Access to this item benefits you](#).

even uses the most advanced INF at each time point from T11 onward.

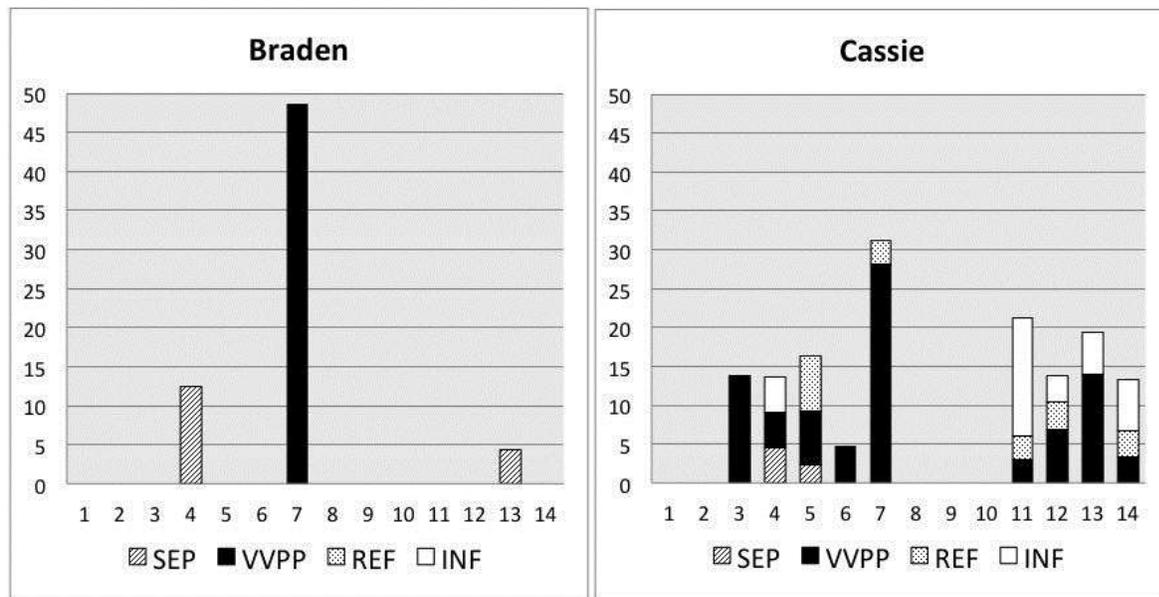


Figure 2. POS frequencies per 100 words for Braden and Cassie

To shed more light onto individual differences between Braden and Cassie, occurrences of one verb form, VVPP, were explored in context. For that purpose, concordance lines (Figure 3) and frequency lists of all VVPP-tagged words were retrieved from these two learners' data. It was found that Cassie begins fairly extensively and consistently using past participles even before instruction, from T3 onward. Although four instances at T3 are accounted for by the form *geboren* ('born') introduced in unanalyzed chunks such as ('I was born'), she uses most past participles in free constructions with present perfect such as *Ich habe nich [sic] beschlossen* ('I have not decided'). After focused instruction, Cassie resumes using past participles at T12 and consistently uses them at each time point. In contrast, Braden uses past participles only at T7, *i.e.* as a direct response to the focused instruction as well as to the writing prompt that contains specific VVPP forms. It is noteworthy that both learners used VVPP forms appropriately (although not always accurately) in present perfect (or, in a few instances, in past perfect) constructions to express past events while rendering oral narration or past events preceding more recent events.

Furthermore, the Type-Token Ratio (TTR) of unique verb forms to the total number of all past participles used was calculated for each learner. It turned out that Cassie used not only more VVPP tokens than Braden but also more unique verb types. For example, at T7, Braden used 9 different verb types out of the total of 17 past participles with the TTR of 0.53, whereas Cassie used 8 types but out of only 10 tokens, with a resulting TTR of 0.8. A high TTR was sustained in Cassie's subsequent VVPP uses. In other words, Braden more frequently repeated the same forms when he used past participles (*e.g.*, *gegangen*, *gesehen*, *gespielt*, Figure 3), whereas Cassie more frequently built past participles from unique verbs

Please share your stories about how [Open Access to this item benefits you](#).

(e.g., *gestorben*, *ersetzt*, *gerettet*, Figure 3). Moreover, most of Braden’s T7 past participles (15 out of 17) were directly emulated from the respective workbook chapter and writing prompt, whereas Cassie emulated only four forms (out of nine) from the workbook at the same time point. Additionally, Cassie frequently used novel VVPP forms before they appeared in the workbook. For example, she (appropriately) used the form *besucht* (‘visited’) at T4 and *gekauft* (‘bought’) at T4 and T5, although both of them were formally introduced only at T7. On top of that, there are examples when Cassie’s VVPP form was emulated from the pedagogical material but the context of use was novel. For instance, the past participle *verbracht* (‘spent’) was used in the workbook as part of the phrase *Zeit verbringen* (‘to spend time’) at T7, whereas Cassie used it at the same time point in a variation of this expression – *den ganzen Tag verbringen* (‘to spend the whole day’).

N	Concordance	N	Concordance
1	NN letztes ADJA Wochenende NN gemacht VVPP	14	meine PPOSAT Eltern NN sind VAFIN gekommen VVPP
2	KON vierzehn CARD Uhr NN Klassen NN gegangen VVPP	15	Ihre PPOSAT BÄrden NN werden VAFIN ersetzt VVPP
3	nach APPR meine PPOSAT Hause NN gegangen VVPP	16	Arbeit NN am APPRART Samstag NN gegangen VVPP
4	Mitbewohners NN Videospiele NN gespielt VVPP	17	an APPR meinem PPOSAT Haus NN geblieben VVPP
5	Halo NN und KON Rock NN Band NN gespielt VVPP	18	ADJA Tag NN auf APPR das ART Sofa NN verbracht VVPP
6	habe VAFIN ich PPER einen ART Film NN gesehen VVPP	19	NN am APPRART Sonntagnacht NN gegangen VVPP
7	. \$. Ich PPER habe VAFIN Cloverfield NN gesehen VVPP	20	NN habe VAFIN ich PPER sie PPER gebraucht VVPP
8	ich PPER mit APPR Freunde NN ausgegangen VVPP	21	PPOSAT Autobatterie NN war VAFIN gestorben VVPP
9	PPER haben VAFIN 'Disc NE Golf NN " \$(gespielt VVPP	22	neustarten NN (\$(' \$(jumpstart' NE) \$(gebraucht VVPP
10	ins APPRART Centennial NN Park NN gespielt VVPP	23	Meine PPOSAT Kinder NN können VMFIN erfahren VVPP
11	habe VAFIN für APPR ein ART Party NN geduscht VVPP	24	NN von APPR einem ART Flugzeug NN gerettet VVPP
12	meine PPOSAT Friends NN Hause NN gegangen VVPP	25	NN gerettet VVPP . \$. Es PPER abgestürzt VVPP
13	habe VAFIN Bier NN und KON Alkohol NN gesoffen VVPP	26	ART Flugzeug NN , \$, aber KON Walt NE gefunden VVPP
14	das ART KU NE Fußball NN Spiel NN gesehen VVPP	27	NN Hurley NE eine ART TV-Skript NN gefunden VVPP
15	meine PPOSAT Friends NN Boden NN geschlafen VVPP	28	Er PPER hat VAFIN viele PIAT Bücher NN gelesen VVPP

Figure 3. VVPP concordances for Braden (left) and Cassie (right)

5. Discussion and implications

This study has sought to investigate the dynamics of word class variety in learner writing in comparison with pedagogical input. All occurrences of selected verb forms considered sophisticated for beginning learners of German were tracked over time, which reflected the writing complexity development of this learner cohort. The general developmental trend is toward a greater variety of word classes, a desirable learning outcome which also generally emulates the instructional progression. This general developmental course is expected as learners slowly enrich their repertoire of grammatical forms in accordance with the pedagogical input they receive. However, although this trend toward increasing variety is predominantly linear in the pedagogic corpus, it is never that ‘clean’ in the learner corpus. Although spikes in the learner use of the focal POS are triggered by the instructional focus in a specific workbook chapter, they are usually higher in the learner corpus. This can be explained by the fact that learners concentrate on a few specific new forms in their writing (frequently reinforced by specific writing prompts), whereas the workbook input remains more evenly distributed. High variability in learner writing was observed throughout the time line (three semesters of study) and was especially

Please share your stories about how [Open Access to this item benefits you](#).

salient when inspected for two individual learner cases. Whereas Cassie's development was more balanced and characterized by a richer variety of both grammatical forms and their lexical content, Braden's profile shows a few high frequency spikes of focal forms expressed by a limited number of lexemes. The observed distributional differences show that learner production is influenced by, but does not directly mirror, the pedagogical input. As the learners in this study were beginners who experienced their first exposure to all grammatical features of their new language, more extended longitudinal studies are needed to observe cumulative input and practice effects.

This study used semi-automatically POS-tagged written corpora and an integrated corpus analysis methodology to provide a multidimensional developmental profile of the writing of a learner cohort in comparison with a corresponding workbook input. Following the method suggested by Granger & Rayson (1998: 130), the study has "shown that automatic profiling can help researchers form a quick picture of the interlanguage of a given learner population". Its findings contribute to the empirical research on L2 complexity by focusing on beginning learners of L2 German and tracking their emergent grammar in correspondence with their very first exposure to new forms in pedagogical materials. In this way, the study helps to account for how and why language competencies develop for specific learners and target languages, in response to particular tasks, teaching, and other stimuli, and mapped against the details of developmental rate, route, and ultimate outcomes. (Norris & Ortega 2009: 557).

Importantly, the study demonstrates a noticeable influence of the pedagogical input, instructional focus, and writing prompt onto learner writing while also pointing out divergences between the learner corpus profile and the workbook corpus profile. In this way, it points to the dynamic nature of language development that cannot be predicted by a predetermined instructional progression alone but is also influenced by a variety of other factors including learner agency.

Based on the results, the following directions for future research can be suggested. First, the method applied here can be used to analyze other learner and pedagogic corpora. Collecting sets of learner profiles from various populations in various settings will give SLA researchers empirical grounding for formulating and testing new developmental hypotheses. Next, the profiles presented here may serve as a stepping stone for more detailed qualitative analyses of the distribution of specific words used by individual learners as mapped against a greater range of POS. Finally, POS analysis can be combined with analyses of other complexity, accuracy, and fluency measures to arrive at truly multidimensional developmental profiles of learner language varieties, preferably using non-linear and dynamic research methods (van Geert & van Dijk 2002).

Vyatkina, N. (2013). Analyzing part-of-speech variability in a longitudinal learner corpus and a pedagogic corpus. In S. Granger, G. Gilquin & F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 479-491. Publisher's official version: <http://www.uclouvain.be/en-404131.html>. Open Access version: <http://hdl.handle.net/1808/11215>.

Please share your stories about how [Open Access to this item benefits you](#).

Acknowledgments

This study was supported in part by the University of Kansas General Research Fund allocations Nos. 2302139 and 2301446. I would like to acknowledge Emily Hackmann for her help in evaluating the tagger accuracy.

References

Aarts, J. & Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (ed.) *Learner English on Computer*. New York: Longman, 132-141.

Borin, L. & Prütz, K. (2004). New wine in old skin? A corpus investigation of L1 syntactic transfer in learner language. In G. Aston, S. Bernardini & D. Stewart (eds) *Corpora and Language Learners*. Amsterdam: John Benjamins, 67-87.

Briggs, J., Di Donato, R., Clyde, M. & Vansant, J. (2008). *Workbook to Accompany Deutsch, Na Klar!: An Introductory German Course*. New York: McGraw-Hill.

Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education* 20, 50-66.

Díaz-Negrillo, A., Meurers, D., Valera, S. & Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* 36, 139-154.

Ellis, N. & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics – Introduction to the special issue. *Applied Linguistics* 27, 558-589.

Ellis, R. & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (ed.) *Planning and Task Performance in a Second Language*. Amsterdam: John Benjamins, 167-192.

Garretson, G. & O'Connor, M. C. (2007). Between the Humanist and the Modernist: Semi-automated analysis of linguistic corpora. In E. Fitzpatrick (ed.) *Corpus Linguistics beyond the Word: Corpus Research from Phrase to Discourse*. Amsterdam: Rodopi, 87-106.

Gouverneur, C. (2008). The phraseological patterns of high-frequency verbs in advanced English for general purposes: A corpus-driven approach to EFL textbook analysis. In F. Meunier & S. Granger (eds) *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins, 223-243.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (eds) *Languages in contrast. Papers from a*

Vyatkina, N. (2013). Analyzing part-of-speech variability in a longitudinal learner corpus and a pedagogic corpus. In S. Granger, G. Gilquin & F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 479-491. Publisher's official version: <http://www.uclouvain.be/en-404131.html>. Open Access version: <http://hdl.handle.net/1808/11215>.

Please share your stories about how [Open Access to this item benefits you](#).

symposium on text-based cross-linguistic studies. Lund: Lund University press, 37-52.

Granger, S. & Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (ed.) *Learner English on Computer*. New York: Longman, 119-131.

Hasko, V. & Meunier, F. (eds) (2013). Capturing L2 development through learner corpus analysis [Special issue]. *Modern Language Journal* 97(S1).

Hawkins, J. & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal* 1, 1-23.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Larsen-Freeman, D. & Cameron, L. (2008). Research methodology on language development from a complex systems perspective. *The Modern Language Journal* 92, 200-213.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 45, 36-62.

Lüdeling, A., Walter, M., Kroymann, E. & Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of Corpus Linguistics 2005*, Birmingham, UK.

Meunier, F. (2010). Learner corpora and English language teaching: Checkup time. *Anglistik: International Journal of English Studies* 21, 209-220.

Meunier, F. & Gouverneur, C. (2009). New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (ed.) *Corpora and Language Teaching*. Amsterdam: John Benjamins, 179-201.

Meurers, W.D. & Müller, S. (2009). Corpora and syntax. In A. Lüdeling & M. Kytö (eds) *Corpus Linguistics*. Berlin: Mouton de Gruyter, 920-933.

Norris, J. & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30, 555-578.

Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing* 1, 79-107.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied*

Vyatkina, N. (2013). Analyzing part-of-speech variability in a longitudinal learner corpus and a pedagogic corpus. In S. Granger, G. Gilquin & F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 479-491. Publisher's official version: <http://www.uclouvain.be/en-404131.html>. Open Access version: <http://hdl.handle.net/1808/11215>.

Please share your stories about how [Open Access to this item benefits you](#).

Linguistics 45, 237-257.

Römer, U. (2004). Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini & D. Stewart (eds) *Corpora and Language Learners*. Amsterdam: John Benjamins, 151-168.

Saville, N. (2010). The English Profile Programme: Background, current issues and future prospects. *Language Teaching* 43, 238-244.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf (last accessed on 10 January, 2012).

Scott, M. (2008). *WordSmith Tools Version 5*. Liverpool: Lexical Analysis Software.

Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernardini & D. Stewart (eds) *Corpora and Language Learners*. Amsterdam: John Benjamins, 45-66.

van Geert, P. & van Dijk, M. (2002). Focus on variability: New tools to study intra- individual variability in developmental data. *Infant Behavior and Development* 25, 340-375.

van Rooy, B. & Schäfer, L. (2002). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies* 20, 325-335.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal* 96, 572-594.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *Modern Language Journal* 97(S1), 11-30.