University of Groningen

Development of chunks in Dutch L2 learners of English

Gustafsson, Hana; Verspoor, Marjolijn

Published in:
Usage-Based Approaches to Language Acquisition and Language Teaching

DOI:
[10.1515/9781501505492-011](https://doi.org/10.1515/9781501505492-011)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
Gustafsson, H., & Verspoor, M. (2017). Development of chunks in Dutch L2 learners of English. In *Usage-Based Approaches to Language Acquisition and Language Teaching* (pp. 235-264). Walter de Gruyter GmbH. <https://doi.org/10.1515/9781501505492-011>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Hana Gustafsson and Marjolijn Verspoor

10 Development of chunks in Dutch L2 learners of English

1 Introduction

Evidence from various research disciplines has demonstrated that a substantial part of native speaker repertoire consists of a wide range of conventionalized expressions (Ellis 2008; Sinclair 1991; Wray 2002) which in this paper are generally referred to as *chunks*. Many chunk types are part of traditional phraseology and clearly recognizable as fixed units, such as *of course*, *living room*, *grow up*, *the sky is the limit*, but those most pervasive in native-like repertoire are simply “normal ways of saying things” (Granger and Paquot 2008: 35; Langacker 2008: 84). These are the preferred ways of expressing certain notions out of all the grammatically correct options available: Compare for example *Will you marry me?* and *Would you like to become my spouse?* (Pawley and Syder 1983). Although many chunks are highly context-specific, in general the proportion of chunks in native-like written language has been estimated at about 50% (Erman and Warren 2000) and it tends to be even higher in native-like spoken language (Ellis 2008: 4).

Since chunks are such a pervasive feature of a native-like repertoire, they are also a crucial aspect of second language (L2) development. Verspoor, Schmid, and Xu (2012) have shown that the number of chunks used in written texts is one of the better measures to distinguish among five L2 English proficiency levels (from beginner to high intermediate). Chunks contribute to fluency and authenticity of L2 use and may also speed up general linguistic development. However, it is precisely this feature of the target language that is often the greatest obstacle for L2 learners. Although the classroom context does focus on chunks, the range is limited (Eyckmans, Boers, and Stengers 2007) and often does not include “normal ways of saying things” as they are not part of traditional phraseology, grammar or lexicon (Langacker 2008: 84). Most importantly, classroom context often provides only limited exposure to authentic input, which from a usage-based perspective is absolutely crucial.

Note: This work was partly supported by the Research Council of Norway through its Centers of Excellence funding scheme, Project number 223 265; and partly by the Center for Language and Cognition Groningen (CLCG), PhD project *Chunks in L2 development: A usage-based perspective*.

Hana Gustafsson, University of Oslo

Marjolijn Verspoor, University of Groningen

From a usage-based perspective, L2 development is primarily driven by frequency and salience of structures in the surrounding input (Ellis and Cadierno 2009). In theory, this means that sufficient exposure to authentic input in the target language – where chunks are very frequent – will drive the learning of chunks of the target language. But even when immersed in authentic input, L2 learners' attention tends to be focused on individual words rather than on word combinations (Wray 2002: 6). Individual words such as *school* are learned separately and earlier than two-word expressions such as *at school*, because for L2 learners single words tend to be more salient as units than multi-word expressions (chunks). Similarly, “preferred ways of saying things” may not be sufficiently salient as a unit in order to be learned as a chunk. The sensitivity to how words go together in authentic language is crucial for the learning of chunks and is most likely heightened by more exposure to authentic input (Ellis and Cadierno 2009). There are numerous individual differences in how L2 learners learn and use chunks (Schmitt and Carter 2004), which is also in line with the usage-based perspective assumption that individual learners construct their L2 differently depending on their exact experience of the L2. Moreover, as Verspoor and Smiskova (2012) maintain, L2 learners may also show great variability in their development of chunks over time. In this study, individual differences will be considered in relation to the amount and kind of input L2 learners are receiving and in relation to individual variability in development over time.

In short, learning chunks is an important part of L2 development and tracking learners' use of chunks can give us valuable insight into the process of L2 development. In this paper, we will report on a data driven study in a group of 22 Dutch L2 English learners; 11 in a semi-immersion instructional (high-input) condition and 11 in regular instructional (low-input) condition. Our main aim is to track the development of chunks in these learners, with a close focus on the potential differences between the high- and low-input conditions. Taking a usage-based perspective, we expect that the learners' use of chunks will reflect the amount and kind of input they are receiving. We expect the learners to use chunks that are generally frequent and/or salient in L2 English and chunks that are frequent and/or salient in their specific L2 environment. The high-input learners are expected to use more chunks and a greater range of chunk types than the low-input learners. To gain insight into the actual developmental process of using chunks over time, we zoom in on two learners selected from our group and conduct a case study from a dynamic usage-based perspective (Verspoor and Smiskova 2012).

Since chunks are a complex linguistic phenomenon and our study of chunks in L2 development is data driven, we will first explain our operationalization of chunks and give a detailed description of the method employed in our study.

Next, we will present and discuss the findings of our longitudinal study of chunk development. With our findings we hope to show that chunks, as we define them, develop differently in high- and low-input learners and that tracking learners' use of chunks can give us important insights into L2 development from a usage-based perspective.

2 Operationalizing chunks

In order to track the development of chunks in learner data we need an exact definition of a *chunk*. However, as a linguistic phenomenon, chunks are notoriously difficult to capture. The exact operationalization depends on the purpose of the study and the nature of the data in question (Wray 2002). Since this study investigates the development of native-like chunks in learner language from a usage-based perspective, we need a highly inclusive definition. In order to build this definition, we followed a cyclical process of definition and identification of chunks in our data (Wray 2002: 19). First, we consulted the learners' writings to see what types of chunks they contain; next, based on this exploratory stage we built a general definition of a chunk; and finally, this data-derived definition then served to identify more instances of chunks in the data. This is a dynamic approach that fits well with a usage-based perspective and our data-driven study.

The exploratory stage revealed that the learners in our study used not only multi-word expressions which are part of traditional phraseology (such as *human body*; *sick and tired*; *boss around*) but also expressions that are the preferred ways of saying certain things (see Smiskova, Verspoor, and Lowie 2012). Compare *when I grow up* and *when I am a grown up adult* in the following two excerpts from learner texts:

- (1) *When I am a grown up adult i would like to be a neurosurgeon. I would like to be a neurosurgeon when i am a grown up because i really like biology and i think i am quite good at it. I also think the human body is very interesting. It's so special how everything is organised so well and that most of the time works.*
- (2) *I seriously have no idea what kind of job I'd like to do. And I'm getting sick and tired of people who ask me that. Just because of that I'm in eight grade, I'm probably supposed to know what I want to do when I grow up... I absolutely don't like it when people boss me around.*

Based on the exploratory stage, we defined a chunk as a conventionalized word sequence expressing a certain concept. The same concept could also be expressed by a word sequence which, while following rules of syntax correctly, is not a

preferred word combination; compare *at school* vs. *in school*; *do homework* vs. *make homework*; *when I grow up* vs. *when I am a grown up adult*. In other words, a chunk for the purposes of this study is defined as a combination of two or more orthographic words¹, which may also include variable slots, expressing an idea (concept) in a conventionalized way. Such definition is in line with Langacker's (2008) notion of units representing normal ways of saying things:

A substantial proportion of what is needed to speak a language fluently tends to be ignored because it is part of neither lexicon nor grammar as these are traditionally conceived. What I have in mind are the countless units representing normal ways of saying things. Native speakers control an immense inventory of conventional expressions and patterns of expression enabling them to handle a continuous flow of rapid speech. While they can certainly be included, I am not referring to lexical items of the sort found in dictionaries, nor even to recognized idioms. At issue instead are particular ways of phrasing certain notions out of all the ways they could in principle be expressed in accordance with lexicon and grammar of the language. These units can be of any size, ranging from standard collocations to large chunks of boilerplate language. These can be fully specific or partially schematic, allowing options in certain positions. (Langacker 2008: 84)

In order to distinguish between different types of chunks identified in the learner data, our aim was to classify the identified chunks in established typological categories. In fact, we found that using a typology of established chunk types was also helpful in the initial identification stage. Based on our definition, we were aiming for a typology of chunks which would include both established phraseological chunk types and “preferred ways of saying things”. Since this is precisely what Granger and Paquot (2008) advocate, we chose to follow their classification approach. They propose the integration of two major approaches: the traditional *phraseological approach*, which is based on linguistic analysis, strictly distinguishes between phraseology and syntax and is mostly concerned with specific phraseological categories, most often non-compositional (idiomatic, semantically opaque, such as *kick the bucket*) and/or syntactically irregular (such as *by and large*). The second is a *corpus-based, frequency-distributional approach*, which has produced extensive evidence of frequently occurring semantically and syntactically regular word combinations, which are not part of traditional phraseology.

The formal typology of phrasemes proposed by Granger and Paquot (2008: 43–44) served as basis for the typology of chunks we used in this study (See Table 1 in Appendix). The original typology has three functional categories: *referential* (chunks which refer only to content and have no pragmatic function),

¹ Compounds which are hyphenated or written together as an orthographic word are also included (e.g. *however*).

textual (chunks with a discourse structuring and organizational function) and *communicative* (chunks with a communicative function, e.g. addressing interlocutors). These functional categories are then subdivided on the basis of structural types and degrees of non-compositionality (idiomaticity). In our study, we further grouped these into organizational levels (phrase or sentence level). Finally, we included “preferred ways of saying things” (namely, conventionalized sentences and conventionalized sentence stems), and additional categories based on the exploratory analysis of our data: structures, variable idioms (Stefanowitsch and Gries 2003: 43) and constructions. Table 1 in Appendix shows the modified typology including examples from our data.

However, when using any typology for a fine categorization of chunks it is important to bear in mind that we are imposing a fixed structure on a fuzzy, emergent phenomenon that may not have entirely fixed categories. The defining characteristics of chunks – such as non-compositionality, fixedness, function – tend to be present to a certain degree. This means that each formal category defined by these characteristics has prototypical examples, but there may also be word sequences in that category that do not display the defining characteristics to the same degree. Compare for instance the varying degrees of non-compositionality (idiomaticity) in *blow a fuse* – *blow your own trumpet* – *blow the gaff* (Granger and Paquot 2008). This is due to the fact that these defining characteristics form a continuum rather than clear-cut categories; and this in turn is caused by the overlap between morphology, syntax, semantics and discourse (Granger and Paquot 2008: 37). Moreover, many chunks are nested, i.e. consisting of smaller, often overlapping chunks (Wray 2002: 28), such as in *[[[The only thing] I [[know [for sure]]], which can complicate the process of counting and classifying chunks.*

Therefore, the typology of chunks presented here is intended as a helpful inventory of prototypes that may be further refined on the basis of more data analysis. Its categories should be perceived as dynamic and open rather than fixed and mutually exclusive; for instance, a chunk can display the defining characteristics of several categories (e.g. *however* is a linking adverbial in its function as well as a compound in its structure).

In resolving these and other complexities involved in researching chunks, it is crucial to closely adhere to the aim and background of the study (Wray 2002: 28). In our case, this involved making decisions in line with the usage-based perspective in order to tease out differences in development between our high- and low-input learners. In the following section, we describe in detail how chunks were identified and classified in this study and how we ensured consistency of the process.

3 The study

This paper reports on two longitudinal studies investigating the development of chunks in learners' language over about 2.5 years in high- and low-input conditions. The first is a group study, in which two groups of learners in two conditions (high- and low-input) are compared at the beginning and the end of the study. The second is a case study, in which the development of chunks is traced of two selected learners across 12 data points over time.

3.1 Input and production frequencies

Our participants are 22 Dutch high school learners who attend the same Dutch school, have a similar socio-economic background, and a similar scholastic aptitude (Verspoor et al. 2010). Both groups have an interest in language as the high-input group opted for a bilingual Dutch-English stream and the low-input group for monolingual (Dutch) stream that includes classical languages. At the start of the study the learners were about 11 years old in their first year at high school. The high-input group attended an education program in which subjects, such as geography and history, were taught in English. This group also had five hours of English as subject taught by a native speaker of English. In the high-input condition, students were exposed to a great amount of spoken English by both non-native and native speakers of English; most of their course materials were the same as used by native speakers. The low-input group attended a Dutch regular education program with all subjects taught in Dutch. This group had 2 hours of English a week also taught by a native speaker of English, but the course materials, even though communicative contained quite a bit of explicit grammar instruction. This group also had 2 hours in Latin and Greek.

The learners were asked to write about once a month on informal topics such as *My new school* or *My vacation*; *Write about the rules at home*; *Do you think they are fair?*; *What do you want to be when you grow up?*; *Write about a film or a book you like*. Most of the writings were written directly on computer, where the word limit was 200 words; some were handwritten in class. There was no time limit for the writings, but the students usually wrote no longer than 10 minutes.

3.2 Identification of chunks in written texts

First, we used researcher intuition to identify multi-word expressions which could match our general definition of a chunk. To help validate such intuitive decisions,

we used Wray's (2008: 113–127) list of diagnostic criteria. Following Wray's instructions (2008: 115), we only used the criteria to validate expressions we had already intuitively identified as chunks, rather than using the criteria to initially identify chunks. If needed, which was mostly in the case of “preferred ways of saying things”, intuitive judgments were further validated with the help of reference corpora. The reference corpora were used to confirm that a multi-word expression we intuitively identified as a chunk indeed frequently occurred in the corpora as a conventionalized unit. This step was based on the underlying notion of frequency, but not on detailed frequency or mutual strength counts.² In principle, this step gives further support to the criterion in point H of Wray's list of diagnostic criteria: “By my judgement, based on direct evidence or my intuition, there is a greater than chance-level probability that the writer will have encountered this precise formulation before in communication from other people” (Wray 2008: 120).

In addition to standardized reference corpora, such as the British National Corpus (BNC) (Davies 2004) and the Corpus of Contemporary American English (COCA) (Davies 2008), we also used the WebCorp search engine (Renouf, Kehoe, and Banerjee 2007), which allowed us to search the World Wide Web as a reference corpus. WebCorp is particularly relevant in our study since we are following young teenagers who tend to have a great deal of exposure to the language of the Internet; moreover, a number of conventionalized expressions that do not occur in standardized corpora are frequently found on the Internet (e.g. *The only thing I know for sure is that* +clause).

As already mentioned, larger chunks were often nested, i.e. they contained smaller, overlapping chunks. The larger chunks were mostly “preferred ways of saying things”, such as *the only thing I know for sure*, which includes a verb complement (*know*+clause) and a particle (*for sure*). In these cases, the smaller chunks were not counted separately, as the composite larger chunk was perceived as a separate form-meaning/function unit in itself.

Finally, all identified chunks were categorized for type following our typology (Table 1 in Appendix). In the cases of chunks which display the defining characteristics of several categories, functional categories were given preference, since they are the more fundamental distinction in our typology. For instance, *however*, which is a compound in its structure as well as a linking adverbial in its function, was classified as a linking adverbial rather than a compound.

To ensure consistency of the identification process, all intuitive judgments were made by the first author, discussed and fine-tuned with the second author

² We were unable to use word association measures as at the time of our study Webcorp did not provide this type of statistical information (Renouf et al. 2007: 53).

and validated as described in this section. To ensure all chunks were categorized consistently, the first author went through the coding process twice.

3.3 Method of analysis

In the cross-sectional study we used the procedure described in the previous section to identify chunks in texts written by the high-input learners ($N = 11$) and low-input learners ($N = 11$) at the start of the study (October 2007) and towards the end of the study (May 2009).³ Next, we established several measures of chunk use in order to tease out differences between the groups and within the groups over time.

First, we recorded the raw token frequencies of all chunk types in Oct 07 and May 09 to see which chunk types were most frequently used in each group and if their distribution changed over time. To discover variety in chunk use, we counted the number of different chunk types per text (chunk types/text). To be able to compare learners and groups, we calculated relative frequencies by taking text length into account: for relative token frequency we calculated the ratio of all chunk tokens per 100 words of text (chunks/100 words); for relative token frequency of each chunk type we calculated the type-token ratio of each chunk type per 100 words (type-token [chunk type]/100 words). To gain more insight into the use of chunks in relation to text length, we calculated the correlation between the length of each text and the number of all chunk tokens in it (correlation text length/chunks). To measure overall “chunk coverage” we calculated the percentage of all words used as part of a chunk in each text (%chunk-words/text). Finally, to capture differences in the development of chunk length we calculated the mean chunk length per text (mean chunk length/text).

Next, we performed statistical analyses on all these measures to see where there might be significant differences in development, both between the groups and within the groups over time (potential change from Oct 07 to May 09).

In the case study we used the same procedure for the identification of chunks texts written by two selected high-input and low-input learners. For each measurement over time (i.e. each collected text) we recorded the number of different chunk types and the raw token frequency of each chunk type. This way we obtained a longitudinal chunk profile for each learner, which allowed us to track their individual development in detail over two years (Oct 07 to Nov 09).

³ Due to subject dropout we had to take May 09 as the end-point of our cross-sectional group study; the data for our microgenetic study was available until November 09.

4 Results group study

The results show that in Oct 07 the low-input group wrote on average significantly longer texts than the high-input group (low-input mean 114.18, high-input mean 65.09 words, $p < .05$) and in May 09 the high-input group wrote on average significantly longer texts than the low-input group (high-input mean 157.72, low-input mean 103 words, $p < .05$). We will first present the raw token frequencies of all chunk types identified in the texts and then the relative frequencies according to each measure.

4.1 Raw frequencies of chunk tokens and types

In total, 18 chunk types⁴ were identified in all the texts together. Figure 1 shows raw token frequencies of each chunk type at the beginning of the study in Oct 07 and Figure 2 at the end of the study in May 09.

Figure 1: October 2007: Raw token frequency of each chunk type in high- and low-input group.

⁴ Not all of the 22 categories of our typology were expected to appear in the data, which indeed they did not (such as slogans, proverbs, similes and complex textual organizers).

Figure 1 shows that the most frequent chunks at the beginning of our study were lexical collocations (e.g. *strong coffee*; *main character*; *first kiss*) followed by particles (e.g. *a lot of*; *at home*; *in English*), structures (e.g. *NUMBER years old*), compounds (e.g. *living room*), verb complements (e.g. *would like to*; *going to*; *have to*; *like -ing*; *think + clause*) and conventionalized sentences (*It's hard to explain*).

Figure 2 shows that at the end of the study the frequency distribution was quite different: verb complements had moved from their fifth place to become the most frequent chunk type; conventionalized sentence stems (e.g. *The only thing I know for sure is that+clause*) and attitudinal formulae (e.g. *I mean*) had moved into the top five most frequent chunk types, and structures moved down from their third place to become much less frequent.

Figure 2: May 2009: Raw token frequency of each chunk type in high- and low-input group.

4.2 Relative frequencies of chunk tokens and types⁵

The relative chunk token frequency: In Oct 07 the high-input group used on average significantly more chunks per 100 words (mean 10.62, $p < .05$) than the low-input group (mean 4.97). Repeated measures ANOVA shows that over time, the measure increased significantly more in the low-input group ($F(1, 20) = 5$, $p < .05$) than in the high-input group. Figure 3 shows the time/group interaction chart. Post hoc pair-wise comparison (paired samples t-test) shows that there was a significant increase between Oct 07 and May 09 in both low-input ($p < .001$) and high input group ($p = .04$). In May 09 there were no significant differences between the groups in the average number of chunks per 100 words ($p > .05$).

Figure 3: Increase over time in chunks/100 words (time/group interaction).

The relative chunk type frequency: In Oct 07 there was no significant difference between the groups in the number of chunk types. Repeated measures ANOVA shows that over time, the number of chunk types increased significantly more in the high-input group ($F(1, 20) = 5$, $p < .05$) than in the low-input group. Figure 4 shows the time/group interaction chart. Post hoc pair-wise comparison

⁵ For detailed results of statistical analyses see Table 2 in Appendix.

(paired samples t-test) shows that the increase between Oct 07 and May 09 was significant in the high-input group ($p = .001$) but not in the low-input group ($p = .07$). In May 09, the high-input group used significantly more chunk types per text (mean 7.0, $p < .001$) than the low-input group (mean 4.55).

Figure 4: Increase over time in the number of chunk types (time/group interaction).

The relative token frequency of specific chunk types (ratio [chunk type]/100 words): Repeated measures ANOVA shows a time/group interaction in the ratio of two chunk types: verb complements and conventionalized sentence stems.

The ratio of verb complements increased significantly more in the low-input group ($F(1, 20) = 5$; $p < .05$) than the high-input group. Figure 5 shows the time/group interaction chart. Post hoc pair-wise comparison (paired samples t-test) shows that the increase in the verb complements ratio between Oct 07 and May 09 was significant in both low-input ($p = .001$) and high-input group ($p = .02$).

The ratio of conventionalized sentence stems increased significantly more in the high-input group ($F(1, 20) = 10$; $p < .05$) than in the low-input group. Figure 6 shows the time/group interaction chart. Post hoc pair-wise comparison (paired samples t-test) shows that the increase in the ratio of conventionalized sentence stems between Oct 07 and May 09 was significant in the high-input group ($p < .001$) but not in the low-input group ($p = .052$).

Figure 5: Increase over time in the ratio of verb complements (time/group interaction).

Figure 6: Increase over time in the ratio of conventionalized sentence stems (time/group interaction).

Correlation between text length and the number of all chunk tokens: In both Oct 07 and May 09, the high-input group shows a moderate to strong significant correlation between text length and the number of chunks in it (Oct: $r = .61$, $p < .05$; May: $r = .87$, $p < .001$); for the low-input group, this is the case only at the end of the study in May 09 ($r = .73$, $p < .05$).

Mean chunk length: In Oct 07 there was no significant difference in the mean chunk length/text between the two groups. Over time, mean chunk length/text increased significantly in both groups (repeated measures ANOVA: $F(1, 20) = 74$, $p < .001$, post hoc pair-wise comparison using paired t-test: high-input $p < .001$, low-input $p < .001$). In May 09, the high-input group used significantly longer chunks (mean 3.13 words, $p < .05$) than the low-input group (mean 2.60 words).

Proportion of chunk-words: In Oct 07, the high-input group had a significantly higher percentage of chunk-words per text (mean 26.8%, $p < .05$) than the low-input group (mean 12.4%). Over time, the proportion of chunks increased significantly in both groups ($F(1, 20) = 34$, $p < .001$; post hoc pair-wise comparison using paired t-test: high-input $p = .008$, low-input $p < .001$). In May 09, the high-input group had a significantly higher percentage of chunk-words per text (mean 45.6%, $p < .05$) than the low-input group (mean 33.2%).

5 Results case study

Two learners were selected as representative of their groups (high- and low-input), in the sense that their development matched most closely the chunk development at the level of their groups. Both learners were judged to be at the same starting level by a panel of independent judges. In this section we report on the results of the case study and in the Discussion section we explore the learners' development qualitatively from a dynamic usage-based perspective. We also relate the results of the case study to the different input conditions and to the results of the group study.

5.1 Low-input learner

The longitudinal profile in Figure 7 shows the different chunk types identified in texts written by the low-input learner (texts collected from Sept 07 to Nov 09). Throughout the study, grammatical collocations such as particles (e.g. *at home*) and verb complements (e.g. *like -ing*) are the most frequent chunk types; in some texts these grammatical collocations are overused (see Figure 7, May 09). The other two most frequently occurring chunk types are lexical collocations and

compounds. After 2.5 years the learner used six chunk types compared to three chunk types at the start of the study. The correlation between the number of chunks and the length of the texts collected over time is rather weak and non-significant ($r = 0.40$, $p > .05$).

Figure 7: Longitudinal chunk profile in a low-input learner (raw token frequency). Numbers and letters in the bar chart indicate raw token frequencies of individual chunk types: a = compounds, b = lexical collocations, c = particles, d = complements, e = phrasal verbs, f = idioms, i = structures, k = constructions, l = conventionalized sentence stems, m = conventionalized sentences (chunk types are labeled alphabetically following Table 1. in Appendix).

5.2 High-input learner

Figure 8 illustrates the range of chunk types used by the high-input learner over time. Grammatical collocations, such as particles (e.g. *at home*) and verb complements (e.g. *like -ing*) are the most frequent, followed by lexical collocations (e.g. *strong coffee*), conventionalized sentence stems (e.g. *One thing I know for sure is that +clause*) and structures (e.g. **number years old**). After 2.5 years the learner used 10 chunk types compared to 1 chunk type at the start of the study. The chunk range includes types with a communicative and discourse organization function (e.g. *I mean; you know; by the way*). There is a strong, significant correlation between the number of chunks and the length of the texts collected over time ($r = .95$, $p < .001$).

Figure 8: Longitudinal chunk profile in a high-input learner. Numbers and letters in the bar chart indicate raw token frequencies of individual chunk types: a = compounds, b = lexical collocations, c = particles, d = complements, e = phrasal verbs, f = idioms, i = structures, j = variable idioms, l = conventionalized sentence stems, m = conventionalized sentences, n = textual prepositions, o = textual conjunctions, p = textual adverbials, r = speech act formulae, s = attitudinal formulae (chunk types are labeled alphabetically following Table 1. in Appendix).

6 Discussion

6.1 Group study

The chunk measures we used in this study have helped tease out a number of significant tendencies in the development of the two groups. The results show that in both groups there is growth over time in all chunk measures. This is not surprising: since chunks are pervasive in language, both groups learn chunks as they learn other aspects of L2 English. However, there are differences in how much some of these measures increased in each group. We will discuss how these differences could be related to the amount and kind of input the learners were receiving. We will also discuss how the results of these measures are related and which measures seem to distinguish best the differences in develop-

ment between the high- and low-input learners. We will also illustrate the group tendencies with the results of the case study.

6.1.1 Raw frequencies

The raw frequency counts show that out of the 22 types established in our typology, 18 types were identified in the learner data, with considerable differences in frequency distribution between the different types (Figures 1 and 2 in the Results section). In general, the overall frequency distribution of chunk types appearing in the data at the start of the study was different than at the end of the study. In October 07 the most frequent chunk types were lexical collocations, particles, structures (short slot-fillers), compounds and complements. Most of these chunk types are short and/or grammatically based and tend to be very frequent in L2 English; they are also the ones most often encountered as part of language instruction. The use of these chunks contributes to L2 accuracy but not as much to general fluency and authenticity. Also, in Oct 07 there were only very few “preferred ways of saying things”, discourse organizers, communicative formulae and idiomatic chunks – the kind of chunks that greatly contribute to fluency and authenticity and are typical of an authentic native-like repertoire. Such findings indicate rather low levels of L2 proficiency at the start of the study.

As the overall use of chunks increased over time and the learners started using new chunk types, the frequency distribution of different chunk types changed. While the short and/or grammatically based chunks were still very frequent, the order of frequency changed as “preferred ways of saying things”, discourse organizers and communicative formulae were more frequently used: in May 09, the most frequent were complements, lexical collocations, particles, conventionalized sentence stems, attitudinal formulae, compounds and conventionalized sentences. Such findings show that as the learners were becoming more proficient and had more contact with the language, they started using the longer chunks that greatly contribute to L2 fluency and authenticity. However, the results also show that some aspects of the frequency distribution and of the changes over time were not the same for both input groups.

6.1.2 Relative frequencies

The relative frequencies show that at the start (Oct 07) the high-input group outperformed the low-input group on two chunk measures: %chunk-words/text and

chunks/100 words. This means that in Oct 07 the high-input group already used significantly more chunks, which could be related to the two months of semi-immersion in authentic input the group had already received at the time of the first writing (as opposed to the low-input group, who had received only 2 hours of English instruction weekly). At that point the high-input group was not using significantly more chunk types than the low-input group.

As the learners were becoming more proficient over time and started using more chunks, the relative measures grew in both input groups; however, there was a significant difference in the growth of two of these measures: the number of chunk types increased significantly more in the high-input group and chunks/100 words increased significantly more in the low-input group. In line with our expectations, the high-input learners developed a greater range of chunk types, since they had more exposure to authentic input. The significantly greater increase in chunks/100 words in the low-input group most likely indicates that there was a slight difference in L2 proficiency between the low-input and high-input learners at the time of the first writing. The progress in the low-input group was then more rapid because they started from lower (even zero) levels than the high-input group, who had already established a certain level of the L2.

The relative frequencies show that at the end of the study (May 09) the high-input group again outperformed the low-input group, this time on three chunk measures. Firstly, the high-input group had a significantly higher proportion of chunks in their texts than the low-input group (% chunkwords/text). In fact, the high-input group had approached native-like proportion of chunks in their texts: on average, 46% of the learners' texts consisted of fixed chunks – quite close to Erman and Warren's (2000) estimation that about 50% of native speaker text consists of chunks; in the low-input group the average percentage was 33%.

Secondly, in May 09 the high-input group used significantly longer chunks (mean chunk length/text) than the low-input group. This contributed to the increased percentage of chunk-words in text: since longer chunks contain more words, the percentage of chunk-words per text is higher. From a usage-based perspective, it is interesting to look at *why* the mean chunk length increased. The increased chunk length reflects our findings about the increase over time in certain chunk types: the high-input group increased their ratio of conventionalized sentence stems (e.g. *The only thing I know for sure is that* +clause) significantly more than the low-input group. In addition, the high-input group developed other types of longer chunks, such as textual sentence stems, attitudinal formulae and conventionalized sentences (e.g. *There are more important things in life*). It was mainly the increase in these “preferred ways of saying things” what contributed to the overall increase in chunk length.

What is also interesting about the longer chunks is that they are nested, i.e. consisting of smaller, often overlapping chunks. For instance, consider the example sentence in (3), which contains a conventionalized sentence stem *The only thing I know for sure*+clause. The stem consists of several overlapping word sequences, which in our study could all be considered chunks. As units, they all occur quite frequently in native-like use of English (for illustration, the corresponding numbers show token frequency per 1 million words in COCA), with the frequency of occurrence decreasing with the increased length of the sequence:

- (3) *The only thing I know for sure is that I want to travel.*
<i>for sure</i>	5706
<i>know for sure</i>	1161
<i>the only thing</i>	7796
<i>the only thing I know</i>	52
<i>the only thing I know for sure</i>	4

While all these sequences could in principle be also counted separately, in our study we only take into account the original 7-word sequence: it is a conventionalized way of expressing a certain notion (a normal way of saying this), which in its entirety can be regarded as a form/meaning/function mapping in Langacker's terms (2008: 84). In a study of L2 development, the ability to use such expression seems a sign of fluency, authenticity and a certain proficiency in the L2. It also shows that the learner is paying attention to how things are normally said, or, to how individual words are combined into authentically sounding clauses and sentences, and finally into a fluent, native-like discourse (Ellis 2001; Pawley and Syder 1983). Since the use of these longer chunks is prominent in the high-input group, we can conclude that a) frequent exposure to authentic input facilitates the learning of longer chunks, b) that high-input learners develop sensitivity for "normal ways of saying things" and that, possibly, c) high-input learners are able to manipulate the L2 in larger form/function/meaning units and/or are able to recall longer word sequences.

Finally, in May 09 learners in the high-input group used significantly more chunk types than the low-input group. This reflects the already mentioned use of discourse organizers, communicative formulae and above all "preferred ways of saying things" and is again in line with our expectations: more authentic input means a wider range of chunk types; more exposure to authentic input means greater opportunity to learn a wider range of chunk types.

However, in May 09 the high-input group also wrote significantly longer texts, which could also influence their scores on some of these measures: longer texts increase the chance that more chunk types are used, and longer texts will

logically contain more chunk tokens. Still, this is not entirely the case in our data: the length of a text and the number of chunks in it do not always correlate highly. Firstly, in Oct 07 the low-input group wrote significantly longer texts by almost a half than the high-input group; still, the high-input group used significantly more chunks. Secondly, the high-input group shows a strong significant correlation between text length and the number of chunks in it both at the start and at the end of the study; for the low-input group, this was only the case at the end of the study. These findings seem to suggest a different type of development in the two input groups: in native speaker repertoire we would also expect text length and number of chunks to correlate – since language is full of chunks, longer text will contain more chunk tokens. This finding could be a sign of more native-like development in the high-input group.

6.1.3 Individual differences

Our cross-sectional study has also revealed individual differences in the use of chunks, which we have noted not only in the statistical analyses (high standard deviations in almost all measures) but also in qualitative analyses of the data. Some of the chunks our learners use are not what one would expect in relation to the assigned topic, nor are they the kind that frequently appears in everyday language, in standardized corpora or in a classroom setting. Still, they are clearly chunks characteristic of the kind of input the learners are regularly exposed to. In an answer to a topic question (*What would you like to be when you grow up?*) two learners responded in very different ways – and used very different chunks. In example (4), we can detect the jargon of computer games, while in (5) we could guess at movies, TV shows and glossy magazines:

- (4) *when i grow up i want to be a **game designer**. then i can make all kinds of games. i already use **game maker 7** where you can **program your own pictures**, you draw them with paint and then you can **let them appear or disappear** and you can **let them fire something or change into something else**. you can **make a sort of super mario** because you can **set the gravity and the jumping height**, but you can also **make a sort of pack man**, you just make a labyrinth and you **let your picture do what you want when you press certain buttons** and you **make other pictures follow you**. then you say that when a **collision** happens **between you and the guy**, you **delete yourself**. there are also programs that are used for making **3D pictures** and where you can **scan your drawings** to paste them on a way that you can **easily make them 3D**.*

- (5) *When I grow up, I want to be famous, I want to be a star, I want to be in movies. When I grow up, I want to see the world, I want to drive nice cars, I want to have groupies. When I grow up, I want to be on TV, that people know me, be in magazines. When I grow older than that, I want to have a nice, rich, husband that pays everything for me. I want to do nothing at all and just lay around our pool all day long. When I grow even older I want to have children and then I hope I will be a good mom. I will make sure they get everything they need, but I will not spoil them. I don't want to have spoiled children, but I do want them to have a good life. That is what I want to be when I grow up.*

In fact, the first few lines of learner text (5) almost exactly correspond with the chorus in a popular song which was traced back via WebCorp: *When I grow up / I wanna be famous / I wanna be a star / I wanna be in movies / When I grow up / I wanna see the world / Drive nice cars / I wanna have groupies / When I grow up / Be on TV / People know me / Be on magazines* (The Pussycat Dolls, LLC 2008: When I grow up). This is valuable evidence of the influence of input on the learner's use of L2 English.

In short, there are clear individual differences in what chunks learners use and how they use them, which can be related to different kinds of input they are exposed to. This is compatible with the usage-based assumption that everyone's experience of language is slightly different: what may be a completely unknown, highly idiomatic and perhaps a rather useless expression for one person (e.g. *super mario, pack man, set the gravity and the jumping height*) may be a frequently encountered and relevant expression for another. Compared to native speakers our Dutch L2 learners receive limited amounts of English input, and each learner may be surrounded by different kinds of input and developing their language in different ways. As a result, the developmental paths of chunks may be very different for each learner. The texts collected during our study provide evidence: some are more dense with chunks than others and the different chunk types are unevenly spread across the texts written by individual learners.

6.2 Case study

The longitudinal chunk profiles of the high and long-input learner (Figures 7 and 8 in the Results section) suggest qualitative differences in development. The high-input learner uses a range of different chunk types and shows considerable variability in the frequency distribution of chunk types over time, while in the low-input learner this variability is much less pronounced. At the

start of the study the low-input learner used more chunks than the high-input learner (compare Figures 7 and 8, Sept 07), but she seems to develop differently and does not experience the same rapid explosion of a wide range of chunk types as the high-input learner. The qualitative differences interpreted here at face value were further analyzed in Verspoor and Smiskova (2012), who argue that from a dynamic usage-based perspective, this difference in variability over time is meaningful and relevant to the learners' L2 development.

The profiles of the two learners also show different patterns in the development of specific chunk types. In the high-input learner, there are interesting relationships over time in the use of lexical collocations, grammatical collocations and "preferred ways of saying things". In the first half of the study she tended to overuse particles and verb complements, while using only a few lexical collocations; in the second half of the study her use of lexical collocations rapidly increased, she started using "preferred ways of saying things" and fewer grammatical collocations. This decrease could be explained by nesting: as the learner starts using longer and more complex conventionalized word sequences, shorter grammatical chunks embedded in them are no longer free-standing and consequently are not counted as such. For the low-input learner, grammatical collocations (verb complements and particles) remain the most frequent types throughout the study, the increase in lexical collocations is much less pronounced than in the high-input learner, and there is no clear emergence of "preferred ways of saying things".

The differing patterns of development can be related to the learners' different input conditions and illustrate in greater detail some of the significant effects identified on group level. The low-input learner, who had less exposure to authentic input, shows a rather limited development of different chunk types (from 3 chunk types to 6), while the high-input learner developed a whole range of different types (from 1 to 10). Next, while the low-input learner uses only very few "preferred ways of saying things", there is a clear emergence of this chunk type in the high-input learner: after 2.5 years, conventionalized sentence stems are the third most frequent chunk type in her text. The rapid increase in lexical collocations in the high-input learner matches the increase in raw token frequency on the group level (although this overall increase in the high-input group was not captured by the ratio type-token colloc/100 words). The high-input learner also uses communicative formulae and discourse organizers (textual and referential chunk types), both of which are an integral part of native-like discourse, while in the low-input learner these chunk types are missing. Finally, while the high-input learner shows a very strong significant correlation between the text length and the number of chunks, the low-input learner only shows a moderate

trend, which is not significant. Just as on the group level, these findings seem to indicate a faster and more native-like development in the high-input learner.

7 Conclusion

In this study we explored the development of chunks from a usage-based perspective in two groups of Dutch L2 learners of English, one in a high- and one in a low-input condition. Using an integrated approach we have captured some significant differences in the development of each group over time. Our findings show that over time both groups used increasingly more chunks and developed a greater range of chunk types, among which were traditionally recognized types and so called “normal ways of saying things” (Langacker 2008); both groups also used increasingly longer chunks. Moreover, in line with our usage-based expectations, our findings also show significant differences between the two input groups. The high-input learners, who had more exposure to authentic input, developed a significantly greater range of chunk types including those with a clear discourse function, greater number of “normal ways of saying things” and a greater proportion of chunk-words per text. At the end of the study, this proportion was on average 46% – fairly close to Erman and Warren’s (2000) estimate of about a 50% proportion of chunks in native speaker text. Such findings lead to interesting conclusions that would benefit from further usage-based research. The significant differences we identified between the two groups are related to recognized features of a fluent, authentic native speaker repertoire: high chunk density, wide range of chunk types of various structure and function, longer and nested chunks. This suggests that the development of chunks in our high-input learners is more native-like than the development in our low-input learners; as a result, the high-input learners are also using their L2 in a more fluent and authentic way. Finally, we have also identified clear individual differences in what chunks the learners are using and how they are using them – this seems to be not only influenced by the amount and kind of input they are receiving, but also by their individual communicative needs.

References

- Davies, Mark. 2004. BYU-BNC: The British National Corpus. Available online at <http://corpus.byu.edu/bnc>.BYU-BNC [On-line].

- Davies, Mark. 2008. The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. Available online at <http://www.americancorpus.org>. COCA [On-line].
- Ellis, Nick C. 2001. Memory for language. In *Cognition and Second Language Instruction*, Peter Robinson (ed.), 33–68. Cambridge: Cambridge University Press.
- Ellis, Nick C. 2008. Phraseology: The periphery and the heart of language. In *Phraseology in Foreign Language Learning and Teaching*, Fanny Meunier, and Sylviane Granger (eds.), 1–13. Amsterdam: John Benjamins.
- Ellis, Nick C., and Teresa Cadierno. 2009. Constructing a second language: Introduction to the special edition. *Annual Review of Cognitive Linguistics* 7: 111–139.
- Erman, Britt, and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20 (1): 29–62.
- Eyckmans, June, Frank Boers, and Helene Stengers. 2007. Identifying chunks: Who can see the wood for the trees? *Language Forum* 33 (2): 85–100.
- Granger, Sylviane, and Magali Paquot. 2008. Disentangling the phraseological web. In *Phraseology: An Interdisciplinary Perspective*, Sylviane Granger, and Fanny Meunier (eds.), 28–49. Amsterdam: John Benjamins.
- Langacker, Ronald W. 2008. Cognitive Grammar as a basis for language instruction. In *Handbook of Cognitive Linguistics and Second Language Acquisition*, Peter Robinson, and Nick C. Ellis (eds.), 66–88. New York: Routledge.
- Pawley, Andrew, and Frances H. Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In *Language and Communication*, Jack C. Richards, and Richard W. Schmidt (eds.), 163–199. New York: Longman.
- Renouf, Antoinette, Andrew Kehoe, and Jay Banerjee. 2007. WebCorp: An integrated system for web text search. In *Corpus Linguistics and the Web*, Nadja Nesselhauf, Marianne Hundt, and Carolin Biewer (eds.), 47–68. Amsterdam: Rodopi.
- Schmitt, Norbert, and Ronald Carter. 2004. Formulaic Sequences in Action: An Introduction. In *Formulaic Sequences: Acquisition, Processing and Use*, Norbert Schmitt (ed.), 1–22. Amsterdam/Philadelphia: John Benjamins.
- Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Smiskova, Hana, Marjolijn Verspoor, and Wander Lowie. 2012. Conventionalized ways of saying things (CWOSTs) and L2 development. *Dutch Journal of Applied Linguistics* 1 (1): 125–142.
- Stefanowitsch, Anatol, and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8: 209–243.
- Verspoor, Marjolijn, and Hana Smiskova. 2012. Foreign Language Development from a Dynamic Usage-Based Perspective. In *L2 writing development: Multiple perspectives*, Rosa M. Manchón (ed.), 17–46. Berlin: Mouton de Gruyter.
- Verspoor, Marjolijn, Monika S. Schmid, and Xiaoyan Xu. 2012. A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing* 21 (3): 239–263.
- Verspoor, Marjolijn, Jennifer Schuitemaker-King, Eva van Rein, Kees de Bot, and Peter Edelenbos. 2010. Tweetalig onderwijs: Vormgeving en prestaties. Onderzoeksrapportage.
- Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.
- Wray, Alison. 2008. *Formulaic Language: Pushing the Boundaries*. Cambridge: Cambridge University Press.

Appendix

Table 1: Typology of chunks (closely based on Granger and Paquot 2008: 43–44).

1. Chunk types with a referential function

1.1 Word and phrase level

a. Compounds	<p><i>sunbathing, two-week holiday, ice-cream</i></p> <p>– made up of two elements which have independent status outside these word combinations. They can be written separately, with a hyphen or as one orthographic word.</p>
b. Lexical collocations	<p><i>heavy rain, take a dive, strong current, pretty hard, real close, go wrong, hurt badly</i></p> <p>– strings of specific lexical items, which co-occur with a mutual expectancy greater than chance and have a semantic dependency relationship; they show compositional structure and do not have pragmatic functions. The “base” of a collocation is selected first by a language user for its independent meaning. The second element is semantically dependent on the “base”. One collocation can be embedded in another, as in (take [[strong] measures])</p>
c. Grammatical collocations: Particles	<p><i>afraid of, involved in, at school, in English</i></p> <p>– restricted combinations of a lexical word and a grammatical word</p>
d. Grammatical collocations: Complements	<p><i>avoid -ing; necessary to; want/going/have/ manage to; go -in; keep -ing; would like to; be able to; know+clause; say that+clause</i></p> <p>– restricted combinations of a lexical word and a complement structure (infinitive, gerund, reflexive pronoun or a nominal sentence)</p>
e. Phrasal verbs	<p><i>blow up, make out, crop up</i></p> <p>– combinations of verbs and adverbial particles; can have varying degrees of non-compositionality</p>
f. Idioms	<p><i>to spill the beans, to let the cat out of the bag, to bark up the wrong tree</i></p> <p>– constructed around a verbal nucleus and characterized by their semantic non-compositionality. Lack of flexibility and marked syntax are further indications of their idiomatic status</p>

- g. Similes *as old as the hills, to swear like a trooper*
– stereotyped comparisons; they typically consist of sequences following the frames “as ADJ as (DET) NOUN” and “VERB like a NOUN”
- h. Irreversible bi- and trinomials *bed and breakfast; kith and kin; left, right and center*
– fixed sequences of two or three word forms that belong to the same part-of-speech category and are linked by the conjunction *and* or *or*
- i. Structures Even ADJ+er than; as ADJ as, a year ago, two meters high
– short slot-fillers containing one or more free slots for a lexical item
- j. Variable idioms *pay a price for -ing; end up -ing*
– have a slot and are to some degree idiomatic (Stefanowitsch and Gries 2003: 222)

1.2 Clause and sentence level

- k. Constructions *The sooner we are finished, the sooner we can go*
– longer slot-fillers containing one or more free slots for a phrase or a clause
- l. Preferred ways of saying things: Conventionalized sentence stems *one thing I know for sure is...; all they can do is...*
– a clause or its fragment whose grammatical form and lexical content is wholly or largely fixed; its fixed elements form a standard label for a culturally recognized concept
- m. Preferred ways of saying things: Conventionalized sentences *It's hard to explain. There are more important things in life. I'm just who I am.*
– a sentence whose grammatical form and lexical content is wholly or largely fixed; its fixed elements form a standard label for a culturally recognized concept

2. Chunk types with a textual function

2.1 Word and phrase level

- n. Textual prepositions *in addition to, apart from*
– grammaticalized combinations of simple prepositions with a noun, adverb or adjective
- o. Textual conjunctions *so that, as if, even though*
– grammaticalized sequences that function as complex conjunctions

- p. Textual adverbials *in other words, last but not least, more accurately, what is more, to conclude, the reason for, however*
 – linking chunks such as polywords, grammaticalized prepositional phrases, adjectival phrases, adverbial phrases

2.2 Clause level

- q. Textual sentence stems *another thing is+clause; it will be shown that+clause; I will discuss+clause*
 – routinized fragments of sentences that are used to serve specific textual or organizational functions and typically involve a subject and a verb

3. Chunk types with a communicative function

- r. Speech act formulae *Good morning; Take care; You're welcome; suggesting (why don't we), concluding (that's all)*
 – preferred ways of performing certain functions such as greetings, compliments, invitations
- s. Attitudinal formulae and sentence stems *in fact, to be honest, it is clear that*
 – used to signal speaker's attitudes towards their utterances and interlocutors
- t. Proverbs and proverb fragments *When in Rome*
 – express general ideas by means of non-literal meaning (metaphors, metonymies, etc.); equivalent to complete sentences but are often abbreviated
- u. Commonplaces *it's a small world; we only live once; the sky is the limit*
 – non-metaphorical complete sentences that express tautologies, truisms and sayings based on everyday experience
- w. Slogans *Make love, not war.*
 – short directive phrases made popular by their repeated use in politics or advertising

Table 2: Statistical analyses of chunk measures in a cross-sectional study (*significant at the .05 level; **significant at the .001 level).

Chunk measures	Oct 07 means (SDs) indep. samples t-test ($\alpha = .05$)		May09 means (SDs) indep. samples t-test ($\alpha = .05$)		Effects (p-values) repeated measures ANOVA ($\alpha = .05$)	
	high-input	low-input	high-input	low-input	time	group
chunk types/text	4.00 (2.45)	3.45 (1.86)	7.00** (0.89)	4.55 (1.86)	.000	.036
% chunk-words/text	26.8* (16.28)	12.40 (7.29)	45.60* (9.82)	33.2 (10.49)	.000	.652
chunks/ 100 words	10.62* (5.64)	4.97 (2.51)	14.29 (2.15)	12.96 (3.74)	.000	.046
mean chunk length/text	1.65 (0.89)	1.02 (0.56)	3.13* (0.40)	2.60 (0.40)	.000	.760
type-token lex.col/100 words	2.62 (1.94)	1.15 (1.09)	3.26 (1.75)	4.02 (2.95)	.006	.063
type-token compl./100 words	1.16 (2.02)	0.26 (0.44)	3.15 (1.20)	4.75 (3.10)	.000	.041
type-token conv.sent stem/100 words	0.25 (0.55)	0.05 (0.17)	2.07 (1.11)	0.56 (0.81)	.000	.004
mean text length	65.09 (20.23)	114.18* (42.83)	157.72* (39.26)	103.00 (35.45)		
Pearson correlation text length/chunks	.612*	.146	.867**	.726*		