LingBaW. Linguistics Beyond and Within 10 (2024), 49–63

Largest-chunking and group formation: Two basic strategies for a cognitive model of linguistic processing

László Drienkó
SzSzC Jáky, Hungary

Abstract

The present study aims at shedding further light on how AGREEMENT GROUPS (AG) processing (e.g. Drienkó 2020a) and LARGEST CHUNK (LCh) segmentation (e.g. Drienkó 2018a) can be combined to model the emergence of language. The AG model is based on groups of similar utterances which enable combinatorial mapping of novel utterances. LCh segmentation is concerned with cognitive text segmentation, i.e. with detecting word boundaries in a sequence of linguistic symbols. Previous cross-linguistic research on French, English, and Hungarian texts (Drienkó 2020b) demonstrated that LCh segmentation is not efficient when words are the basic segmentation units and utterances are the target sequences. However, almost all utterance boundaries were identified at the expense of inserting relatively many extra boundaries. These extra boundaries delineated reoccurring fragments for building longer utterances. The present analysis of English mother-child data confirms previous findings that in spite of the relatively low efficiency of word-based LCh segmentation with respect to utterance boundaries, LCh segments can still prove to be useful word combinations for AG processing. Furthermore, compared with the previous experiments, the data suggest higher boundary precision (42%) and higher coverage (85%). These findings, on the one hand, support the claim that LCh fragments can be useful in linguistic processing (with AGs), and, on the other hand, are in line with a view that mother-child language facilitates processing more than other speech contexts.

Keywords: Cognitive computer modelling; segmentation; syntactic processing; language acquisition

1. Introduction

The AG language processing model, initially proposed by Drienkó (2014), adopts a distributional approach, relying on word distribution to group utterances. Harris (1951, 1952) pioneered distributional methods in linguistics, considering *contexts* for linguistic items. Kiss

LingBaW. Linguistics Beyond and Within, e-ISSN: 2450-5188
DOI: 10.31743/lingbaw.18008

(1973) introduced a word categorisation model using cluster analysis, expanded by Redington et al. (1998). Finch et al. (1995) adopted a similar method to assign categories to word sequences, i.e. to phrases. Mintz (2003) formalized context using *frequent frames*, i.e. preceding and succeeding words, while Weisleder and Waxman (2010) explored, besides Mintz's *mid-frames*, the usefulness of *end-frames* for categorisation. Additionally, St. Clair et al. (2010) argued for *flexible frames*. Cameron-Faulkner et al. (2003) found framing effects in language acquisition, which findings were confirmed by cross-linguistic results in e.g. Stoll et al. (2009). AGs can be viewed as combinations of such framing contexts. Wang and Mintz (2010) claim that grammatical relations are more consistent within frequent frames than in bigrams, which accords well with our view that AGs represent linguistic relations. Bannard and Matthews (2008) suggest that children tend to store word sequences in memory during language acquisition. The organisation of such stored utterances into groups based on similarity is a key concern of the AG model.

Early work on speech segmentation is exemplified by Harris (1955) where statistical cues were used to predict linguistic unit boundaries. Saffran et al. (1996) demonstrated statistical information availability in language acquisition. Various other cues like syllable distribution and prosody also affect speech segmentation strategies (Mattys et al. 2005; Cutler et al. 1987; Cutler et al. 1988; Thiessen et al. 2007; Bagou et al. 2002). The LCh method, as proposed in Drienkó (2016a), offers a quantitative approach based solely on linguistic structure, taking no advantage of further cues like stress or metrical features.

The structure of the current study is organized as follows. Sections 1.1 and 1.2 offer a brief introduction to agreement groups and LCh segmentation. Section 1.3 addresses the issue of combining word-based LCh segmentation with AG processing, setting the context for our analysis consisting of three computer experiments. Section 2 presents the empirical results obtained from the experiments. Section 3 discusses the significance of the findings in relation to linguistic modelling. Section 4 provides concluding remarks summarizing the key points of the study.

1.1. Agreement Groups

The concept of *agreement groups* and *agreement groups coverage* has been explored in various studies as a distributional approach to modelling linguistic processing. Drienkó (2014) demonstrated that agreement groups, which are groups of 2-5 word long utterances differing from a base utterance by only one word, can account for a certain percentage of novel utterances in English mother-child speech. These AGs may facilitate categorization (lexical/syntactic, semantic) and could potentially serve as the basis for actual agreement relations. Similar findings were reported for Hungarian and Spanish in Drienkó (2013a). To handle longer utterances, the notion of *coverage* was introduced in Drienkó (2013b, 2015, 2016b). The coverage apparatus aims to identify 2-5-word long fragments within an input utterance and map them onto agreement groups. In the case of English mother-child utterances from the Anne files of the Manchester corpus (Theakston et al. 2001) in the

CHILDES database (MacWhinney 2000), the author found average coverage values of 78% and 83% for the *continuous* and *discontinuous* cases, respectively.

The central objective of the AG approach is to arrange the utterances of a linguistic corpus into groups differing in only one word from a given utterance. These AGs serve as the basic processing units of the model for mapping utterances. Novel utterances are mapped onto AGs of *familiar* utterances, viz. the utterances of a training corpus. An utterance can be mapped onto an AG if it can be obtained by choosing words from the subsequent columns of a corresponding hypothetical table for the AG, where each column represents an utterance position and contains all the words of the AG occurring in the corresponding position. For example, the agreement group AG1 in (1) licenses the novel utterances in (2), i.e. each utterance in (2) can be mapped onto AG1. For a given AG, the novelty of mappable utterances is graded in proportion to in how many words an utterance in question differs from the words of the 'base' utterance of the AG. Utterance *a boy laughs*, e.g. differs from the base utterance *the girl talks* in all the three positions – cf. the boldface words – while the other three utterances, *a girl laughs, the boy laughs* and *a boy talks,* involve only two positional differences. Note that novelty for a particular AG begins with two positional differences since the utterances within an AG already involve difference in one position from the base utterance.

(1) AG1
the girl talks
the girl **laughs**
a girl talks
the **boy** talks

(2) *a girl laughs*
the boy laughs
a boy talks
a boy laughs

Besides the immediate AG-mapping level, the AG model assumes a *coverage* mechanism that processes utterances as combinations of shorter utterances. The processing task of the coverage mechanism consists in trying to establish a COVERAGE STRUCTURE for an utterance by identifying 2-5-word long fragments in it that can be directly mapped onto AGs. Suppose we want to process utterance *a little boy laughs* and our store of AGs, besides AG1 in (1), also includes AG2 as shown in (3). We can obtain a corresponding coverage structure by identifying fragments *a boy laughs* and *little boy* which fragments can be mapped onto groups AG1 and AG2, respectively. Cf. Table 1. Note that the AG model allows for discontinuous mapping. Fragment *a boy laughs* is discontinuous in utterance *a little boy laughs* owing to the inserted word *little*.

(3) AG2
big girl
little girl
big **boy**

Table 1: *Coverage structure for 'a little boy laughs'*

a	***little***	**boy**	***laughs***	
a		boy	laughs	AG1 (discontinuous fragment)
	little	boy		AG2

We associate a 100% coverage value with coverage structures where each utterance position is covered, as in Table 1. However, there might be utterance positions that cannot be covered by AG-compatible fragments. As the coverage structure in Table 2 illustrates, the two groups AG1 and AG2 in our running example do not suffice to completely cover utterance *a little boy often laughs* due to the fact that no AG can be found for mapping any utterance fragment containing the word *often*. Thus, the coverage value for *a little boy often laughs* is 4/5 = .8 (80%) since four utterance positions of five are covered.

Table 2: *Coverage structure for 'a little boy often laughs'*

a	***little***	**boy**	**often**	***laughs***	
a		boy		laughs	AG1 (discontinuous fragment)
	little	boy			AG2

The AG model operates on two basic levels of linguistic processing. The first level involves direct mappings onto AGs for handling holophrases, shorter utterances, or *formulaic* expressions. The second level requires more computational effort as it involves finding legal (AG-compatible) fragments (Level 1 operation) and then selecting an optimal combination of fragments to ensure grammaticality. This duality is reflected in the coverage structures of utterances. Drienkó (2020a) discusses additional dualistic properties of the AG framework and highlights its relevance to research on cognitive linguistic processing. This includes topics such as generalization, categorization, a semantic/syntactic interpretation of the *less-is-more* principle in Newport (1990), its relationship to U-shaped learning (Strauss 1982) and "vocabulary spurt" (e.g., Ganger & Brent 2004), parallels with the dual-process model of Van Lancker Sidtis (2009), lateralization of formulaic and *analytical* speech (e.g., Sidtis et al. 2018), neurolinguistic processing (Bahlmann et al. 2006), and the processing of complex linguistic structures such as long-distance dependencies, crossing dependencies, or embeddings (also discussed in Drienkó 2016b).

1.2. Largest-Chunk segmentation

The Largest Chunk (LCh) segmentation algorithm as proposed in Drienkó (2016a, 2018) for inferring utterance fragment boundaries looks for locally maximal chunks that occur at least twice in a sequence of linguistic units, fundamentally, letters of the alphabet. Some elementary segmentation examples are listed as (4i)-(4iii). The input sequence (4i), *abcabc*, e.g. is segmented as *abc abc*, since the largest chunk that occurs twice is *abc*. In (4ii) there is only one

c, so the *ab* chunks are locally maximal. Example (4iii) illustrates how some *nested dependency* structures can be captured by LCh segmentation.

(4) i abcabc → abc abc
ii abcab → ab c ab
iii abcdefefcdab → ab cd ef ef cd ab

The segmentation results are interpreted in terms of four precision metrics: INFERENCE PRECISION (IP), ALIGNMENT PRECISION (AP), REDUNDANCY (R), and BOUNDARY VARIABILITY (BV). The definitions are given under (5). Note the interdependence of the precision values: IP × R = AP, since (cib/aib) × (aib/acb) = cib/acb.

(5) Inference Precision = cib/aib
(correctly inferred boundaries/all inferred boundaries)

Redundancy = aib/acb
(all inferred boundaries/all correct boundaries)

Alignment Precision = cib/acb
(correctly inferred boundaries/all correct boundaries)

Boundary variability = $\Sigma\Delta f_i$/aib
(the average distance, in characters, of an inferred boundary from the nearest correct boundary)

For a simplistic illustration of how Inference Precision is obtained consider the toy corpus of two utterances {*toby is, toby in*}. When the basic segmentation units are the letters of the text, the LCh algorithm outputs the segments *tobyi*, *s*, *tobyi*, and *n* as in (6). Since 2 boundaries are correct of all the 4 inferred boundaries – viz. the boundaries after *s* and *n* – Inference Precision is 2/4 = 0.5. Recall that IP is defined as the proportion of correctly inferred boundaries, *cib,* to all inferred boundaries *aib*. Cf. (5).

(6) tobyistobyin → tobyi s tobyi n

When segmentation is based on syllables, we anticipate higher precision values since no erroneous syllable-internal boundaries can be inferred. The LCh segments for our toy corpus {*toby is, toby in*} would be *to-by-*, *is-*, *to-by-*, and *in-*, cf. (7). Since all the four inferred boundaries are correct, IP = 4/4 = 100%.

(7) to-by-is-to-by-in- → to-by- is- to-by- in-

The LCh algorithm, as described in Drienkó (2017), was used to segment utterances in English, Hungarian, Mandarin, and Spanish. The algorithm achieved an IP range of 53% - 66% when segments were based on letters. However, when syllables were used as the basic units of segmentation, the IP values significantly improved. In Drienkó (2018a) the IP range for syllables was found to be 77% - 95%, with an average of 86%. This suggests that using syllables as units of segmentation leads to higher precision in boundary inference.

The LCh segmentation strategy aligns with Peters' (1983) approach to language acquisition, where learners extract large chunks from the speech stream and form the 'ultimate' units of language by segmenting and fusing relevant chunks. The results also

support a *less-is-more* interpretation (Newport 1990), indicating that less detail in utterance structure – syllables versus letters – may facilitate higher precision in boundary inference.

The LCh strategy allows for direct quantitative results based solely on the linguistic structure of the text, without relying on additional cues such as stress or metrical features. However, it is worth noting that LCh segmentation may be compatible with other cognitive strategies and can be aided by cognitive cues. In fact, Drienkó (2018b) reported that utterance boundary information enhances LCh segmentation, which aligns with research on infant word segmentation and, in particular, with the Edge Hypothesis of Seidl et al. (2006) suggesting that extraction of target words is facilitated by utterance boundaries.

1.3. Word-based largest chunks for Agreement Groups processing

The AG model assumes that language learners have access to clearly defined utterance boundaries in their training corpus. However, this assumption does not align well with real-life language acquisition, where learners are exposed to continuous speech without explicit boundary markers. Previous research suggests that word boundaries can be detected with high precision using the LCh strategy, particularly in the case of syllable-based segmentation (Drienkó 2017, 2018a). If we assume that language learners have a tool for detecting word boundaries, such as syllable-based LCh segmentation, it may be valuable to explore how this segmentation strategy can be useful when considering the word as the basic unit of text. It is possible that the strategy could identify recurring word combinations that correspond to phrases and utterances. These "phrases" (or speech fragments) could then be input to the group formation algorithm of the AG model. The resulting set of AGs could be used for syntactic processing of new utterances, conditioning a cognitive computer model for the emergence of language that relies on LCh segmentation, AG formation, and their associated mapping mechanisms.

Some cross-linguistic results were reported in Drienkó (2020b) testing the LCh+AG ("syntax out of a stream of words") approach against the short novel *Le Petit Prince (The Little Prince)* by Antoine de Saint-Exupéry in three languages: French, English, and Hungarian. It was concluded that LCh segmentation is not very efficient when words are the basic segmentation units and utterances are the target sequences. However, almost all utterance boundaries were identified at the expense of inserting relatively many extra boundaries. These extra boundaries delimited reoccurring fragments that could be used for producing coverage structures for longer utterances. The present study explores a different register, mother-child language, in order to see how linguistic context affects the insights that the combination of largest-chunking and AG formation yields.

In the experiments, the input corpus of utterances was transformed into a sequence of words by removing utterance boundaries, and the resulting word sequence was segmented using the LCh segmentation algorithm. The word combinations (largest chunks) obtained in the first stage were then used to generate AGs. Finally, these AGs were used to map utterances from a novel section (test set) of the original corpus, allowing for testing of coverage. It is important to note that, for computational reasons, utterance boundaries were included in the test set, which means that our

results may underestimate the model's coverage potential since word combinations spanning utterance boundaries were not considered. Additionally, to gain a more detailed understanding of the processing mechanisms, the corpus was divided into three parts, and three separate experiments were conducted. The results of these experiments are presented in Section 2.

2. The experiments

In our experiments we used the Anne files of the Manchester corpus (Theakston et al. 2001) in the CHILDES database (MacWhinney 2000). The Anne section of the corpus contains 68 files, 1a through 34b, each file consisting of the tapescript of a 30-minute mother-child session. The dataset was divided into three subsets – files 1a-11b, 12a-22b, 23a-32b – and coverage was measured separately for each. To obtain utterance fragments, we reduced the data subsets even further. We regarded each 60-minute mother-child session as a short text, i.e. a sequence of words without utterance boundaries, and segmented them via the LCh segmentation algorithm. However, for a given coverage experiment, segments from all its 60-minute sessions were considered. For instance, in Experiment 1 the first collection of segments came from files 1a and 1b, the second collection from 2a and 2b, etc., and the segments of all the eleven collections were used to form AGs. Coverage was then tested on file 12a, corresponding with the next 30-minute mother-child session. In Experiment 2 the first collection of segments came from files 12a and 12b, the last collection from 22a and 22b, and coverage was tested on file 23a. Finally, in Experiment 3, segments were obtained from sessions 23 through 32 and coverage was measured on file 33a.

2.1. Experiment 1

In Experiment 1, after merging *a* and *b* sessions, we obtained LCh segments from files 1 through 11. Table 3 shows the precision metrics for the segmentation procedure. Recall that Inference Precision (IP) represents the proportion of correctly inferred boundaries (cib) to all inferred boundaries (aib), i.e. IP = cib/aib, Redundancy (R) is computed as the proportion of all the inferred boundaries to all the correct (original) boundaries (acb), i.e. R = aib/acb, Alignment Precision (AP) is specified as the proportion of correctly inferred boundaries to all the original boundaries, i.e. AP = cib/acb, and Boundary Variability (BV) designates the average distance, in characters, of an inferred boundary from the nearest correct boundary. Cf. (5). Here we specifically include BV_{wo} for measuring the average of the distance from the nearest correct boundary in words, since the basic textual unit in the experiments of this study is the word.

Table 3: *LCh segmentation precision results for Experiment 1*

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	Avr.
IP	0.472	0.496	0.496	0.487	0.474	0.475	0.430	0.435	0.460	0.389	0.463	**0.461**
R	1.623	1.484	1.482	1.437	1.615	1.632	1.783	1.770	1.577	1.951	1.593	**1.538**
AP	0.766	0.736	0.735	0.700	0.766	0.775	0.767	0.771	0.726	0.759	0.738	**0.749**
BV	4.863	4.417	4.407	4.652	4.608	4.832	5.162	5.430	5.234	6.027	4.747	**4.943**
BV_{wo}	1.045	0.922	0.925	0.979	0.966	1.006	1.069	1.121	1.083	1.266	0.978	**1.011**

Overall, we obtained 33179 segment tokens from the 11 sessions, 16519 of which were multiword segments, i.e. segments containing at least two words. The distribution of multiword segments with respect to their lengths measured in words is sketched in Figure 1.

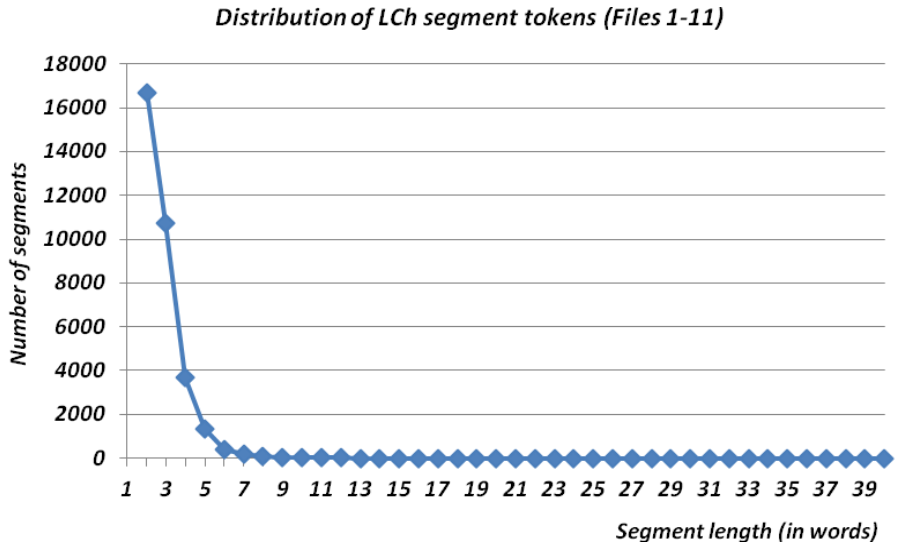

Figure 1: *The distribution of multiword LCh segments from files 1-11 with respect to their lengths*

Of all the 16519 multiword utterance fragments we selected those which contained at most five words, and these were used for the formation of AGs. The distribution of the 5905 two-to-five-word-long segment types in terms of their lengths measured in words is given as Figure 2. The number of words (types) that occurred in the 5905 segments was 953.

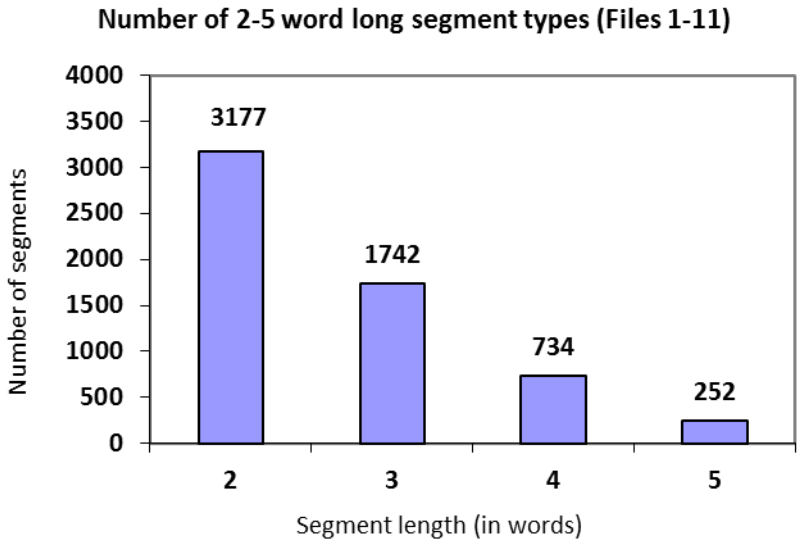

Figure 2: *The distribution of two-to-five-word-long LCh segment types from files 1-11 in terms of segment length measured in words*

Since each utterance fragment had its own group, there were 5905 AGs. The utterances in session 12a were used for testing the coverage potential of this 5905-group AG system. There were 565 utterance types in file 12a, 43 of which being one-word utterances. We applied the

coverage apparatus to the 522 multiword utterances in 12a. Via dividing the sum of the coverage values for the individual utterances in the test file by the number of utterances we obtain average coverage. The average coverage value for Experiment 1 was 457.3/522 = 87.6%. If we assume, in accordance with our word-based LCh segmentation procedure, that all words are "known" to the AG system, coverage becomes somewhat higher since one-word utterances in the test set can trivially be covered by themselves. Thus, by also taking the 43 one-word utterances into consideration, we get (457.3 + 43 = 500.3)/(522 + 43 = 565) = 88.5% as average coverage.

2.2. Experiment 2

In Experiment 2 the first collection of segments came from files 12a and 12b, the last collection from 22a and 22b, and coverage was tested on file 23a. After merging *a* and *b* sessions, we obtained LCh segments from files 12 through 22. Table 4 shows the precision metrics for the segmentation procedure. Recall that IP = cib/aib, R = aib/acb, AP = cib/acb, and BV designates the average distance, in characters, of an inferred boundary from the nearest correct boundary. BV_{wo} gives the average distance measured in words.

Table 4: *LCh segmentation precision results for Experiment 2*

	12	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	Avr.
IP	0.417	0.424	0.395	0.412	0.387	0.417	0.407	0.442	0.429	0.438	0.425	**0.417**
R	1.835	1.745	1.955	1.878	1.983	1.888	1.922	1.724	1.718	1.751	1.75	**1.747**
AP	0.764	0.741	0.773	0.773	0.767	0.787	0.782	0.762	0.738	0.767	0.743	**0.763**
BV	5.580	5.499	6.473	5.475	6.058	5.775	5.764	5.316	5.026	5.229	5.184	**5.580**
BV_{wo}	1.152	1.137	1.342	1.137	1.246	1.177	1.164	1.097	1.04	1.068	1.068	**1.11**

Overall, we obtained 35619 segment tokens from the 11 sessions, 17864 of which were multiword segments, i.e. segments containing at least two words. The distribution of multiword segments is sketched in Figure 3.

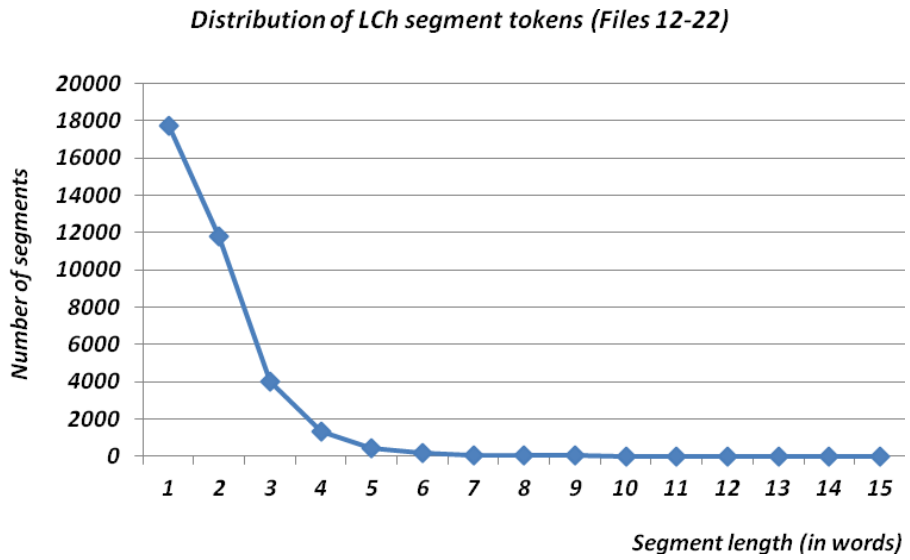

Figure 3: *The distribution of multiword LCh segments from files 12-22 with respect to their lengths*

Of all the 17864 multiword utterance fragments we selected those which contained at most five words, and these were used for the formation of AGs. The distribution of the 6197 two-to-five-word-long segment types is given as Figure 4. The number of words (types) that occurred in the 6197 segments was 1038.

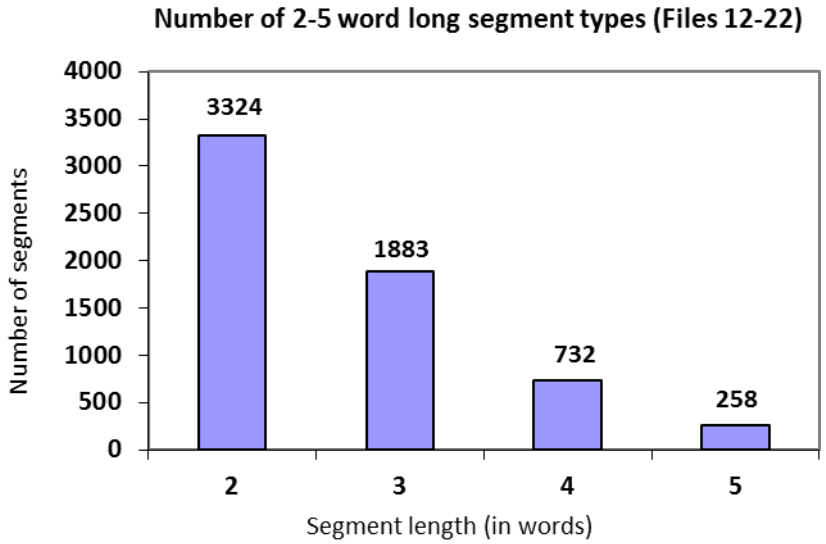

Figure 4: *The distribution of two-to-five-word-long LCh segment types from files 12-22 in terms of segment length measured in words*

The utterances in session 23a were used for testing the non-discontinuous coverage potential of the 6197 AGs. There were 526 utterance types in file 23a, 51 of which were one-word utterances. We applied the coverage apparatus to the 475-multiword subset of 23a. The average coverage value for Experiment 2 was 407.8/475 = 85.8%. If we take the 51 one-word utterances into consideration average coverage becomes (407.8 + 51 = 458.8)/(475 + 51 = 526) = 87.2%.

2.3. Experiment 3

In Experiment 3 we obtained LCh segments from files 23 through 32, after merging *a* and *b* sessions. Coverage was tested on file 33a. Table 5 shows the precision metrics for the segmentation procedure.

Table 5: *LCh segmentation precision results for Experiment 3*

	23	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	Avr.
IP	0.38	0.367	0.392	0.399	0.389	0.392	0.369	0.387	0.377	0.376	**0.383**
R	1.997	2.124	1.928	1.867	1.909	2.004	2.069	2.015	2.102	2.023	**2.01**
AP	0.759	0.780	0.755	0.745	0.742	0.785	0.763	0.781	0.793	0.760	**0.759**
BV	6.183	6.56	5.772	5.768	5.982	6.106	6.322	6.053	6.453	6.404	**6.293**
BV_{wo}	1.290	1.341	1.195	1.212	1.242	1.263	1.307	1.252	1.329	1.335	**1.312**

We obtained 34649 segment tokens from the 10 sessions, 16761 of which contained more than one word. The distribution of multiword segments is sketched in Figure 5.

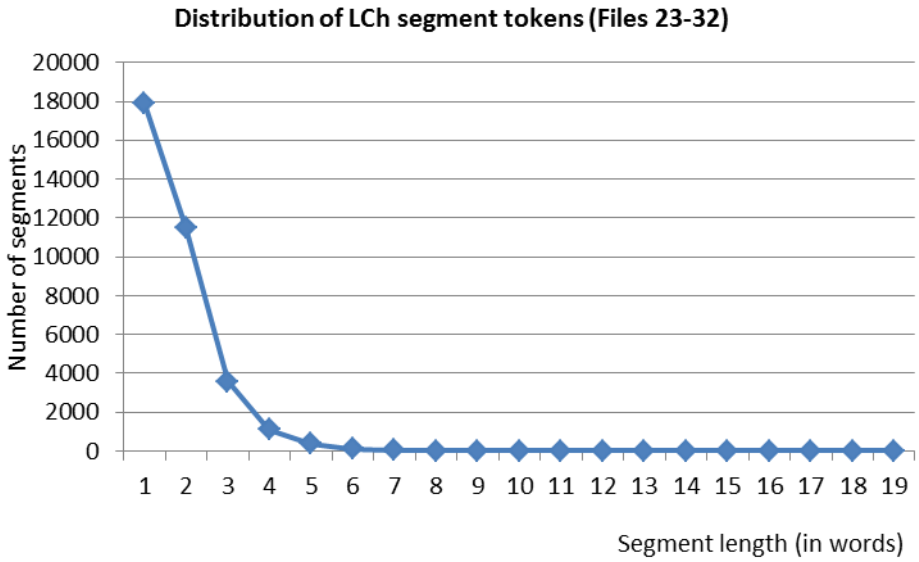

Figure 5: *The distribution of multiword LCh segments from files 23-32 with respect to their lengths*

Of all the 16761 multiword utterance fragments we selected those which contained at most five words, and these were used for the formation of AGs. The distribution of the 5805 two-to-five-word-long segment types is given as Figure 6. The number of words (types) that occurred in the 5805 segments was 1065.

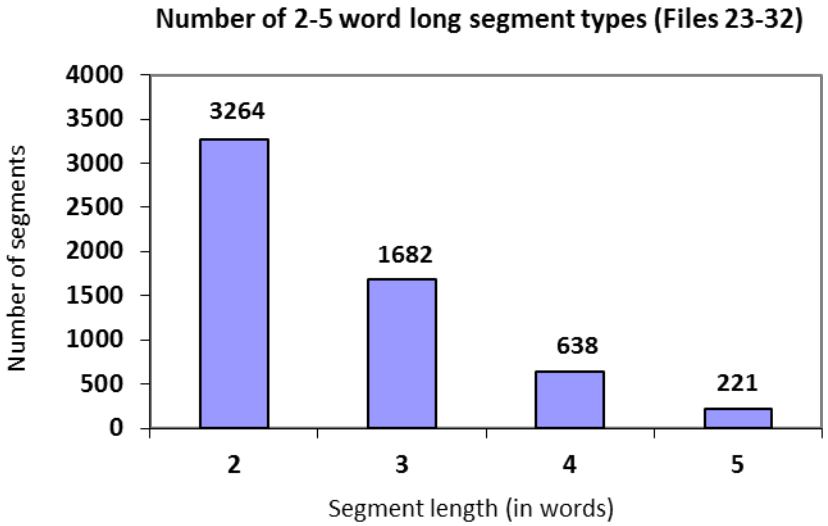

Figure 6: *The distribution of two-to-five-word-long LCh segment types from files 23-32 in terms of segment length measured in words*

The utterances in session 33a were used for testing the non-discontinuous coverage potential of the 5805 AGs. There were 515 utterance types in file 33a, 49 of which were one-word utterances. We applied the coverage apparatus to the 466-multiword subset of 33a. The average coverage value for Experiment 3 was 375.15/466 = 80.5%. If we take the 49 one-word utterances into consideration average coverage becomes (375.15 + 49 = 424.15)/(466 + 49 = 515) = 82.4%. Table 6 presents the average results from all the three experiments.

Table 6: *Overall average segmentation precision and coverage results*

	Datasets			
	1-11	**12-22**	**23-32**	**Average**
Avr. IP	0.461	0.417	0.383	0.420
Avr. R	1.538	1.747	2.01	1.765
Avr. AP	0.749	0.763	0.759	0.757
Avr. BV	4.943	5.580	6.293	5.605
Avr. BV_{wo}	1.011	1.11	1.312	1.144
Average coverage	0.876	0.858	0.805	0.846

3. Discussion

The Inference Precision values show that the proportion of correctly inferred boundaries of all inferred boundaries is about 40%, 42% overall. This suggests that the LCh segmentation mechanism, as compared with former findings (e.g. Drienkó 2017, 2018a), is only moderately robust when words are the basic units for segmentation and utterance boundaries are to be inferred. Nevertheless, the other precision parameters reveal significant features of the LCh strategy that condition the emergence of useable word combinations for syntactic processing. First of all, the distance of an erroneously inferred boundary is, on average, merely 1.144 words (5.605 characters) from the nearest correct utterance boundary, i.e. by shifting the erroneous boundary ca. one word to the left or to the right we reach a correct utterance boundary. Secondly, Alignment Precision is relatively high. The 75.7% average value indicates that about three quarters of the original utterance boundaries are correctly identified. Perhaps most importantly, the relatively high AP value is achieved via inserting extra boundaries. These additional boundaries are incorrect with respect to utterance edges. However, they mark out reoccurring word sequences that can be used as building blocks for utterances. The 1.765 average Redundancy value shows that roughly twice as many boundaries are inferred as would be strictly necessary to identify the original utterances.

The coverage results of our experiments, averaging 84.6%, are fairly impressive, especially when compared to the relatively low IP values. The fact that, on average, over 80% of an utterance can be covered by fragments output by the LCh module of the processing system indicates that LCh segmentation may be a promising mechanism for providing useful word combinations, or "phrases", i.e. building blocks for syntactic processing. In fact, the 84.6% average value is higher than our previous result, 78%, obtained with the 2-5-word-long utterances in the training corpus (Drienkó 2013b), which indicates that LChs provide a comparable, or even better basis for AG processing. The high coverage value also suggests that the formation of groups, AGs in particular, can be a successful strategy for creating a syntactic mapping apparatus. Thus, our results would be in line with a usage-based model of the

emergence of linguistic capacities supported by two fundamental cognitive strategies – LCh segmentation and the formation of AGs.

Compared with the *Le Petit Prince* experiment in Drienkó (2020b), the data here suggest higher boundary precision and higher coverage: 42% and 85%, respectively, versus 16% and 45% in the former experiments. Cf. Table 7. Thus, besides drawing attention to largest-chunking and AGs, our findings also highlight the role of mother-child language in facilitating linguistic processing.

Table 7: *Comparison of the results with those from the Le Petit Prince experiment*

	Le Petit Prince The Little Prince Kis herceg **Average**	**Anne** **Average**
Avr. IP	0.16	0.420
Avr. R	5.6	1.765
Avr. AP	0.9	0.757
Avr. BV	17.02	5.605
Avr. BVwo	–	1.144
Avr. cov. (cont.)	0.45	0.846
Avr. cov. (discont.)	0.54	–

4. Conclusions

The primary objective of the present paper was to explore the viability of combining word-based LCh segmentation with AG processing. We reported empirical results from experiments with CHILDES mother-child data. It was found that word-based segmentation is not robust for inferring utterance boundaries, IP is around 40%. Nevertheless, the majority of utterance boundaries can be found, AP ≈ 76%, via inserting redundant boundaries, R ≈ 1.76. The resultant wealth of segments conditions the emergence of utterance components, or building blocks, that can be organised into AGs. Thus LCh segments prove to be useable word combinations for linguistic processing. As reflected in the coverage values, such building blocks can account for, on average, some 80% of the test utterances, which makes our approach a promising processing framework. Thus, the 'LCh+AG' approach can be regarded as a usage-based model of the emergence of linguistic capacities based on two fundamental cognitive strategies, LCh segmentation and AG formation. As the present results are quantitatively superior to previous findings from literary texts, besides drawing attention to largest-chunking and AGs, our findings also highlight the role of mother-child language in facilitating linguistic processing. In the experiments only non-discontinuous fragments were allowed for AG coverage. Previous research (Drienkó 2015, 2020b) suggests that discontinuous fragments

improve coverage results. Consequently, the 84.6% average coverage value that we report here might have been higher if discontinuous fragments had also been considered.

References

Bagou, O., C. Fougeron, and U. H. Frauenfelder. 2002. Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language. *Speech Prosody 2002*, Aix-en-Provence, France, April 11–13, 2002.

Bahlmann, G., and A. D. Friederici. 2006. Hierarchical and linear sequence processing: An electrophysiological exploration of two different grammar types. *Journal of Cognitive Neuroscience* 18(11): 1829–1842.

Bannard, C., and D. Matthews. 2008. Stored word sequences in language learning. *Psychological Science* 19(3): 241–248.

Cameron-Faulkner, Th., E. Lieven, and M. Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive Science* 27: 843–873.

Cutler, A., and D. M. Carter. 1987. The predominance of strong initial syllables in English vocabulary. *Computer Speech and Language* 2: 133–142.

Cutler, A., and D. G. Norris. 1988. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance* 14: 113–121.

Drienkó, L. 2013a. Distributional cues for language acquisition: a cross-linguistic agreement groups analysis. Poster presentation for the *11th International Symposium of Psycholinguistics, Tenerife, Spain* 20–23 March, 2013.

Drienkó, L. 2013b. Agreement groups coverage of mother-child language. Talk presented at the *Child Language Seminar, Manchester, UK,* 23–25 June, 2013.

Drienkó, L. 2014. Agreement groups analysis of mother-child discourse. In G. Rundblad, A. Tytus, O. Knapton, and C. Tang (eds.), *Selected Papers from the 4th UK Cognitive Linguistics Conference,* 52–67. London: UK Cognitive Linguistics Association. http://www.uk-cla.org.uk/proceedings/volume_2_36/36-32

Drienkó, L. 2015. Discontinuous coverage of English mother-child speech. Talk presented at the *Budapest Linguistics Conference,* Budapest, Hungary, 18–20 June, 2015.

Drienkó, L. 2016a. Discovering utterance fragment boundaries in small unsegmented texts. In A. Tanács, V. Varga, and V. Vincze (eds.) *XII. Magyar Számítógépes Nyelvészeti Konferencia.* (12th Hungarian Computational Linguistics Conference), 273–281. http://rgai.inf.u-szeged.hu/mszny2016/

Drienkó, L. 2016b. Agreement groups coverage of English mother-child utterances for modelling linguistic generalisations. *Journal of Child Language Acquisition and Development – JCLAD* 4(3): 113–158.

Drienkó, L. 2017. Largest chunks as short text segmentation strategy: a cross-linguistic study. In A. Wallington, A. Foltz, and J. Ryan (eds.), *Selected Papers from the 6th UK Cognitive Linguistics Conference*, 273–292. The UK Cognitive Linguistics Association. http://www.uk-cla.org.uk/files/downloads/15_drienko_273_292.pdf

Drienkó, L. 2018a. Largest-Chunk strategy for syllable-based segmentation. *Language and Cognition* 10(3): 391–407.

Drienkó, L. 2018b. The effects of utterance-boundary information on Largest-Chunk segmentation. Talk presented at the *20th Summer School of Psycholinguistics,* Balatonalmádi, Hungary, 10–14 June, 2018.

Drienkó, L. 2020a. Agreement Groups and dualistic syntactic processing. In A. Haselow, and G. Kaltenböck (eds.), *Grammar and cognition: Dualistic models of language structure and language processing,* 310–354. John Benjamins Publishing Company.

Drienkó, L. 2020b. Word-based largest chunks for Agreement Groups processing: Cross-linguistic observations. *Linguistics Beyond and Within (LingBaW)* 6(1): 60–73. https://doi.org/10.31743/lingbaw.11831

Finch, S., N. Chater, and M. Redington, 1995. Acquiring syntactic information from distributional statistics. In J. P. Levy, D. Bairaktaris, J. A. Bullinaria, and P. Cairns (eds.), *Connectionist models of memory and language*, 229–242. UCL Press: London.

Ganger, J., and M. R. Brent. 2004. Reexamining the Vocabulary Spurt. *Developmental Psychology* 40(4): 621–632.

Harris, Z. S. 1951. *Methods in structural linguistics*. Chicago, IL, US: University of Chicago Press.

Harris, Z. S. 1952. Discourse analysis. *Language* 28(1): 1–30.

Harris, Z. S. 1955. From phoneme to morpheme. *Language* 31: 190–222.

Kiss, G. R. 1973. Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation* 7: 1–41.

MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mattys, S. L, L. White, and J. F. Melhorn. 2005. Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General* 134(4): 477–500.

Mintz, T. H. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90(1): 91–117.

Newport, E. L. 1990. Maturational constraints on language learning. *Cognitive Science* 14: 11–28.

Peters, A. 1983. *The units of language acquisition*. Cambridge, Cambridge University Press.

Redington, M., N. Chater, and S. Finch. 1998. Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science* 22 (4): 425–469.

Saffran, J. R., R. N. Aslin, and E. L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294): 1926–8.

Seidl, A., and E. K. Johnson. 2006, Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science* 9: 565–573.

Sidtis, J. J., D. V. Sidtis, V. Dhawan, and D. Eidelberg. 2018. Switching Language Modes: Complementary Brain Patterns for Formulaic and Propositional Language. *Brain connectivity* 8(3): 189–196.

St. Clair, M. C., P. Monaghan, and M. H. Christiansen. 2010. Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition* 116(3): 341–360.

Stoll, S., K. Abbot-Smith, and E. Lieven. 2009. Lexically Restricted Utterances in Russian, German, and English Child-Directed Speech. *Cognitive Science* 33: 75–103.

Strauss, S. 1982. Ancestral and descendent behaviours: The case of U-shaped behavioural growth. In T. G. Bever (ed.), *Regressions in mental development: Basic phenomena and theories,* 191–220. Hillsdale, NJ: Lawrence Erlbaum Associate, Inc.

Theakston, A. L., E. V. Lieven, J. M. Pine, and C. F. Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language* 28(1): 127–52.

Thiessen, E. D., and J. R. Saffran. 2007. Learning to Learn: Infants' Acquisition of Stress-Based Strategies for Word Segmentation. *Language Learning and Development* 3(1): 73–100.

Van Lancker Sidtis, D. 2009. Formulaic and novel language in a 'dual process' model of language competence: evidence from surveys, speech samples, and schemata. In R. L. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley (eds.), *Formulaic Language: Volume 2. Acquisition, loss, psychological reality, functional applications*, 151–176. Amsterdam: Benjamins Publishing Co.

Wang, H., and T. H. Mintz. 2010. From linear sequences to abstract structures: Distributional information in infant-direct speech. In J. Chandlee, K. Franich, K. Iserman, and L. Keil (eds.), *Proceedings Supplement of the 34th Boston University Conference on Language Development 34 Online Proceedings Supplement.* Somerville, MA: Cascadilla Press. http://www.bu.edu/bucld/proceedings/supplement/vol34/

Weisleder, A, and S. R. Waxman. 2010. What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English. *Journal of Child Language* 37(5): 1089–108.

Cite this article as:

Drienkó, L. (2024). Largest-chunking and group formation: Two basic strategies for a cognitive model of linguistic processing. *LingBaW. Linguistics Beyond and Within, 10*, 49–63.