LingBaW

*Linguistics Beyond And Within 6 (2020), 60-73*

# Word-based largest chunks for Agreement Groups processing: Cross-linguistic observations

László Drienkó
*SZSZC-Jáky Székesfehérvár, Hungary*

**Abstract**
The present study reports results from a series of computer experiments seeking to combine word-based Largest Chunk (LCh) segmentation and Agreement Groups (AG) sequence processing. The AG model is based on groups of similar utterances that 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. Our observations are based on the text of *Le petit prince* (*The little prince*) by Antoine de Saint-Exupéry in three languages: French, English, and Hungarian. The data suggest that word-based LCh segmentation is not very efficient with respect to utterance boundaries, however, it can provide useful word combinations for AG processing. Typological differences between the languages are also reflected in the results.

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

## 1. Introduction

The AG language processing model as proposed in Drienkó (2014) is a usage-based distributional framework where groups of utterances are formed according to the distribution of words in a given corpus. Distributional linguistic research can be traced back at least to Harris (1951, 1952). In Harris's work the contexts, or environments, of a linguistic element were used to determine the distribution of the element in question. The contexts of words or phrases are particularly helpful in categorisation research based on cluster analysis (e.g. Kiss 1973, Redington *et al.* 1998, Finch *et al.* 1995), where context is typically formalised by context vectors. In Mintz (2003) a context, or *frame,* is provided by words that immediately precede or follow a given target element and a *frequent frame* is a context occurring with a frequency above an arbitrary threshold value. Weisleder and Waxman (2010) propose the utilization of *end-frames* with utterance-end information for categorisation. St. Clair *et al.* (2010) suggest that *flexible frames*, with bigram information from contexts, are more suited for categorising than only frequent frames. Item-based phrases in language acquisition research, as framed by words in initial positions, constitute a specific type of context (e.g. Cameron-Faulkner *et al.* 2003, Stoll

*et al.* 2009). AGs may be viewed as superimpositions of such contextual frames. According to Wang and Mintz (2010) "grammatical relations between words are more consistent in individual frequent frames than in bigrams" and "words within a frequent frame are especially "close" syntactically" (6, 8). Such views accord with our assumption that the "agreement relations" encoded in AGs represent syntactical/grammatical relations.

Early work on speech segmentation is exemplified by Harris (1955). His research focussed on statistical characteristics of language, fundamentally on successor frequencies, which he used for predicting word or morpheme boundaries. As documented by Saffran, Aslin and Newport (1996), infants may indeed be supported by statistical characteristics of speech in acquiring language. Research on speech segmentation has also demonstrated that several lexical and sub-lexical language-related cues play an important part in language acquisition (e.g. Mattys, White and Melhorn 2005). These cues can be utilised by various segmentation strategies. Metrical segmentation (Cutler and Carter 1987; Cutler and Norris 1988), for instance, is based on the distribution of strong and weak syllables. Also, infants can rely on stress patterns (Thiessen and Saffran 2007), or lengthening of speech sounds and/or rising in acoustic frequency (Bagou, Fougeron, and Frauenfelder 2002) for segmenting language. The LCh segmentation strategy does not employ such cues. It only needs information on the succession of linguistic elements in a particular text.

The structure of this paper is as follows: Sections 1.1 and 1.2 provide a short introduction to AGs and LCh segmentation. Section 1.3 sets the scene for the experiments by considering the issue of the possible combination of word-based LCh segmentation with AG processing. In Section 2, we present our empirical findings. In Section 3, we discuss the significance of the results with respect to linguistic modelling. Section 4 contains some concluding remarks.

### *1.1. Agreement Groups*

The AG model of language processing is a usage-based distributional framework operating with memorised groups of similar utterances, and cognitive mapping mechanisms. Thus a collection of familiar/known utterances enables the processing of novel word sequences. Formally, an AG can be regarded as a hypothetical table for concatenating linguistic units, where columns in the table represent (agreement) categories, and any element (word) in a column can be concatenated with any other in the next column.

The idea of *agreement groups* and *agreement groups coverage* was presented in a series of works as a distributional approach to modelling linguistic processing. Drienkó (2014) showed that agreement groups, i.e. groups of 2–5 word long utterances differing from a base utterance in only one word, can account for a certain percent of novel utterances of English mother-child speech, may facilitate categorisation (lexical/syntactic, semantic), and might serve as a basis for 'real' agreement relations. The findings were confirmed cross-linguistically by Hungarian and Spanish data in Drienkó (2013a). For the processing of longer utterances, the notion of *coverage* was introduced in Drienkó (2013b, 2015, 2016b). The coverage apparatus seeks to identify 2–5 word long fragments of an input utterance and map them onto AGs. By applying the AG coverage method to mother-child speech (Anne sessions, Manchester corpus: Theakston *et al.*

2001) from the CHILDES corpora, (MacWhinney 2000), it was found that the continuous and the discontinuous cases yielded, respectively, 78% and 83% average coverage values.

The essence of the AG approach lies in forming groups differing in only one word from a given utterance. In fact, each utterance of the training set has its own group. For instance, the training corpus (1) yields the AGs under (2).[1]

(1)

| | | | | | | |
|---|---|---|---|---|---|---|
| the dog | the cat | big dog | big cat | white dog | the big dog | |
| the white dog | the big cat | the dog laughs | the cat laughs | the dog cries | cat laughs | dog laughs |

(2)

| G1: | G2: | G3: | G4: | G5: | G6: | |
|---|---|---|---|---|---|---|
| <u>the dog</u> | <u>the cat</u> | <u>big dog</u> | <u>big cat</u> | <u>white dog</u> | <u>the big dog</u> | |
| the cat | the dog | big cat | big dog | big dog | the big cat | |
| big dog | big cat | white dog | the cat | the dog | the white dog | |
| white dog | | the dog | | | | |

| G7: | G8: | G9: | G10: | G11: | G12: | G13: |
|---|---|---|---|---|---|---|
| <u>the white dog</u> | <u>the big cat</u> | <u>the dog laughs</u> | <u>the cat laughs</u> | <u>the dog cries</u> | <u>cat laughs</u> | <u>dog laughs</u> |
| the big dog | the big dog | the dog cries | the dog laughs | the dog laughs | dog laughs | cat laughs |
| | | the cat laughs | | | | |

We think of groups as hypothetical tables as defined by the utterance length for the group (number of columns in the table), and the maximum number of words occurring in an utterance position (number of rows). We say that an utterance is compatible with a group (i.e. can be mapped on a group) if it can be obtained by choosing words from the subsequent columns of the corresponding hypothetical table. Although the novel utterance *white cat*, e.g., is not an utterance of the training set, it can be mapped on the *the dog* group, G1, or on the *big dog* group, G3. The assignment of 'agreement categories' is done with reference to groups and utterance positions, cf. (3). Categories *G9_3* and *G11_3* for *cries*, for instance, indicate that the word occurs in Group 9 and Group 11 in the third word position within the corresponding utterances.

(3)

| | |
|---|---|
| the: | G1_1, G2_1, G3_1, G4_1, G5_1, G6_1, G7_1, G8_1, G9_1, G10_1, G11_1 |
| big: | G1_1, G2_1, G3_1, G4_1, G5_1, G6_2, G7_2, G8_2, |
| white: | G1_1, G3_1, G5_1, G6_2, G7_2 |
| dog: | G1_2, G2_2, G3_2, G4_2, G5_2, G6_3, G7_3, G8_3, G9_2, G10_2, G11_2, G12_1, G13_1 |
| cat: | G1_2, G2_2, G3_2, G4_2, G6_3, G8_3, G9_2, G10_2, G12_1, G13_1 |
| laughs: | G9_3, G10_3, G11_3, G12_2, G13_2 |
| cries: | G9_3, G11_3, |

The COVERAGE STRUCTURE of an utterance is a tabular visualisation of a configuration of AGs onto which the fragments of the utterance in question can be mapped. For instance, Table 1 shows the possible fragments that can cover sentence *the big white dog laughs*. In Table 2 the words are represented by their agreement categories directly indicating which groups are involved.

[1] Examples from Drienkó (2017a).

***Table 1:*** *Schematic coverage structure for* the big white dog laughs

| **the** | **big** | **white** | **dog** | **laughs** |
|---|---|---|---|---|
| | | | dog | laughs |
| the | | | dog | |
| | big | | dog | |
| | | white | dog | |
| the | big | | dog | |
| the | | white | dog | |
| the | | | dog | laughs |

***Table 2:*** *Coverage structure with category information for* the big white dog laughs

| **the** | **big** | **white** | **dog** | **laughs** |
|---|---|---|---|---|
| | | | G12_1 | G12_2 |
| | | | G13_1 | G13_2 |
| G1_1 | | | G1_2 | |
| ... | | | ... | |
| G5_1 | | | G5_2 | |
| | G1_1 | | G1_2 | |
| | ... | | ... | |
| | G5_1 | | G5_2 | |
| | | G1_1 | G1_2 | |
| | | G3_1 | G3_2 | |
| | | G5_1 | G5_2 | |
| G6_1 | G6_2 | | G6_3 | |
| G7_1 | G7_2 | | G7_3 | |
| G8_1 | G8_2 | | G8_3 | |
| G6_1 | | G6_2 | G6_3 | |
| G7_1 | | G7_2 | G7_3 | |
| G9_1 | | | G9_2 | G9_3 |
| G10_1 | | | G10_2 | G10_3 |
| G11_1 | | | G11_2 | G11_3 |

The AG model assumes two basic levels of linguistic processing. The first level corresponds to direct mappings onto AGs for processing holophrases, shorter utterances, or "formulaic" expressions. The second level requires more computational effort since firstly legal (i.e. AG-compatible) fragments have to be found (Level 1 operation), then an optimal combination of fragments must be selected in order to effect grammaticality. This duality is reflected in the coverage structures of utterances. Further dualistic properties of the AG framework are communicated in Drienkó (2018a, 2020) along with contact points for research on cognitive-linguistic processing including generalisation, categorisation, a semantic/syntactic categorical interpretation of the *less-is-more* principle of Newport (1990) and its relationship to U-shaped learning (Strauss, 1982) and vocabulary spurt (e.g. Ganger and Brent 2004), parallelisms with the dual-process model of Van Lancker Sidtis (2009), lateralization of formulaic and analytical speech (e.g. Sidtis, Sidtis, Dhawan, and Eidelberg 2018), neurolinguistic processing (Bahlmann *et al.* 2006), and the processing of complex linguistic structures such as long-distance dependencies, crossing dependencies, or embeddings (cf. also Drienkó 2016b).

### *1.2. Largest-Chunk segmentation*

The LCh segmentation algorithm as proposed in Drienkó (2016a) searches for a succession of language chunks in an unsegmented sequence of linguistic symbols, which chunks are locally maximal in length and occur minimally twice in the whole sequence. To quantify the empirical results, four precision values are computed: INFERENCE PRECISION (IP), ALIGNMENT PRECISION (AP), REDUNDANCY (R), and BOUNDARY VARIABILITY (BV). As an immediate example, consider the toy corpus {*mary is, mary it*} consisting of two utterances. When the basic segmentation units are the characters of the text, the LCh algorithm outputs the segments *maryi*, *s*, *maryi*, and *t* as in (4). Since 2 boundaries are correct of all the 4 inferred boundaries – viz. the boundaries after *s* and *t* –, IP is 2/4=0.5. Note that *IP=cib/aib*, i.e. the number of correctly inferred/inserted boundaries (*cib)* divided by the number of all inferred/inserted boundaries (*aib*).

(4) maryismaryit → maryi s maryi t

When segmentation is based on syllables, we expect higher precision since boundaries cannot be erroneously inferred syllable-internally. The LCh segments for our example corpus {*mary is, mary it*} would be *ma-ry-*, *is-*, *ma-ry-*, and *it-*, cf. (5). Now IP=4/4=100%, since each of the four original boundaries is inferred correctly.

(5) ma-ry-is-ma-ry-it- → ma-ry- is- ma-ry- it-

In the cross-linguistic analysis of Drienkó (2017b), letter/character-based LCh segmentation was applied to utterances from English, Hungarian, Mandarin, and Spanish. The analysis yielded a 53% – 66% IP range, averaging 59%. Drienkó (2018b) examined how the precision values are affected when syllables are the basic segmentation units. It was found that syllable-based LCh segmentation results in considerably higher IP values, within an interval of 77%–95%, averaging 86%.

The LCh segmentation strategy may be compatible with the approach of Peters (1983) where a key role in language acquisition is played by segmenting and fusing linguistic chunks extracted from a continuous stream of speech. The LCh segmentation results might also suggest an analogy with the less-is-more interpretation of the data in Newport (1990), i.e. with the claim that certain cognitive skills may develop at the expense of others. In our case, boundary inference is more efficient when the processing of syllable structure (characters) is suppressed, i.e. when the syllable is taken to be the basic segmentation unit. Although the LCh strategy does not require cues like, for instance, metrical features, or stress patterns, it may be compatible with other cognitive strategies, and it can be aided by cognitive cues. In Drienkó (2018c) it was reported that LCh segmentation is enhanced by utterance boundary information, which fact is congruent with findings from infant word segmentation research. Indeed, the Edge Hypothesis of Seidl and Johnson (2006), in particular, assumes that utterance boundaries may provide an important cue in segmentation.

### *1.3. Word-based Largest Chunks for Agreement Groups*

The AG model tacitly assumes that utterance boundaries are readily available to the language learner, i.e. the training corpus consists of utterances with their well-defined boundaries. However, this is an over-optimistic attitude with regards to real-life natural language acquisition. The learner of a language is normally exposed to continuous speech without evident boundary markers. Previous research findings (Drienkó 2017b, 2018b) indicated that word boundaries can be detected via the Largest Chunk strategy with fairly high precision, especially for the syllable-based case. Assuming, then, that the language learner has a tool for detecting word boundaries (e.g. syllable-based LCh segmentation) it might be insightful to examine, as a next step, how the LCh segmentation strategy can be useful when the word is taken to be the basic textual unit. It might be expected that the strategy can detect reoccurring word combinations corresponding to phrases and utterances. These "phrases" (or rather speech fragments), in turn, could be input to the group formation algorithm of the AG model. Finally, the resultant body of AGs could condition a mapping mechanism for novel word sequences. Thus there could be a cognitive computer model for the emergence of language, basically building on two cognitive capacities, LCh segmentation, and AG formation together with the concomitant mapping mechanisms.

The present study reports results from a series of experiments seeking to combine word-based LCh segmentation with the AG utterance processing apparatus. In the experiments, first, the input corpus of utterances was transformed into a sequence of words by deleting punctuation symbols, i.e. utterance boundaries, and the resultant word sequence was segmented by the LCh segmentation algorithm.[2] In the next phase, the collection of word combinations (largest chunks) obtained in the first stage was used for producing AGs. Finally, the resultant AGs were used for mapping utterances of a novel section (test set) of the original corpus, i.e. for testing coverage. For computational reasons, we decided to include utterance boundaries in the test set. This means that our results quantitatively underestimate the coverage potential of the model in that word combinations possibly spanning utterance boundaries are ignored.

## 2. The experiments

Our observations are based on the text of *Le petit prince* (*The little prince*) by Antoine de Saint-Exupéry (1943a,b,c) in three languages: French, English, and Hungarian. The book contains 27 chapters. For each language, we utilised Chapters 1–26 for producing text segments whereas Chapter 27 was used for testing the coverage potential of AGs. In the segmentation phase, the text was divided into five subparts – Chapters 1–5, 6–10, 11–15, 16–21, and 22–26 – and each subpart was segmented separately. However, for a given language, segments from all the five subtexts were considered. For instance, in Experiment 1 the first collection of segments came

[2] Since the texts contain long and complex sentences, we chose to identify boundaries demarked by punctuation symbols including e.g. the comma, colon, or brackets, with utterance boundaries. That means that in the present study the term 'utterance boundary' should rather be understood as also subsuming clause or phrase boundaries besides sentence boundaries.

from Chapters 1–5 of the French text, the second collection from Chapters 6–10 etc., and the segments of all the five collections were used to form AGs. Coverage was then tested on Chapter 27. The same holds for Experiments 2 and 3 with the English and Hungarian version of the book, respectively.

### *2.1. Experiment 1: French*

In Experiment 1 the LCh segments were obtained from Chapters 1 through 26 of the original French text. Overall, there were 9665 segment tokens, 3522 types, provided by the 5 datasets.

Table 3 shows the precision metrics for the segmentation procedure. Note that *IP=cib/aib*, i.e. the number of correctly inferred/inserted boundaries, *cib,* divided by the number of all inferred/inserted boundaries, *aib*; *R=aib/acb,* i.e. the number of all inferred/inserted boundaries, *aib,* divided by the number of all correct, original, boundaries, *acb*; *AP=cib/acb,* i.e. the number of correctly inferred/inserted boundaries divided by the number of all the correct boundaries; and *BV* stands for the average distance between an inferred boundary and the nearest correct one, measured in characters.

***Table 3:*** *Segmentation precision and coverage results for Experiment 1*

| ***Le petit prince*** | | | | | | |
|---|---|---|---|---|---|---|
| | 1–5 | 6–10 | 11–15 | 16–21 | 22–26 | Average |
| IP | 0.13 | 0.15 | 0.18 | 0.16 | 0.16 | 0.16 |
| R | 7.11 | 6.16 | 4.92 | 5.41 | 5.55 | 5.83 |
| AP | 0.94 | 0.91 | 0.89 | 0.89 | 0.88 | 0.90 |
| BV | 20.91 | 18.57 | 15.76 | 17.07 | 17.13 | 17.89 |
| Average coverage (cont.) | 0.58 | | | | | |
| Average coverage (discont.) | 0.66 | | | | | |

Of all the 3522 segment types 1112 were multiword segments containing at most five words. These 1112 two-to-five-word-long segments were used for the formation of AGs. They contained 585 word types. Since each segment had its own group, there were 1112 AGs. The text of Chapter 27 was used for testing the coverage potential of this 1112-group AG system. The chapter consists of 37 sentences. In order to minimise computational costs sentence boundaries were retained, as well as boundaries demarcated by other punctuation symbols, e.g. commas and colons. One-word utterances were excluded from the analysis as meaningless for syntactic processing since AG-utterances minimally consist of two words. The test set eventually contained 70 text fragments which were input to the coverage evaluation procedure. By coverage we mean the percentage of utterance positions covered by at least one fragment mappable on some AG. For instance, assuming that utterance fragments *the dog, clever creature,* and *is a creature* can be mapped on some AGs, the coverage value for utterance '*the dog is a clever creature*' is 4/6 = 67% since four of the six utterance positions are covered by fragments *the dog,* and *clever creature*. This is the non-discontinuous case. In the discontinuous case, we would say that coverage is 100%, as the *is* and *a* positions of the sentence could be covered

discontinuously by *is a creature*, cf. Tables 4 and 5 displaying the continuous and discontinuous coverage structure for the utterance *the dog is a clever creature*.

***Table 4:*** *Coverage structure for 'the dog is a clever creature' (continuous fragments only)*

| **the** | **dog** | **is** | **a** | **clever** | **creature** |
|---|---|---|---|---|---|
| the | dog | | | | |
| | | | | clever | creature |

***Table 5:*** *Coverage structure for 'the dog is a clever creature' (discontinuous fragments allowed)*

| **the** | **dog** | **is** | **a** | **clever** | **creature** |
|---|---|---|---|---|---|
| the | dog | | | | |
| | | | | clever | creature |
| | | is | a | | creature |

Via dividing the sum of the coverage values for each utterance in the test set by the number of utterances in the test we obtain average coverage. The average coverage value from Experiment 1 was 40.36 / 70 = 57.6% for the continuous case and 46.44 / 70 = 66.3% for the discontinuous case, cf. Table 3.

### *2.2. Experiment 2: English*

In Experiment 2 the LCh segments came from Chapters 1 through 26 of the English translation of the book. Overall, there were 9316 segment tokens, 3046 types, provided by the 5 datasets. Table 6 shows the precision metrics for the segmentation procedure.

***Table 6:*** *Segmentation precision and coverage results for Experiment 2*

| **The Little Prince** | | | | | | |
|---|---|---|---|---|---|---|
| | 1–5 | 6–10 | 11–15 | 16–21 | 22–26 | Average |
| IP | 0.12 | 0.15 | 0.18 | 0.16 | 0.16 | 0.15 |
| R | 6.99 | 6.07 | 4.83 | 5.44 | 5.32 | 5.73 |
| AP | 0.87 | 0.89 | 0.87 | 0.89 | 0.85 | 0.87 |
| BV | 19.18 | 16.77 | 14.47 | 16.56 | 15.77 | 16.55 |
| Average coverage (cont.) | 0.58 | | | | | |
| Average covergae (discont.) | 0.67 | | | | | |

Of all the 3046 segment types 1140 were multiword segments containing at most five words. These 1140 two-to-five-word-long segments were used for the formation of AGs. They contained 483 word types. The text of Chapter 27 was used for testing the coverage potential of the 1140-group AG system. Due to the retention of punctuation-effected boundaries, the 37 sentences of the chapter were represented by 66 text fragments. The average coverage value in Experiment 2 was 38.45 / 66 = 58.3% for the continuous case and 44.23 / 66 = 67% for the discontinuous case, cf. Table 6.

### *2.3. Experiment 3: Hungarian*

In Experiment 3 the LCh segments were provided by Chapters 1 through 26 of the Hungarian translation of the book. Overall, we obtained 9260 segment tokens, 4053 types from the 5 datasets. Table 7 shows the precision metrics for the segmentation procedure.

***Table 7:*** *Segmentation precision and coverage results for Experiment 3*

| **Kis herceg** | | | | | | |
|---|---|---|---|---|---|---|
| | 1–5 | 6–10 | 11–15 | 16–21 | 22–26 | Average |
| IP | 0.12 | 0.18 | 0.22 | 0.20 | 0.19 | 0.18 |
| R | 7.59 | 5.36 | 4.18 | 4.29 | 4.79 | 5.24 |
| AP | 0.94 | 0.97 | 0.92 | 0.88 | 0.95 | 0.93 |
| BV | 23.86 | 17.32 | 12.81 | 14.62 | 14.51 | 16.62 |
| Average coverage (cont.) | 0.20 | | | | | |
| Average coverage (discont.) | 0.28 | | | | | |

Of all the 4053 segment types 533 were multiword segments containing at most five words. The 533 two-to-five-word-long segments were used for the formation of AGs. They contained 416 word types. Chapter 27 was used for testing the coverage potential of the 533-group AG system. Due to the retention of punctuation-effected boundaries, the 37 sentences of the chapter were represented by 84 text fragments. The average coverage value in Experiment 3 was 16.51 / 84 = 19.6% for the continuous case and 23.57/ 84 = 28.06% for the discontinuous case, cf. Table 7. Table 8 presents the average results from all the three experiments.

***Table 8:*** *Overall average segmentation precision and coverage results*

| | **PP** | **LP** | **KH** | **Average** |
|---|---|---|---|---|
| Average IP | 0.16 | 0.15 | 0.18 | 0.16 |
| Average R | 5.83 | 5.73 | 5.24 | 5.6 |
| Average AP | 0.90 | 0.87 | 0.93 | 0.9 |
| Average BV | 17.89 | 16.55 | 16.62 | 17.02 |
| Average covergae (cont.) | 0.58 | 0.58 | 0.20 | 0.45 |
| Average covergae (discont.) | 0.66 | 0.67 | 0.28 | 0.54 |

## 3. Discussion

The Inference Precision (IP) values show that the number of correctly inferred boundaries as compared to the number of all inferred boundaries is rather low, 16%, on average. This suggests that the LCh segmentation mechanism, as compared to previous results (Drienkó 2017b, 2018b), is not very efficient when words are the basic segmentation units and utterances are the target sequences, i.e. utterance boundaries are to be inferred. However, the other precision values reveal further features of the LCh strategy that make it capable of providing useable word combinations for syntactic processing. As the 90% average Alignment Precision (AP) value indicates, almost all of the utterance boundaries are correctly identified. The high AP value is

achieved via inserting extra boundaries. The 5.6 average Redundancy value shows that more than five times as many boundaries are inferred as would be strictly necessary to identify the original utterances. The extraneous boundaries are incorrect with respect to utterance edges. Nevertheless, they delineate reoccurring word sequences that can be used as building blocks for utterances. As reflected in the coverage values, such building blocks, or "phrases" can account for, on average, ca. 50% of the text.

For each language, the coverage value is higher when discontinuous fragments are permitted in processing. This fact echoes the findings in Drienkó (2015) claiming that discontinuous fragments in the coverage mechanism enhance the coverage potential of the AG model.

The data also reflect typological differences between the languages involved in the experiments. While the segmentation metrics are remarkably similar across languages, the 20% and 28% coverage values for Hungarian stand in clear contrast to the corresponding values for French and English, well over 50%, cf. Table 8. Since Hungarian is a highly inflectional language with relatively free word order, words and utterances are less likely to reoccur in the same form as in English or French. As repetitions are vital for LCh segmentation, just as similarity of word combinations is a key determinant in the formation of AGs, it can be expected that languages with a high degree of word-form variation and/or variable word order require more extensive training input in order to achieve the same level of efficiency of AGs. In other words, while Chapters 1–26 of the English and French texts provide enough similar segments for the resultant AGs to achieve relatively high coverage, that is not the case for Hungarian. The French and English training texts provided 1112 and 1140 word combinations, i.e. AGs, respectively. For Hungarian, the number of AGs was 533, ca. half the number of AGs for either French or English. Note that the differences seem to suggest a correlation between AG space and coverage. The same line of argumentation may be valid in explaining why the coverage values are slightly higher for English (58.3%, 67%) than for French (57.6%, 66.3%), assuming that French has richer inflectional morphology and freer word order. The English-French-Hungarian ranking also mirrors the morphological complexity of the languages as assumed to negatively correlate with the number of native speakers (e.g. Koplenig 2019): English, the least inflectional language, is spoken by the most native speakers, whereas the most inflectional, Hungarian, has rather few native speakers.

Boundary Variability (BV) is rather high, the average value is 17.02. This means that, on average, the distance between an inferred boundary and the nearest true one is about 17 characters. Via dividing BV by the average word length for the particular language we get an estimation of $BV_{wo}$, i.e. 'Boundary Variability measured in words'. Table 9 shows the average $BV_{wo}$ values calculated as BV/WL, where WL stands for '(average) word length'. The data reveal that, across the languages, the 17.02 character-based average BV corresponds to an average distance of 3.6 words from the nearest correct boundary. Again, dissociation can be observed in terms of language types. Hungarian can express grammatical dependencies within a sentence inflectionally. For instance, a single verb can refer to the subject and/or the object. In English and French, explicit parts of speech are needed for the subject or the object. Such facts suggest that Hungarian needs fewer albeit longer words to build a sentence or utterance. Fewer words in utterances, in turn, imply less chance to err in boundary inference (cf. also the IP and AP

values in Table 8). For instance, with a two-word-long utterance there is only one possibility to make an error, i.e. when a boundary is inserted between the two words. The distance of the incorrect boundary from either the boundary before the first word or the boundary after the second word is then one word. With a four-word-long utterance, an incorrectly inserted boundary in the middle would be two words away from either the left or the right correct boundary. Thus growing utterance length involves utterance positions that can possibly increase $BV_{wo}$. Consequently, the lower $BV_{wo}$ value for Hungarian than for either English or French, and the lower $BV_{wo}$ value for French than for English might ultimately be ascribed to morphological differences affecting utterance length.

***Table 9:*** *Average BV, word length (WL) and* $BV_{wo} = BV/WL$

| | **PP** | **LP** | **KH** | **Average** |
|---|---|---|---|---|
| Average BV | 17.89 | 16.55 | 16.62 | 17.02 |
| Average WL | 4.6 | 4.2 | 5.3 | 4.7 |
| Average $BV_{wo}$ | 3.88 | 3.94 | 3.13 | 3.6 |

## 4. Conclusions

The primary purpose of the present study was to investigate, cross-linguistically, the viability of combining word-based LCh segmentation with AG processing. We reported empirical results from experiments with the text of *The Little Prince*. It was found that word-based segmentation is not particularly efficient for inferring utterance boundaries, IP is ca. 16%. However, the majority of utterance boundaries can be reconstructed, AP≈90%, by way of inserting redundant boundaries, R≈5.6. The resultant abundance of segments, in turn, conditions the emergence of utterance components, or building blocks, that can be organised into AGs. Thus LCh segmentation provides useable word combinations for syntactic processing. As reflected in the coverage values, such building blocks, or "phrases" can account for about 50% of the test texts, on average, rendering our approach a promising processing framework. The data also highlight typological differences between the languages involved.

Our findings may be considered preliminary and need further validation against more extensive corpora. One step in that direction could be the analysis in Drienkó (in review) based on English mother-child utterances, with coverage over 80%. If it turns out to be adequately supported by empirical data, the 'LCh+AG' approach can offer a footing for establishing a usage-based model/theory of the emergence of language capacities built around two fundamental cognitive strategies, LCh segmentation and AG formation. The model might also be compatible with traditions in language acquisition research. Erickson and Thiessen (2015), e.g., conceptualise statistical learning as consisting of two major processes, Extraction and Integration. Extraction refers to statistical chunking whereas Integration involves similarity-weighted aggregation over chunks. Our LCh segments implicitly reflect the statistical-distributional structure of a sequence of symbols (words, in the present work) whereas the grouping of the segments into AGs is dictated by their distributional similarity-statistics.

## 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.

Cameron-Faulkner, Th., E. Lieven, 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.

de Saint-Exupéry, A. 1943a. *Le petit prince*. Édition du groupe "Ebooks libres et gratuits". Retrieved from https://www.ebooksgratuits.com/pdf/st_exupery_le_petit_prince.pdf.

de Saint-Exupéry, A. 1943b. *The little prince*. English translation by Jeff Mcneill. TheVirtualLibrary.org. Retrieved from https://thevirtuallibrary.org/index.php/en/?option=com_djclassifieds&format=raw&view=download&task=download&fid=14329.

de Saint-Exupéry, A. 1943c. *A kis herceg*. Hungarian translation by György Rónay. Retrieved from http://mek.oszk.hu/00300/00384/00384.pdf.

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. Retrieved from 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. Takács, V. Varga, and V. Vincze (eds.), *XII. Magyar Számítógépes Nyelvészeti Konferencia* (*12th Hungarian Computational Linguistics Conference*), 273–281. Retrieved from https://rgai.inf.u-szeged.hu/sites/rgai.sed.hu/files/MSZNY2016_web_ISO_B5.pdf.

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. Retrieved from http://jclad.science-res.com/archives_full_issu/Vol%204%20issue%203%20FULL%20ISSUE.pdf.

Drienkó, L. 2017a. Agreement groups processing of context-free utterances: Coverage, structural precision, and category information Talk presented at the *2nd Budapest Linguistics Conference,* 1–3 June 2017, Budapest, Hungary.

Drienkó, L. 2017b. 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. Retrieved from http://www.uk-cla.org.uk/files/downloads/15_drienko_273_292.pdf.

Drienkó, L. 2018a. Agreement groups and dualistic syntactic processing. Talk presented at the "*One Brain – Two Grammars? Examining dualistic approaches to language and cognition"* international workshop, 1–2 March 2018, Rostock, Germany. Retrieved from https://independent.academia.edu/LaszloDrienko/Conference-Presentations.

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

Drienkó, L. 2018c. 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. (in review). Largest-chunking and group formation: two basic strategies for a cognitive model of linguistic processing.

Drienkó, L. 2020. 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.

Erickson, L. C., and E. D. Thiessen. 2015. Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review* 37: 66–108.

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: 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: l–41.

Koplenig, A. 2019. Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society Open Science* 6: 181274. Retrieved from https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.181274.

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

Mattys, S. L, L. White, 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): 1926–1928.

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. Van Lancker 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, 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.), *Boston University Conference on Language Development 34 Online Proceedings Supplement*. Somerville, MA: Cascadilla Press. Retrieved from 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.