1   Segmenting words and stems in an artificial language

2   Georgia Loukatou[1] & Alejandrina Cristia[2]

3   [1] Stanford University

4   [2] Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes

5   Cognitives, ENS, EHESS, CNRS, PSL University

Abstract

How do learners break up the spoken stream into smaller, recombinable units? We investigate segmentation among adults exposed to artificial languages. Words in these languages are composed of stems and affixes, and thus both words and stems are potential segmentation units. We ask whether adults segment out stems, and whether they perform better on a stem or a word segmentation task. During the exposure phase, adult participants (N = 52; age > 18 years) watch videos of animals performing simple intransitive actions, while hearing sentences that describe those scenes. During a subsequent test phase, participants are given a two-alternative forced choice designed to assess the acceptability of words versus part-words, and stems versus part-stems.

Keywords: TO ADD

Word count: XXX words

Segmenting words and stems in an artificial language

TODO

- add the refs already cited

- flesh out the sections we already discussed

- perhaps reorganize intro slightly around two strands of literature

- draft results

- draft discussion

Introduction

Many researchers believe that segmentation, the process whereby meaningful units are identified from running speech, is an important task when learning a language [saffran, frank, etc]. Studying this process is difficult in natural language, because learning the form and meaning of such smaller units may benefit from both perception and production experience, particularly in adulthood. In this study, we aim to shed light on the way that adult learners segment fluent sentences by exposing them to an artificial language with a complex morphological structure, in a design that builds on two lines of research introduced next: Learning of forms, and learning of grammatical units.

Both lines of work use artificial languages, created by the researchers. It is assumed that some learning mechanisms are shared between artificial and natural language learning (Gómez & Gerken, 2000; Reber, 1967). When used and interpreted properly, artificial languages can help obtain better experimental control over the input to which learners are exposed (Fedzechkina et al., 2016; Folia et al., 2010), and isolate specific learning factors (Hayakawa et al., 2020), especially for segmentation.

Learning of forms.   In this line of work, the accent is on how people cut up fluent speech into units that are then available in memory, in a sort of lexicon. The artificial

languages used tend to be relatively simple, having only a handful of words and clear rules for between-word transitions, and are presented unimodally, typically with no meaning associated to the sentences. For instance, in Saffran et al. (1996), adults were exposed to an artificial language consisting of six trisyllabic words. The only cues available for word segmentation were the transitional probabilities between syllables (transitional probabilities between syllables spanning a word boundary were lower). The words were concatenated into fluent paragraphs and word boundaries were removed. The participants heard the text read by a speech synthesizer, with no acoustic word (or sentence) boundary cues, with an exposure lasting 21 minutes. At test, they were presented with the trisyllabic words and matched foils in a two-alternative forced choice. A preference for the words over the foils was interpreted as evidence that the participants had culled out the sequences from running speech, and thus learned the forms of the words.

The current standard for artificial language studies on segmentation is to focus on words as the target level of segmentation (e.g., Cunillera et al., 2010; Finn & Hudson Kam, 2008; Frank et al., 2013; Karuza et al., 2013; Kurumada et al., 2013; Saffran et al., 1997; Thiessen & Erickson, 2013; Tyler & Cutler, 2009 - but see segmentation of multi-word units by Siegelman & Arnon, 2015).[1] In most such studies, the words composing the artificial language have no sub-units (stems or affixes) that could be found within other words. Participants seem to be able to segment words out of artificial languages based on different cues (transitional probabilities cues, Saffran et al., 1996; mapping to word referents, Cunillera et al., 2010). Moreover, segmented units seem to be available for other linguistic processes. For instance, adult learners generalise non-adjacent dependencies (Frost & Monaghan, 2016), map the word forms to word referents (Cunillera et al., 2010) and learn the overall linguistic structure (Siegelman & Arnon, 2015). There has been an effort to

_____

[1] Numerous studies address segmentation and morpheme processing in infants and young children. Since the experiments in the present study bear only on adults, for whom learning mechanisms may be to a certain extent different from those found in children, we do not review that extensive literature.

investigate word segmentation in artificial languages with structures that resemble more those of natural languages (e.g. variable word length and frequency, Frank et al., 2010; Hoch et al., 2013; Kurumada et al., 2013; Schuler et al., 2017). One important feature of human languages is that words may be composed of smaller meaningful units – such as stems and affixes. Words' and stems' edges coincide in English much of the time, but not in many other languages. For example, in Modern Greek, adjective or verb stems can never be bare (without affixes). The verb , is composed by the verb stem (which means "sleep") and the suffix indicating the first person singular. The noun is composed by the noun stem (which means "magician") and the affix indicating plural number and male gender. Incidentally, both affixes and could also be found as free words in speech, because they are homophones with a preposition and a determiner, respectively. The fact that some stems never appear bare without other affixes is a cross-linguistically frequent feature, and could also be captured by artificial languages, and discovered by listeners when segmenting meaningful units. However, we are unaware of a study specifically testing whether listeners do segment out stems separately from affixes in this kind of study.

Learning of grammatical units

In a different line of research focused on learning grammatical units, artificial languages tend to be more complex than in the other line of research summarized above, and participants typically are exposed to the language multimodally, with visual scenes representing the content of segmented sentences (e.g., Finley & Newport, 2010; Finley & Wiemers, 2013; Hudson Kam & Newport, 2009). For example, in Hudson Kam & Newport, 2009, participants were presented with a language having over 30 nouns, 12 verbs, a negation particle, and two different determiners, which were the main focus of the study, and which appeared in the order determiner-noun. Participants learned the language over 8 sessions, each over 25 minutes long, by listening to each sentence while watching the associated video, and then repeating the sentence. The test session contained vocabulary

tests covering the language's nouns, as well as a sentence completion test, where the participant watched an event video and heard the main verb, having to produce the rest of the sentence. Results revealed that participants had been able to segment noun stems from the preceding determiners, even though exposure had been to full sentences, with no breaks between words.

Other studies in this line of work reveal that adults perform such segmentation even for separating stems and affixes …

TO CONTINUE: the grammatical units are & generalization when shown systematic paradigms & stimuli where word boundaries are salient). mention L2 learning, decomposition, & maybe production

TODO (explain that what is missing is crossing of the 2 lits: formation of a "long term lexicon" via segmentation from fluent speech as in first line of work, & segmentation of meaningful & grammatical parts as in second line of work) Stems, like words, are building blocks of language and the ability to extract them is fundamental. Nonetheless, it is still unclear whether humans focus on bigger (words) or smaller (stems) blocks, or whether they process both levels in a complementary way. Despite the wealth of evidence that stems may also be readily segmented, no previous study has looked at stem segmentation, even though the aforementioned literature strongly suggests that humans should process both words and stems in language learning. Consequently, it is still unclear whether humans can segment stems out of running speech in an artificial language, and whether, if exposed to a language with both words and stems, they would succeed better in segmenting one or the other. how do we learn subword units? is it only a secondary/deeper task? chunking/synthesis (REF) versus splitting/analysis (REF)? the role of items found in isolation? If we find that there is no difference in performance, perhaps there is no difference in representation: discussion regarding decomposition of items that are represented in the lexicon (Embick).

This study

The goal of this study is to inform questions concerning the relationship between language segmentation and language structure, by investigating segmentation in two meaningful unit levels, words and stems, among adults exposed to artificial languages. Participants hear sentences of an artificial language, where words contain two or three morphemes. Their preference for words versus non-words and for stems versus non-stems will then be measured. The languages were designed to resemble natural languages such that content words are composed by stems and affixes. For the same reason, words were embedded in sentence frames, with more than one word each sentence. Transitional probabilities between syllables are controlled for, as explained in the Methods section. QUESTIONS AND PREDICTIONS MAKE REF TO 2AFC, EXPLAIN One crucial difference with respect to previous studies is that participants are not based on transitional probabilities or other such cues (e.g. length, frequency) to choose the correct answer, as these cues are controlled for (both words/stems and non-words/non-stems in each test trial have the same cues)..

Specifically, three key questions and predictions addressed in this study are:

1. Do participants segment out whole words in a language where there are affixes? If participants segment out words, then they will choose words more than non-words – items that are not structurally words or stems. We predict that participants will do so, because word segmentation has been successful in previous artificial language studies where participants were presented with full sentences and because participants strip affixes in structure learning studies. However, it is possible that participants will fail to do so in our experiment, which combines the difficulty of complex morphology with passive exposure to unsegmented sentences. Should this occur, we can run a control in which participants are exposed to an artificial language that does not have systematically complex words.

2. Do participants segment out stems? If participants segment out stems, then they will choose stems more than non-stems – items that are not structurally words or morphemes. We predict that participants will do so, because of the structure learning studies summarized above. However, if participants do not show a preference for stems over foils in this study, but they do show a preference for words over matched foils, this may indicate that stem segmentation occurs later and/or requires more exposure and/or requires more active manipulation of the language material. Before concluding so, we should consider our next question.

3. Which units, words or stems, are better segmented? To test this question, we will compare the two preference tests mentioned above in terms of preference. If results are stronger for words, then this may provide initial support for "analysis" theories; if results are stronger for stems than words, then this would be consistent with "synthesis" theories. It is also possible that there is no difference between the two, which would be consistent with the idea that words and stems are processed similarly.

The study is preregistered in OSF: https://osf.io/fuydc. We report fully on the design and planned analyses, and report results for one pilot.

Methods

All materials and scripts can be found in
https://osf.io/pzcs2/?view_only=4167d91834b34ba6bc84618c294a35bc.

Stimuli.   Each participant is assigned to one of four artificial languages, A, B, C or D. The languages were generated using different orders of concatenation of 18 syllables that are meaningless in French ("glu," "sin," "ga," "kli," "ten," "ko," "blu," "tun," "man," "blo," "ti," "gle" ,"da," "pun," "go," "kan," "fen" and "bi"). Each syllable only has one use in the vocabulary. A counterbalancing procedure was used to create the languages, as follows.

Language A: Syllables "glu," "sin," "ga," "kli" and "ten" form three noun stems, the

169 next five syllables "ko," "blu," "tun," "man" and "blo" form three verb stems, syllables

170 "ti," "gle" form the optional elements, "da," "pun" form the noun and verb singular, "go,"

171 "kan" form the plural affixes and "fen" and "bi" form the aspect affixes.

172 Language B was generated by inverting syllable order: syllables "bi," "fen,""kan,"

173 "go," "pun" were used for the noun stems, "da," "gle,""ti," "blo,""man" for the verb stems,

174 ""tun," "blu" for the optional elements, "ko," "ten" for the noun and verb singular, "kli,"

175 "ga" for plural and "sin," "glu" for aspect).

176 Language C was created by inverting the order of syllables from the middle to the

177 beginning, and then from the end to the middle: "blo," "man," "tun," "blu," "ko" form the

178 noun stems, "ten," "kli," "ga," "sin," "glu" the verb stems, "bi" and "fen" the optional

179 elements, "kan" and "go" the singular, "pun" and "da" the plural, and "gle" and "ti" the

180 aspect.

181 Language D was created by starting from the middle to the end, and then from the

182 beginning to the middle: "blo," "ti," "gle," "da," "pun" were used for noun stems, "go,"

183 "kan," "fen," "bi," "glu" for verb stems, "sin," "ga" for optional elements, "kli," "ten" for

184 singular, "ko," "blu" for plural and "tun," "man" for aspect.

185 TODO ADD FIGURE 1 CAPTION: Utterance structure

186 Language elements were assigned to meanings, such that in each language one noun

187 stem meant "dog," another "cat," another "sheep"; one verb stem meant "walk" , another

188 "jump," another "flip." One noun affix indicated singular, another plural. One verb affix

189 indicated singular, another plural, another a continuous action that extends over time and

190 another an instantaneous action. At this point, then, items are akin to dictionary entries,

191 having a phonological form and a meaning.

192 These items were then combined into sentences. The structure of each sentence is

193 portrayed in Figure 1. Each sentence consists of a noun stem with its noun number affix

194 and of a verb stem with an aspect affix and a verb number affix. The sentence can have

one optional element at the beginning, or a different optional element at the end (akin to an adverb because it can occur at the beginning or the end, but having no meaning).

There were 96 sentences in each language generated for the training phase. The full compositionality is 3 noun one-or-two syllable stems x 2 noun number affixes x 3 verb one-or-two syllable stems x 2 aspect affixes x 2 number affixes. The same sentences may appear with no disrupters, one opening disrupter or one closing disrupter. Some sentences are held out, i.e. are logically possible but not presented. This is so that it can allow for certain requirements to be respected during the test phase, mainly item frequency and transitional probabilities between syllables, as will be presented in the next paragraph. In addition, we generated 14 'false' items for the test phase, to be paired with 'correct' stem or whole word items. A correct noun word is a noun stem with an affix indicating number. A correct verb word is a verb stem with an affix indicating aspect and an affix indicating number. Stems are presented bare. The false items include fragments of sentences heard during the train phase, which do not form a stem or an entire word. These fragments may be a verb stem or the second syllable of a verb stem paired with only an aspect affix, the second syllable of a noun stem paired with a noun number affix, a noun number affix paired with a verb stem, or a noun number affix paired with a verb stem and an aspect affix. In total, the test phase consists of 30 paired-forced-choice trials. Fourteen trials have word and non-word pairs (six nouns versus non-nouns + eight verbs versus non-verbs). Sixteen trials have stem and non-stem pairs (eight noun stems versus non-noun stems + eight verb stems versus non-verb stems).

One requirement is having the same sum of forward transitional probabilities (TPs) between syllables of word and non-word items, as well as stem and non-stem items that will be paired in a test trial. Statistical learning using TPs is a primary source of evidence for segmentation in laboratory experiments for infants (Estes & Lew-Williams, 2015; R. L. Gomez, 2012; Pelucchi et al., 2009; Romberg & Saffran, 2010) and adults (Frank et al., 2010; Perruchet & Desaulty, 2008; Toro et al., 2005). In those studies, participants segment

222 out words based on TPs.

223   However, since in this experiment we do not test the well-established use of TPs in

224 segmentation, the TP information should be controlled for in the test. To this end, we

225 adjusted the combination of stems and affixes during training. Stems have TP=1. When

226 comparing a stem to a non-stem, some meaningless pair of syllables appearing next to each

227 other should also have TP=1. This happens by presenting some stem always with the same

228 affix. For example, the noun stem ('sheep') appears systematically in plural, and the verb

229 stem ('flip') appears systematically with a non-progressive affix. For these stems, TP

230 between stem and affix equals 1. One test example presents the participant with a choice

231 between the correct verb "walk" (one-syllable verb stem + progressive affix + verb plural

232 affix) and a false item: plural noun affix + one-syllable verb stem + progressive affix. Both

233 choices are equally frequent in the training set, and TPs between syllables are 0.5. Another

234 test includes a choice between the noun stem "sheep" (including a two-syllable noun stem)

235 and a false choice, the second syllable of the noun stem and a noun plural affix. Both

236 choices are equally frequent in the training set, and TPs between syllables are 1.

237   The text was converted to speech using the mac Speech Synthesis tool. Specifically,

238 the Mexican voice "Paulina" was used because the voice has a flat prosody to avoid

239 accidental insertion of prosodic cues. The speed of speech was fixed to 140 words per

240 minute after pilot listening revealed that the default speed for this voice was very slow.

241   For the training phase, the audio sentences were paired with videos of scenes, which

242 represent the meaning of the sentence. For example, if a sentence contains a 'dog' noun

243 stem, and a 'walk' verb stem, the video would show a dog walking. The noun stem is

244 followed by a number suffix, and the video shows one or two dogs. The verb stem is

245 followed by an aspect suffix, and the video shows a dog walking continuously or once.

246 Scenes of the videos are pictured in Figure 2. Videos are not shown during the test phase,

247 and stimuli are presented purely auditorily.

₂₄₈   TODO ADD FIGURE Figure 2. Video samples from the study. In the first video, a

₂₄₉ cat walks. In the second video, two cats are walking.

₂₅₀   **Procedure.**   The experiment was administered online, using the Labvanced

₂₅₁ Platform. There are two phases, which we call training and test. The training stimuli

₂₅₂ (audiovideos) are presented as utterances with clearly marked utterance boundaries, as

₂₅₃ cued both by silence and the end of a scene. The participant listens to a sentence while

₂₅₄ watching a video portraying the action described in the sentence. Participants are not given

₂₅₅ the sentence in writing, nor any breakdown of the different words, nor stems versus suffixes.

₂₅₆   During the training period, we ensure attention through the use of catch trials. Eight

₂₅₇ questions testing the attention of the participant were added throughout the training

₂₅₈ period. Four questions ask which animal was seen in the last video. Three options are

₂₅₉ proposed that the participant has to choose from: a dog, a cat and a sheep. Four more

₂₆₀ questions ask which action was performed by the animal in the last video. Three options

₂₆₁ are proposed: walk, jump and flip.

₂₆₂   Once the training period is over, participants are asked the following: 'Vous allez

₂₆₃ maintenant entendre 2 sons. Quel son ressemble le plus à la langue que vous venez

₂₆₄ d'entendre? Ne réfléchissez pas trop et allez-y avec votre instinct!" ("You are now going to

₂₆₅ hear two sounds, which one sounds better for the language you just heard? Don't overthink

₂₆₆ it and follow your gut"). They then need to press a button after hearing each trial: the left

₂₆₇ button if they think that the first item works better, or the right button for the second

₂₆₈ sound. Participants first perform a short mock test demonstrating the test process. They

₂₆₉ have to choose between pairs of french words, were one word is very frequent, and the other

₂₇₀ one is very rare.[2] The participants are asked which sound resembles more to the French

₂₇₁ language, and they can continue with the real test once they have answered correctly

––––––

[2] The frequencies were retrieved from http://www.lexique.org/, a database of 140,000 French words and
their information collected from diverse corpora (New et al, 2007).

272 (selected the most frequent word) in two successive trials. They then perform the test

273 trials.

274   At each trial, the participant only hears the stimuli, and thus there is no video. One

275 can say that each trial presents them with a "correct" option and a foil. The correct option

276 is a correctly segmented word or stem; the foil is an item that spans a word and/or stem

277 boundary. Each test item has at least 2 syllables, to avoid missing or mishearing a (very

278 short) sound. Paired items have the same length, but also result in the same sum of

279 (forward) Transitional Probabilities, and have the same frequency (or, rarely, the foil has a

280 larger frequency than the correct option). This way, no other cue could affect the

281 preference of one versus the other item, other than the preference for a correctly segmented

282 unit versus a non-unit. The whole experiment lasts 25 minutes. The study was initially

283 designed to be tested in the lab, however, due to the current sanitary situation, it was

284 updated for online testing.

285   Prior to the onset of the COVID-19 crisis, two pilot studies were conducted, each

286 with eleven participants. Participants in the first pilot brought up three issues; the

287 successive presentation of similar sounds during the test phase, the length of the

288 familiarization period, and the absence of marked difference in verb aspect in the videos,

289 e.g. they considered the 'walk' and 'walking' videos as describing the exact same action.

290 These issues were successfully addressed in a second pilot. First, the repetition of the same

291 stimulus or of a stimulus with the same stems is now avoided in successive trials. Second,

292 an engaging message to the participant appears in the screen after completing 25%, 50%

293 and 75% of the training, congratulating them for completing the corresponding part of the

294 study. Third, the duration of the action in the videos with progressive actions was

295 increased, and a "narrator" figure was added next to the videos.

296   The online version of the experiment was tested by four participants. Based on their

297 feedback, the following two extra steps were taken when transferring the experiment online.

298 First, eight questions testing the attention of the participant were added throughout the

299 training period. Four questions ask which animal was seen in the last video. Three options

300 are proposed that the participant has to choose from: a dog, a cat and a sheep. Four more

301 questions ask which action was performed by the animal in the last video. Three options

302 are proposed: walk, jump and flip. No feedback is given on the correctness of the answer.

303 Second, right before the test section, the participants take a short mock test demonstrating

304 the test process. They have to choose between pairs of french words, were one word is very

305 frequent, and the other one is very rare. The participants are asked which sound resembles

306 more to the French language, and they can continue with the real test, once they have

307 answered correctly (selected the most frequent word) in two successive trials.

308 **Participants.** There were 52 participants (XX male), aged XX years (range

309 XX-XX), and having French as their native language based on self report. Participants

310 were recruited with the Prolific crowd-sourcing platform (www.prolific.co) and the RISC

311 participant platform (www.risc.cnrs.fr/). They were paid 10 euros for their participation.

312 We estimated that 52 participants would be sufficient, given that previous similar

313 studies using artificial language segmentation experiments with adults found and average

314 effect size of 0.46 in terms of the word versus foil two-alternative forced choice (Cunillera et

315 al., 2010; Frost & Monaghan, 2016; Hoch et al., 2013; Perruchet & Desaulty, 2008; Toro et

316 al., 2005; Tyler & Cutler, 2009 - a table with effect sizes and their average can be found in

317 Supplementary material).

318 Eligible for inclusion were participants 1. who completed the experiment (meaning

319 that they answered all test questions), 2. who answered correctly at least 4 of the attention

320 questions, and their proportion of correct answers is not lower than two standard

321 deviations from the mean group proportion, 3. whose average decision time during test was

322 not lower than two standard deviations from the group average. Given these exclusion

323 criteria, we continued recruitment until we achieved the required size, replacing

324 participants with matching conditions as needed.

Statistical Analysis

For the analysis, we will fit a generalized linear mixed effect model using the lme4 library in R (R studio team, 2015). The participants' answers (right/wrong) will be the dependent variable and the level (stem/word) will be a fixed variable. The random effect of the participant is included, with level and number of trial (1st, 2nd…) as random slopes.

Pilot Data

Participants of the second pilot adapted well to the particularities of the complex linguistic structure, and considered both minimal meaningful units and words when listening to this unknown language. On average, 64% of their responses on word recognition (SD=0.186) and 61% of their responses on stem recognition (SD=0.16) were correct.

With respect to the three key questions, we observed that: 1. based on the intercept in a regression with the word level as baseline, the word trial coefficient was 0.62 (SE=0.28), with p-value=0.024. 2. based on the intercept of a regression with the stem level as baseline, the stem trial coefficient was 0.47 (SE=0.342), with p-value=0.003. 3. Based on the level as a fixed effect in a regression with the stem level as baseline, the coefficient for level is 0.162 (SE=0.248, p-value=0.540). Thus, recognition of words versus stems is not significantly different in this pilot. The trend is for more accurate responses for word trials than stem trials.

In a further exploratory analysis, we asked whether there is a difference in segmentation between verbs and nouns. The two word classes may not be overall processed in a similar way. Verbs could be more difficult to segment than nouns, for example due to the presence of more morphemes per word on average. We investigated this by including type (noun /verb) as a fixed effect variable and as a random slope. In the pilot results, the type coefficient was -0.33 (SE=0.239, p-value=0.169). The trend was for more accurate responses for nouns than verbs, but the result was non-significant.

The pilot results establish the feasibility of the proposed experiment.

351 **Results**

352 **Discussion**

Acknowledgments

Data, code and materials availability statement

All data, code, and materials are available from https://osf.io/5qspb/

References