Evaluating the relative importance of wordhood cues using statistical learning

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13429

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13429

Evaluating the Relative Importance of Wordhood Cues Using Statistical Learning

Elizabeth Pankratz, Simon Kirby, Jennifer Culbertson

Centre for Language Evolution, Department of Linguistics and English Language, University of Edinburgh

Received 17 July 2023; received in revised form 22 January 2024; accepted 27 February 2024

Abstract

Identifying wordlike units in language is typically done by applying a battery of criteria, though how to weight these criteria with respect to one another is currently unknown. We address this question by investigating whether certain criteria are also used as cues for learning an artificial language—if they are, then perhaps they can be relied on more as trustworthy top-down diagnostics. The two criteria for grammatical wordhood that we consider are a unit's free mobility and its internal immutability. These criteria also map to two cognitive mechanisms that could underlie successful statistical learning: learners might orient themselves around the low transitional probabilities at unit boundaries, or they might seek chunks with high internal transitional probabilities. We find that each criterion has its own facilitatory effect, and learning is best where they both align. This supports the battery-of-criteria approach to diagnosing wordhood, and also suggests that the mechanism behind statistical learning may not be a question of either/or; perhaps the two mechanisms do not compete, but mutually reinforce one another.

Keywords: Wordhood criteria; Statistical learning; Sequence learning; Distributional cues

1. Introduction

How to reliably identify the span of a "word" has provoked much debate (e.g., Dixon & Aikhenvald, 2002a; Haspelmath, 2011; Julien, 2007; Tallman, 2020; Wray, 2015). Typically, a unit is called a "word" if it meets a number of grammatical and phonological criteria (Dixon,

Correspondence should be sent to Elizabeth Pankratz, Centre for Language Evolution, Department of Linguistics and English Language, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD. E-mail: ecpankratz@gmail.com

2009). For example, some criteria ask: is the unit able to move around freely and appear in different environments? Do its component parts always occur in the same order without interruption? Is the unit phonotactically legal? Are pauses permitted within the unit or only at its boundaries?

Such wordhood criteria are normally fairly unanimous in languages like English—languages in which word-internal structure is relatively sparse. In languages with more morphology, though, the criteria often disagree about whether a span should be considered a "word" (Haspelmath, 2011; Tallman, 2020). This raises the question: how should we adjudicate between different criteria? Which ones should be relied on more, if any?

We propose that one way to identify which criteria should be preferred as diagnostic tools is to take into account the role they play in learning. Typically, the criteria are applied top-down to determine the status of a unit whose span is already known (Julien, 2007)—but if learners also use the same cues bottom-up to identify units from scratch, then this could indicate that they are cognitively important, and this is worth considering in an empirically oriented linguistic theory.

For example, studies like Shukla, Nespor, and Mehler (2007); Endress and Mehler (2009); Endress and Hauser (2010) have found that phonological cues (e.g., prosody, between-word pauses) tend to be relied on for learning wordlike units more than distributional cues are. This suggests that perhaps phonological criteria should be the descriptive linguist's first port of call. To flesh out the distributional side of the picture, in this study, we offer a novel demonstration of the relative reliability and cognitive importance of two common criteria for grammatical wordhood: mobility and internal immutability.

This cognitive approach offers a different way to tackle the question of whether "words" are even a sensible thing to try to identify. A "word" seems primarily to be a unit of orthography (Haspelmath, 2011; Julien, 2007; Tallman, 2020; Wray, 2015), so is it futile to study wordhood outside of the written medium?

One view of wordhood criteria is that they are actually aiming to identify the minimum free forms of language that are represented by language users (Anderson, 1992; Bloomfield, 1933), which do exist outside of writing and may or may not overlap with the orthographic word (Wray, 2002). This approach thus aims to sidestep the ontological debate about "words" and instead focus on evaluating the cues learners use to identify recombinable free units in their language.

Here, we study two of the primary grammatical wordhood criteria proposed in the literature and the role they play for learning multisyllabic units in an artificial language. We do this using a widely used method—statistical learning experiments—which have shown that learners who encounter a continuous stream of input (be it auditory, visual, or tactile) are able to identify the recurring wordlike units within it (Saffran, Aslin, & Newport, 1996a; Saffran, Newport & Aslin, 1996b). We engineer the input in our experiment such that the two criteria point to different spans of syllables, and then test which spans are better learned.

1.1. Statistical properties of wordhood criteria

Of the many possible criteria for identifying words (see, e.g., Anderson, 1992; Aronoff, Meir, Padden & Sandler, 2004; Bickel, Hildebrandt & Schiering, 2009; Bloomfield, 1933;

Dixon & Aikhenvald, 2002a; Dixon, 2009; Julien, 2007; Mansfield, 2021; Reichling, 1935; Tallman & Epps, 2020; van Wyk, 1968), two arise again and again. For one, words can move around freely and occur in many different environments. This criterion, which we call "mobility," has also been called, inter alia, "positional mobility" (Saffran et al., 1996b), "syntagmatic mobility" (Dixon & Aikhenvald, 2002b; Reichling, 1935; van Wyk, 1968), "independent distribution" (Boas, 1911; Julien, 2007), and "flexible linear position" (Mansfield, 2021). For another, words have a consistent internal structure across contexts and occurrences. We call this criterion "internal immutability" (Reichling, 1935; van Wyk, 1968), and it covers criteria that have also been referred to as "uninterruptibility" (Langacker, 1972; Saffran et al., 1996b), "cohesiveness" (Dixon & Aikhenvald, 2002b), "fixed ordering" (Dixon & Aikhenvald, 2002b), "internal cohesion" (Bloomfield, 1933; Julien, 2007), and "internal stability" (Saffran et al., 1996b).

These are high-level theoretical criteria, but they have analogues in the low-level distributional properties of syllable sequences (Harris, 1954, 1955). The mobility criterion can be thought of in terms of the variability at word boundaries. Specifically, units that are mobile have less predictable transitions from their word-final syllable to other word-initial syllables. In canonical statistical learning terms, the mobility criterion corresponds to word boundaries having low transitional probabilities (TPs). (A TP is a conditional probability: given an observed syllable, what is the probability that a particular syllable will then follow?) The internal immutability criterion also has an analogue in these terms: it can be understood as saying that syllables within words have completely predictable transitions. Put differently, word-internal TPs are high (e.g., equal to 1).

These two properties fall together in typical statistical learning experiments (Isbilen & Christiansen, 2022), but by including an analogue to morphological structure in the languages we test (see Section 2), we can tease them apart.

1.2. Statistical learning

In the tradition of Saffran et al. (1996a, 1996b), statistical learning experiments test the role of distributional cues for the learning and segmentation of continuous input. In their seminal experiments, Saffran and colleagues showed that both adult and infant participants were able to identify syllable triplets—that is, sequences of three syllables—that recurred within the continuous stream of input. Over the last decades, many hundreds of experiments have used this paradigm to replicate and extend this result (for recent reviews, see Frost, Armstrong & Christiansen, 2019; Siegelman, 2020, and Isbilen & Christiansen, 2022).

Most of those experiments use the same language structure as Saffran and colleagues: several different triplets (of syllables, images, etc.) that are repeated, concatenated, and shuffled to produce the input stream (Isbilen & Christiansen, 2022). For languages with this structure, the TPs between triplets are always low, and the TPs within triplets are always high. Thus, the same spans are identified by both distributional cues: both low TPs at their boundaries (corresponding to the mobility criterion discussed above), and high TPs within them (corresponding to internal immutability).

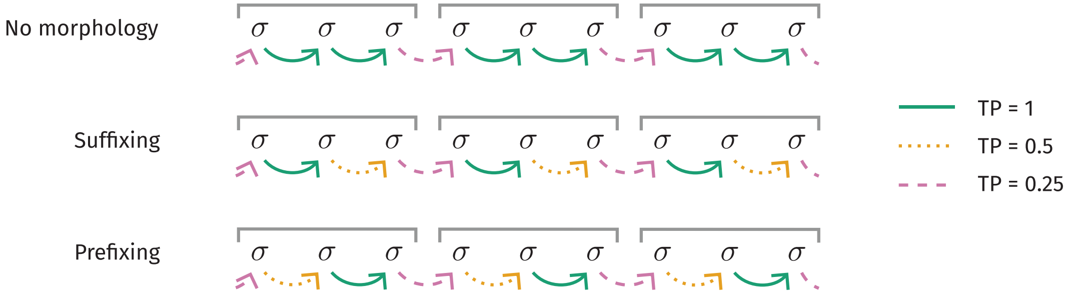

Fig. 1. Three artificial languages with different morphological structures and, therefore, different TP patterns between syllables σ. Brackets span the syllable triplets, green arrows indicate TPs of "stems" (high at TP = 1), dotted orange lines indicate TPs of "affixes" (medium at TP = 0.5), and dashed pink lines indicate TPs of "word" boundaries (low at TP = 0.25).

While these distributional cues are confounded in many statistical learning experiments, there is in fact a long-standing debate about which of the two is the main source of information exploited by learners. The original proposal by Saffran and colleagues laid out an account based around low TPs, and suggested that learners use the variability at unit boundaries to identify where the units are to be separated (see also, e.g., Gervain & Guevara Erra, 2012). However, support has also been growing for an account which argues that learners pay attention to those units that co-occur most frequently or reliably, that is, those with high internal TPs. This is usually referred to as the chunking account (for theoretical support, see Christiansen & Chater, 2016; Isbilen & Christiansen, 2020, and for modeling work, see Goldwater, Griffiths & Johnson, 2009; Orbán, Fiser, Aslin & Lengyel, 2008; Frank, Goldwater, Griffiths & Tenenbaum, 2010). These two accounts—low TPs and chunking—correspond to the mobility and internal immutability criteria, respectively.

2. The experiment

2.1. Design: Three language structures

We test the learning of three different language structures, schematized in Fig. 1. The first language has the most typical format: it consists of syllable triplets with no internal structure. We call this the "no morphology" language. The second language is the "suffixing" language: all syllable triplets consist of a two-syllable stem followed by a one-syllable affix. Finally, in the "prefixing" language, all syllable triplets begin with the one-syllable affix, followed by the two-syllable stem. The TPs at triplet boundaries are 0.25; at morpheme boundaries, 0.5; and otherwise 1.

If people learn units based on their mobility (similar to the TP account), then in all languages, triplets should be learned better than the pairs of syllables they contain. This is because the triplets are bounded on both sides by the lowest TPs of 0.25, while the syllable pairs are only bounded on one side or the other by this low TP. And if people learn units based on their internal immutability (similar to the chunking account), then sequences that do not span a morpheme boundary, that is, those with TPs of 1, should be learned better than sequences containing a nondeterministic transition. This means that in the no-morphology language, triplets should again be learned better, while in both languages with morphology, the syllable pairs that constitute the "stems" should come out on top.

The use of both suffixing and prefixing languages also allows us to consider the influence of a possible suffixing preference on learning (Hawkins & Cutler, 1988; Himmelmann, 2014; Martin & Culbertson, 2020). For example, our English-speaking participants may find the suffixing language easier to learn, since it has the same morphological structure as English (see, e.g., Martin & Culbertson, 2020). If this is the case, then we might observe greater overall accuracy on the suffixing language than on the prefixing language. We may also observe that participants are differently sensitive to the nondeterministic transitions in these two languages: perhaps participants are able to identify and learn the two-syllable stems in the suffixing language better than in the prefixing language, since they are accustomed to segmenting words this way. Our statistical modeling will estimate both of these possible effects.

2.2. *Materials*

2.2.1. *The language*

We present the languages both visually and auditorily (see Section 2.3). Each language draws on the same set of syllables: *bu* [bu], *cu* [ku], *da* [da], *je* [ʒe], *ko* [ko], *lu* [lu], *mu* [mu], *qe* [ke], *qi* [ki], *ru* [ru], *vu* [vu], and *xo* [zo]. Some of these syllables appear non-English (e.g., *qe*, *xo*, *je*), but to ensure their auditory perceptibility, we mapped them to phonemes familiar to English speakers. These syllables create the triplets shown below:[1]

- No morphology: *kodaqe*, *qimuvu*, *buxoje*, *ruculu*
- Suffixing: *buxo-lu*, *buxo-je*, *rucu-lu*, *rucu-je*, *qimu-qe*, *qimu-vu*, *koda-qe*, *koda-vu*
- Prefixing: *cu-xoje*, *bu-xoje*, *cu-qimu*, *bu-qimu*, *vu-kolu*, *ru-kolu*, *ru-daqe*, *vu-daqe*

Note that there are four triplets in the no-morphology language and eight in the languages with morphology. In the latter case, the languages have four stems and four affixes, just in different combinations. Importantly, the number of appearances of the stems and affixes is the same overall as the number of appearances of the triplets in the no-morphology language.

2.2.2. *The input stream*

In typical statistical learning experiments, the underlying TPs between triplets are not necessarily reflected in the observed transitional frequencies, since the input streams are assembled randomly. At the population level, of course, proportions of observed transitions should approach the TPs, but for individual participants, this may not be the case. For our study, though, we wanted every participant's observed proportion of transitions between syllables to match the TPs exactly. This is important because this property of the input maps to the mobility criterion, and it is unclear what conclusions could be drawn if every participant received slightly different information about how variable the triplet boundaries are. We, therefore, created input streams such that the proportion of transitions between all syllables precisely match the underlying TPs. To do this, we had to diverge in two ways from the customary method of creating input streams.

For one, typically each triplet is repeated the same number of times. In a language with four triplets repeated 24 times each, say, the input stream is 96 triplets long, but it contains only 95 transitions—one transition between a triplet-final and a triplet-initial syllable would be observed less than all the others. To yield 96 transitions in which every triplet-final syllable can transition to every triplet-initial syllable six times, our input stream contains 97 triplets. [2]

Another way that our input stream differs from previous work is that we must permit a triplet to be immediately followed by itself (in contrast to, e.g., Bogaerts et al., 2016; Elazar et al., 2022; Endress & Mehler, 2009; Saffran et al., 1996a, 1996b; Siegelman, Bogaerts & Frost, 2017). If we did impose the nonrepetition constraint, then the between-triplet TPs would vary across language structures. For example, in the no-morphology language, the final syllable *qe* in *kodaqe* could only transition to the three other initial syllables and not back to *ko*, yielding TPs of 0.33. But in the prefixing language, say, the final syllable *je* in *cu-xoje* could still transition to *cu* 1/7th of the time (since there are two triplets with the *cu* prefix), and each of the other three prefix syllables 2/7ths of the time. To avoid this, we opted not to implement this constraint on randomization.

All in all, our input streams consist of 291 syllables, forming 97 triplets. In the no-morphology language, each of the four triplets is repeated 24 times, and then the first triplet of the input stream is repeated once more at the end. In the suffixing and prefixing languages, each of the eight triplets is repeated 12 times (plus a repetition of the first triplet of the stream at the end); this works out to every stem and every suffix being seen 24 times (or 25, for one stem-suffix pair), like the no-morphology triplets. All triplets appear in a quasi-random order that satisfies the constraint that every triplet transitions to every other one exactly six times.

2.2.3. The audio

The audio was created using the speech synthesizer eSpeak (Duddington, 2012). We individually synthesized each syllable with eSpeak's default fundamental frequency of 82 Hz and the synthesizer's Kurdish voice. Syllable durations range from 311 to 396 ms, similar to, for example, the 250–350 ms of Siegelman et al. (2018). To create the multisyllabic audio for the test phase, we concatenated the existing audio files of the individual syllables.

2.3. Procedure

The experiment consists of a familiarization phase, followed by a testing phase. In the familiarization phase, participants are presented both visually and auditorily with a continuous stream of 291 syllables (see Section 2.2.2 for more details). Each syllable is shown on screen, and its audio is simultaneously played. Once the audio finishes, the next syllable is shown and its audio is played. The familiarization phase lasts for just under 2 min. The motivation for showing both auditory and visual information simultaneously (unlike many comparable experiments that use either audio or visual material, but not both), and also for keeping the familiarization phase fairly brief (compared to, e.g., the 3 min of Siegelman et al., 2018 and the 7 min of Elazar et al., 2022), is because this part of the experiment is very passive, and we wanted to keep our online participants from losing focus.

Following familiarization, the test phase requires participants to make a series of two-alternative forced choice judgments between sequences they saw in the input stream ("targets") and unseen sequences created by shuffling the language's syllables into an order that never appeared in the input stream ("foils").

We tested participants both on syllable triplets and on syllable pairs. The four target triplets were observed triplets in the language, and the eight target pairs were the two syllable bigrams that each target triplet contains. For example, in the suffixing language, example targets would be the triplet *buxolu* and the pairs *buxo* and *xolu* (the former pair is the stem; the latter spans the morpheme boundary). For the suffixing and prefixing languages, we selected one triplet with each of the four stems, such that each stem and each affix were tested once.

The foils are syllable sequences that never appeared in any input stream, for example, *kojemu*, *buqe*, and *jeda*. The same foils were used for all three languages.

In each trial, the order of target and foil are randomized; one is presented alone on screen with concurrent audio, then the other joins it, and then buttons appear below for participants to make their choice (similar to the procedure in, e.g., Siegelman et al., 2018). Across the whole testing phase, each target and each foil were repeated twice (but each target–foil pair appeared only once), yielding 24 trials.

2.4. Participants

We recruited 150 participants from Prolific's pool of self-reported native English speakers resident in the United Kingdom. Fifty participants were allocated to each language. They were each paid £2 for their participation.

This experiment was approved by the University of Edinburgh's PPLS ethics committee (ref. 259-2122/2).

3. Results and analysis

Fig. 2 shows participants' accuracy in each of the three language structures, for syllable triplets and pairs. Many participants are clustered around chance level, suggesting that the task was challenging and that many people did not learn much—a result shared with many other statistical learning experiments (Siegelman et al., 2017). But considering the data on aggregate, accuracy is still greater than chance (see below), and we can also identify some trends. For example, performance is best overall in the no-morphology language when participants are tested on triplets (top left in Fig. 2); this is striking because it is the standard setup of most statistical learning experiments, and also the condition in which both of the cues for wordhood line up. Specifically, triplets in this language are groups of syllables surrounded by low TPs—that is, high mobility—which themselves have high TPs—that is, high internal immutability.

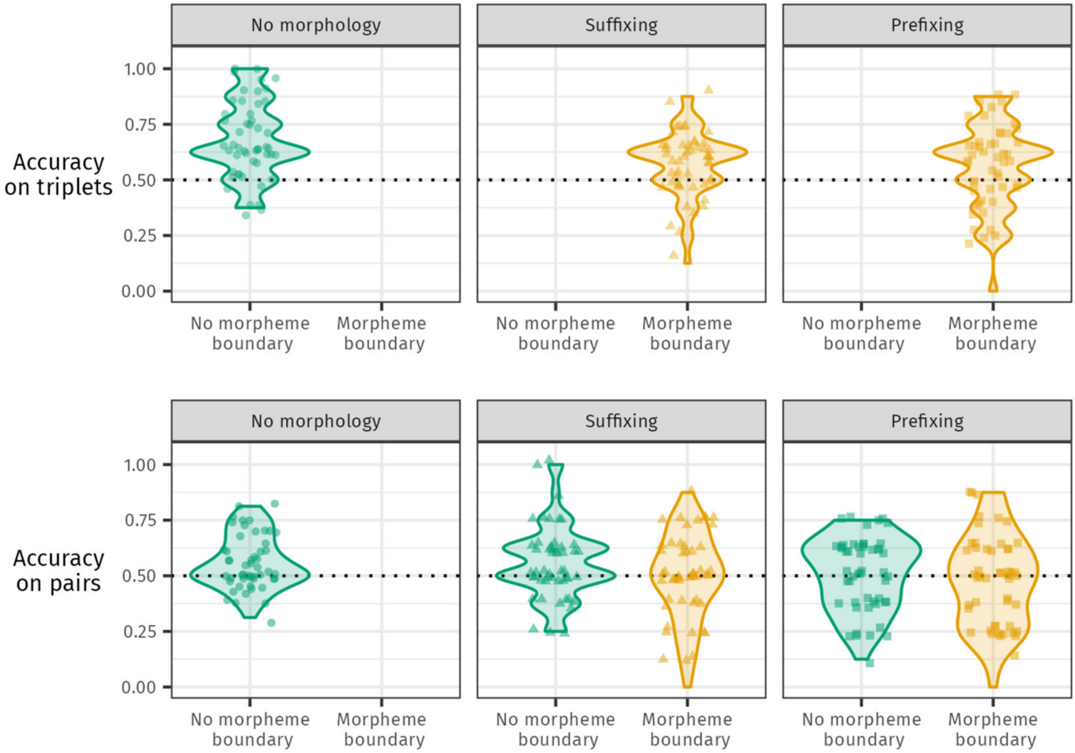

Fig. 2. Each participant's mean accuracy for different kinds of targets (triplets vs. pairs; targets with and without morpheme boundaries) in all three languages. Blank spots reflect combinations of variables that did not exist in the given language (e.g., the suffixing language does not contain any triplets without a morpheme boundary). The horizontal dotted line indicates chance. Accuracy is highest overall for triplets in the language without morphology, where both wordhood cues align.

We used brms in R (Bürkner, 2017; R Core Team, 2023) to fit a Bayesian linear regression model with a Bernoulli likelihood and a logit link function to these data. The model estimates the log-odds of a correct response based on the following predictors: language structure (no morphology/suffixing/prefixing), the length of the target item (triplet/pair), whether the target contains a morpheme boundary (present/absent), and an interaction between affix type (suffixing/prefixing) and morpheme boundary. The language structure variable is Helmert-coded, so that we can compare the no-morphology language (coded as 2/3) to the mean of both languages with morphology (–1/3), as well as the suffixing language (coded as 1/2) to the prefixing language (–1/2). Target length was sum-coded (triplet as 1/2 and pair as –1/2), as was morpheme boundary (a present boundary as 1/2, and an absent one as –1/2). The interaction between suffixing versus prefixing and morpheme boundary was also coded using ±1/2. We included a continuous predictor containing the centered log frequencies of the syllable transition within the target (see Footnote 1), and for triplet targets, we used the higher-frequency transition of the two they contain. Finally, the model contained by-participant adjustments to

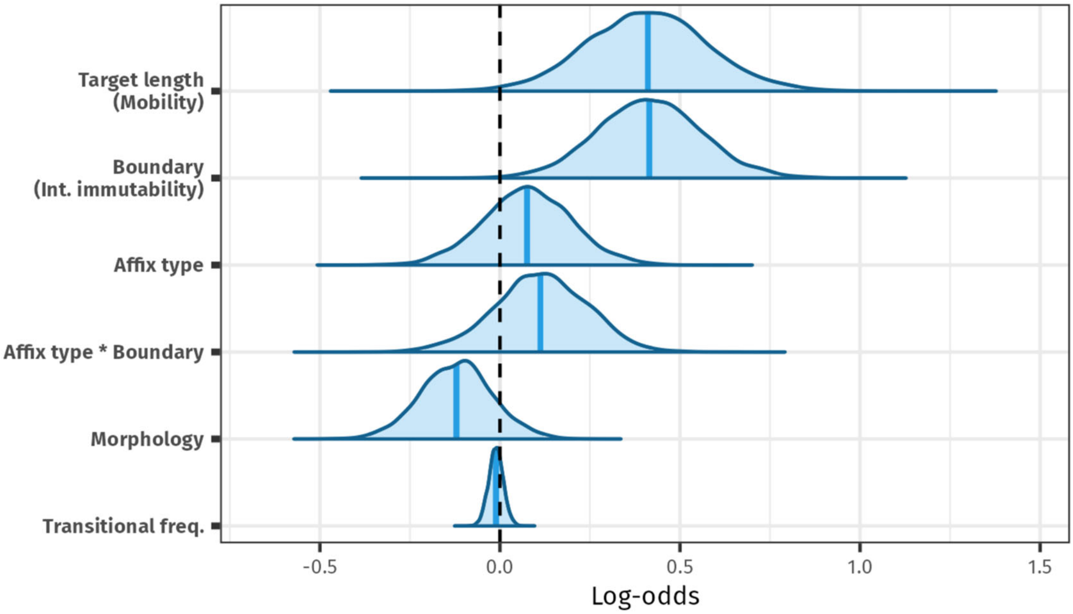

Fig. 3. Posterior distributions of the model's slope coefficients. Blue vertical lines represent posterior means, and the shaded regions cover the 95% CrI. Only for the effects of Target length and Boundary are the CrIs fully on one side of zero; the model is 95% certain that both of those predictors have a positive association with accuracy, while for all the others, it is uncertain about their association with the outcome. The predictor Target length corresponds to the mobility criterion, and Boundary corresponds to internal immutability.

Table 1
Posterior means and 95% CrIs for the fixed effects, given in log-odds space

	Posterior mean	95% CrI (lower)	95% CrI (upper)
Intercept	0.32	0.12	0.53
Target length (Mobility)	0.41	0.00	0.82
Boundary (Int. immutability)	0.41	0.06	0.77
Affix type	0.08	−0.23	0.37
Affix type * Boundary	0.11	−0.21	0.42
Morphology	−0.12	−0.36	0.12
Transitional freq.	−0.01	−0.06	0.04

the intercept and to the slopes of target length and morpheme boundary, as well as by-target adjustments to the intercept. Weakly regularizing priors were chosen using prior predictive checks. The model converged, as indicated by all Rhats = 1.00.

The model's posteriors are visualized in Fig. 3 and summarized in Table 1. The model's intercept reflects the grand mean, that is, the overall log-odds of a correct response. With its 95% Credible Interval spanning [0.12, 0.53], the model indicates that with 95% certainty, overall accuracy is (slightly) greater than chance. (A log-odds of 0 is equivalent to a probability of 0.5, the chance level for a two-alternative forced choice task; all of the values within the

95% CrI are above zero, indicating that the most probable values for the intercept parameter are all above chance.)[3]

To consider the roles of the wordhood criteria of mobility and internal immutability, we now focus on the posteriors for the two parameters Target length and Boundary, respectively.

The mobility criterion would predict that the mobile units, that is, those with high variability (low TPs) at either edge, are learned better than the sub-parts of those units. In our experiment, these corresponded to the triplet targets (see Fig. 1), which were contrasted with the pair targets in the variable Target length. The mean of this parameter's posterior is 0.41 log-odds (95% CrI: [0.00, 0.82]). That the 95% CrI comprises only positive values suggests that there is a 95% probability that this effect is positive, when all other predictors are at their means: in other words, we can be fairly certain that performance is better on triplets than on pairs. However, the CrI bordering on zero means that the model considers very small, possibly null, effects to also be plausible.

Slightly narrower, and thus slightly more certain, is the posterior distribution estimated for Boundary. This parameter corresponds to the internal immutability criterion, which would predict that units that are internally always the same, that is, contain no morpheme boundaries and have internal TPs of 1 (see Fig. 1), are better learned. The Boundary parameter's posterior mean is also estimated at 0.41 log-odds, but its 95% CrI spans only the range [0.06, 0.77]. We can be more certain that this effect is positive, when all other predictors are at their means: targets without morpheme boundaries—that is, internally immutable targets—are learned better than targets that contain a morpheme boundary. However, the effect is still quite small.

Turning to the possibility of a suffixing preference in these data, we consider the parameters estimated for Affix type and Affix type * Boundary. For one, if the suffixing language is learned better on the whole than the prefixing language, we would expect a positive effect of Affix type. Although there is slightly more posterior probability mass on the positive side of zero, there is also plenty on the negative side; the model, therefore, considers both positive and negative effects to be plausible. In other words, we cannot be certain in which direction this effect may go. The same holds for the interaction between Affix type and the presence of a morpheme boundary. All in all, the model's estimates are not consistent with a clear suffixing preference in this task.

Another uncertain estimate is produced for the effect of Morphology; the model did not find a clearly positive or negative difference in accuracy between the no-morphology language and the languages with morphology.

Finally, we note the model's extremely narrow posterior around zero for Transitional frequency. This indicates a high degree of certainty that targets that contained higher-frequency transitions in English corpus data were not responded to more accurately than low- or zero-frequency transitions.

4. Discussion

This study investigated the influence on learning of two wordhood criteria from the literature: mobility and internal immutability. We found that each criterion is individually likely to be positively associated with successful statistical learning, though their individual effects are fairly small—it is when they both line up that accuracy is highest. Because both criteria seem to serve as cues for learning, this supports their use as part of a diagnostic toolkit for identifying what language users may represent as their language's minimal free forms.

Interestingly, our model indicates that both cues are approximately equally facilitatory (though the posterior estimate of the role of internal immutability was slightly more certain). We, therefore, are not able to say that one is a stronger cue than the other, or that one should be "trusted" more than the other. However, the additive nature of these cues does support the common approach of diagnosing wordhood using a battery of criteria, rather than a single criterion.

As Haspelmath (2011, pp. 59–60) notes, the danger in this approach is when particular criteria are hand-picked to be able to tell a clearer story. Perhaps with further research along the lines of the present study, linguists could develop a more standardized, cognitively founded battery of tests. As mentioned above, due to the apparent primacy of phonological cues like prosody and between-word pauses (Endress & Mehler, 2009; Endress & Hauser, 2010; Shukla et al., 2007), those phonological criteria could be weighted most heavily, followed then equally by distributional cues like mobility and internal immutability.

4.1. The suffixing preference

A suffixing preference could have revealed itself in this experiment either with better performance overall on the suffixing language compared to the prefixing language, or with better identification and learning of stems in the suffixing language compared to the prefixing one. The model is quite uncertain about either of these effects. However, this analysis indicates that perhaps with more participants, the statistical learning methodology could be used to further study preferences for suffixing versus prefixing in users of different languages.

4.2. The mechanisms behind statistical learning

Above in Section 1.2, we drew a connection between the two wordhood criteria we test and two mechanisms that may underlie successful statistical learning. The mobility criterion appears to relate to an account of statistical learning in which learners seek low TPs at unit boundaries; the internal immutability criterion relates to learners "chunking" items together based on their frequent co-occurrence.

Interestingly, this experiment's results show near-equal support for these two accounts. Perhaps considering low TPs and chunking as compatible and mutually supporting mechanisms, rather than distinct and competing ones, might come closer to capturing how people succeed at statistical learning.

5. Conclusion

Many different criteria come into question when trying to diagnose whether a given span in a language counts as a "word." Here, we have suggested that for a criterion to be trusted as a top-down wordhood diagnostic, it should also be used as a bottom-up cue in learning. We have used the statistical learning paradigm to test the individual impact of two common criteria for grammatical wordhood, mobility and internal immutability, and we have found that they both influence learning in an additive way. This supports their use as equally weighted components of a diagnostic set of criteria. Further, the connection we draw between these wordhood criteria and the low-TP-based versus chunking accounts of statistical learning suggests that learners may use a combination of approaches, rather than one or the other, to learn and identify wordlike units in their language.

Acknowledgments

This work was supported by an ESRC award ES/P000681/1 (EP) and a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship 752-2021-0366 (EP). We thank the two anonymous reviewers for their helpful input.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/gfmz7/?view_only=a5e7a614f9d0490cab62cc739173d3f8.

Notes

1 Our goal with this set of syllables was to reduce the influence of participants' English L1 (Bogaerts, Siegelman, & Frost, 2016; Elazar et al., 2022; Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018). We intended to select syllables such that all transitions between them would have a frequency of zero in a large web corpus, ENCOW16A-NANO (Schäfer, 2015; Schäfer & Bildhauer, 2012). Due to a bug that only considered the frequency of word-boundary transitions, several syllable transitions do actually have frequencies greater than zero. We, therefore, include the log transitional frequency as a predictor in the statistical model. To foreshadow our result, we find no influence of transitional frequency on learning.

2 It is possible that the greater frequency of one of the triplets in the input stream increases its learnability, but we did not observe this in the data. And in any case, the difference made by one extra observation of a triplet (25 observations, rather than 24) is proportionally smaller than the difference made by missing one transition (which would show one transition five times, rather than six).

3 For further discussion of the Bayesian approach to statistical inference, in which reasoning is not based on testing a null hypothesis but rather on quantifying uncertainty, see, for example, Vasishth and Gelman (2021).

References

Anderson, S. R. (1992). *A-Morphous Morphology*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

Aronoff, M., Meir, I., Padden, C., & Sandler, W. (2004). Morphological universals and the sign language type. In G. Booij & J. Van Marle (Eds.), *Yearbook of Morphology 2004* (pp. 19–40). Dordrecht: Springer.

Bickel, B., Hildebrandt, K. A., & Schiering, R. (2009). The distribution of phonological word domains: A probabilistic typology. In J. Grijzenhout & B. Kabak (Eds.), *Phonological domains: Universals and deviations* (pp. 47–75). De Gruyter Mouton.

Bloomfield, L. (1933). *Language*. New York: Holt.

Boas, F. (1911). Introduction. In F. Boas (Ed.), *Handbook of American Indian Languages, Part 1*, Bureau of American Ethnology, Bulletin 40 (pp. 1–83). Washington, DC: Government Printing Office.

Bogaerts, L., Siegelman, N., & Frost, R. (2016). Splitting the variance of statistical learning performance: A parametric investigation of exposure duration and transitional probabilities. *Psychonomic Bulletin & Review*, *23*(4), 1250–1256.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Dixon, R. M. W. (2009). *Basic linguistic theory 2: Grammatical topics*. Oxford: Oxford University Press.

Dixon, R. M. W., & Aikhenvald, A. Y. (2002a). *Word: A cross-linguistic typology* (1st ed.). Cambridge University Press.

Dixon, R. M. W., & Aikhenvald, A. Y. (2002b). Word: A typological framework. In R. M. W. Dixon & A. Y. Aikhenvald (Eds.), *Word: A cross-linguistic typology* (1st ed.) (pp. 1–41). Cambridge University Press.

Duddington, J. (2012). eSpeak text to speech. espeak.sourceforge.net.

Elazar, A., Alhama, R. G., Bogaerts, L., Siegelman, N., Baus, C., & Frost, R. (2022). When the "tabula" is anything but "rasa": What determines performance in the auditory statistical learning task? *Cognitive Science*, *46*(2), e13102.

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177–199.

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351–367.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128–1153.

Gervain, J., & Guevara Erra, R. (2012). The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, *125*(2), 263–287.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2–3), 146–162.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, *31*(2), 190–222.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, *45*(1), 31–80.

Hawkins, J. A., & Cutler, A. (1988). Psycholinguistic factors in morphological asymmetry. In J. A. Hawkins (Ed.), *Explaining language universals* (pp. 280–317). Oxford: Blackwell.

Himmelmann, N. P. (2014). Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language*, *90*(4), 927–960.

Isbilen, E. S., & Christiansen, M. H. (2020). Chunk-based memory constraints on the cultural evolution of language. *Topics in Cognitive Science*, *12*(2), 713–726.

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical learning of language: A meta-analysis into 25 years of research. *Cognitive Science*, *46*(9), 1–35.

Julien, M. (2007). On the relation between morphology and syntax. In G. Ramchand & C. Reiss (Eds.), *The Oxford Handbook of Linguistic Interfaces*, Oxford Handbooks in Linguistics (pp. 353–382). Oxford: Oxford University Press.

Langacker, R. W. (1972). *Fundamentals of linguistic analysis*. New York: Harcourt Brace Jovanovich.

Mansfield, J. (2021). The word as a unit of internal predictability. Technical report.

Martin, A., & Culbertson, J. (2020). Revisiting the suffixing preference: Native-language affixation patterns influence perception of sequences. *Psychological Science*, *31*(9), 1107–1116.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750.

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reichling, A. (1935). *Het Woord: Een Studie Omtrent de Grondslag van Taal En Taalgebruik*. Nijmegen: Berkhout.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)* (pp. 28–34).

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 486–493). Istanbul, Turkey: European Language Resources Association (ELRA).

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*(1), 1–32.

Siegelman, N. (2020). Statistical learning abilities and their relation to language. *Language and Linguistics Compass*, *14*(3), 1–19.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, *177*, 198–213.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, *49*(2), 418–432.

Tallman, A. J. R. (2020). Beyond grammatical and phonological words. *Language and Linguistics Compass*, *14*(2), e12364.

Tallman, A. J. R., & Epps, P. (2020). Morphological complexity, autonomy, and areality in western Amazonia. In P. Arkadiev & F. Gardani (Eds.), *The complexities of morphology* (pp. 230–264). Oxford University Press.

van Wyk, E. B. (1968). Notes on word autonomy. *Lingua*, *21*, 543–557.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2015). Why are we so sure we know what a word is? In J. R. Taylor (Ed.), *The Oxford Handbook of The Word*, Oxford Handbooks in Linguistics (pp. 725–750). Oxford: Oxford University Press.