Highlights

Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled' from memory or compositionally constructed on the fly?

Luca Onnis, Falk Huettig

- Is language processing compositional or preassembled?
- We reassessed prior evidence for multi-word storage with corpus data
- Prediction and retrodiction appear to be important influences on multi-word processing
- Multi-word units vs compositional construction is a dual route process
- Forward and backward predictability are both informative of lexical cohesiveness

Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled' from memory or compositionally constructed on the fly?

Luca Onnis[a], Falk Huettig[b,c]

[a]*School of Social Sciences, University of Genoa, Genoa, Italy*
[b]*Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*
[c]*Centre for Language Studies, Radboud University, , Nijmegen, The Netherlands*

Abstract

An important debate on the architecture of the language faculty has been the extent to which it relies on a compositional system that constructs larger units from morphemes to words to phrases to utterances on the fly and in real time using grammatical rules; or a system that chunks large preassembled, stored units of language from memory; or some combination of both approaches. Good empirical evidence exists for both 'computed' and 'large stored' forms in language, but little is known about what shapes multi-word storage / access or compositional processing. Here we explored whether predictive and retrodictive processes are a likely determinant of multi-word storage / processing. Our results suggest that forward and backward predictability are independently informative in determining the lexical cohesiveness of multi-word phrases. In addition, our results call for a reevaluation of the role of retrodiction in contemporary language processing accounts (cf. Ferreira and Chantavarin 2018).

Keywords: frequency effects, prediction, postdiction, retrodiction, stored sequences

1. Multi-word storage and compositionality

Are frequent and larger language units (e.g. *it was really funny*) constructed online using compositional rules or can they be retrieved as 'preassembled' stored chunks from long-term memory? This question has received much attention recently because it has been thought to elucidate be-

tween competing theoretical accounts of language processing. On one side of the debate there are several influential theoretical frameworks of human information processing that claim that linguistic structure is the consequence of 'emergent' processes: 1) usage-based accounts of language processing (e.g., Goldberg 2006) according to which whole chunks are taken directly from the input to be stored in the mind, and 2) exemplar models of stored knowledge (e.g., Nosofsky 1988) that assume that we store examples in memory rather than forming abstract generalisations, i.e. linguistic structures 'emerge' from experienced patterns in the input. If frequent multi-word sequences were represented and used routinely as chunks (rather than compositionally computed online) then this would provide support for notions that argue that language processing involves the processing of dynamic patterns at different grain sizes (Elman, 2009) rather than stable lexical (word-like) units.

On the other side of the debate there are approaches that assume an essential role for the computation of compositional multi-word phrases (e.g., Pinker and Ullman 2002). Compositional approaches do not deny that some longer phrases can occasionally be stored, for example idioms (e.g. *kick the bucket*) could be stored as a whole, but the debate is unresolved about whether a very large number of frequent multi-word phrases (e.g. *it was really funny*) are computed in real time from their component words, or are instead stored and retrieved as a whole chunk.

More and more researchers (e.g., Snider and Arnon 2012, cf. Bod 2006) have started to question a strict distinction between compositionally constructed vs. stored longer phrase units. Jackendoff (e.g., Jackendoff 2002) for examples argues in this regard that the ease or speed with which a rule may be activated relative to stored phrases plays a role in how 'freely productive' it is. Further work is needed to elucidate among competing accounts of multi-word processing. The present study aims to contribute to this endeavour.

2. Multi-word frequency effects

Frequency effects seem ubiquitous in language (Pfänder and Behrens, 2016): forms and structures that are highly frequent are acquired and processed faster than infrequent ones, both in comprehension and production. Crucially, such processing advantage is often taken as a signature of the fact that the language units are accessed as 'precompiled' from memory, and not computed on the fly. To the extent that frequency effects apply to the lexicon,

they would be consistent with a division of labour whereby compositional mechanisms do the independent syntactic work of assembling morphemes and words (Pinker, 2015; Ullman, 2016, 2004). However, the detection of so-called frequency effects for larger units of language such as grammatical phrases that include lexical and syntactic items has been proposed as evidence that language is much less compositional. Consistent with such suggestions, Bannard and Matthews (2008) proposed that children store more than individual words in memory based on their results that young children were significantly more likely to repeat frequent sequences such as *a drink of milk* correctly than to repeat infrequent sequences such as *a drink of tea*. Such a view is consistent with the notion that compositional constructions only emerge gradually during child development (e.g., Tomasello 2000).

There are however similar data with adult participants. Arnon and Snider (2010) for example found that adults responded faster to higher frequency than lower frequency phrases in a phrase-recognition task. Adult speakers' recognition times for *we have to talk* for instance were faster than for *we have to sit*, with the latter having lower overall frequency as a four-word unit than the former.

In the following section we first consider some possible conceptual objections to a theoretical distinction between 'stored' and 'computed' linguistic forms. Then, in the next section we ask whether the documented phrase-frequency effects for multi-word phrases may emerge from dynamic online processes driven by context predictability rather than phrase frequency per se. The subsequent corpus analyses indeed support the view that frequency effects for multi-word sequences are effects of online prediction and retrodiction in disguise. We find evidence that (forward and backward) transitional probabilities at multiple levels (which may contribute to the overall high frequency of the entire multi-word sequence) could support sequential, compositional processing rather than chunk-based processing. In the Discussion section, we then consider which cognitive and neural mechanisms could give rise to predictability effects on multi-word sequences. This allows us to reappraise the debate on 'stored' versus 'computed' forms by proposing an alternative framework that can account for facilitative processing effects on combinatoriality. Finally, we discuss some limitations of the present approach in particular with regard to hierarchical syntactic compositional parsing approaches.

2.1. Some conceptual inadequacies concerning a strict dichotomy

First, considering conceptual inadequacies, we conjecture that a theory of language processing that relied on a very large number of memorised pre-existing chunks would face difficulties accounting for graded effects of lexical access. Indeed, Arnon and Snider (2010); Snider and Arnon (2012) showed that their documented frequency effects for four-word sequences occurred across the frequency range and was thus a gradient one.

Secondly, and consequently, if frequency effects are graded it is difficult to establish an empirical threshold for what multi-word sequences should be retrievable whole versus being compositionally computed online. Given that frequency is a continuous variable in language, and the logarithm of frequency is linearly related to reaction times in various psycholinguistic tasks, a dichotomous categorization of lexical items in stored versus non-stored / compositionally computed sequences is hard to achieve.

Third, the frequency distribution of linguistic items – including multi-word sequences – while being continuous is highly non-linear and skewed (Zipf, 1949). The vast majority of sequences (or n-grams in technical parlance) are positioned in the long tail of infrequent and rare events. This would practically leave most of the language of interest outside the benefits of mental storage, and would thus be of little theoretical relevance in explaining how the entirety of language works. A theory of weak memory storage for such a large number of sequences would have to account for what else holds language together in processing such sequences besides a weak frequency effect.

A fourth consideration is that while storage of single lexical items is large, storage of unique 2-, 3-, 4-grams, and so on, is even larger by several magnitudes, as evidenced by large scale n-gram corpus analyses, including our own below. And this state of affairs does not even consider non-adjacent n-grams such as *in X opinion*, where *X* can be replaced by a personal pronoun (*in my/your/their/ opinion*) or a noun in genitive form (*in teachers' opinion*). Most language in fact has been characterised in terms of partially matching sequences, which may have gaps or open slots (Kolodny et al., 2015).

Relatedly, as a fifth consideration most frequent linguistic patterns are composed of sequences of varying degree of compositionality and abstraction (e.g., *more than Y know**, where *Y* is an open slot that can be filled by various pronouns and nouns, and the verb stem *know** agrees morphologically with *Y* and can take different tense forms).

As a sixth and final point, phrases can be part of linguistic patterns of different sizes, just like syllables can be part of different words. For instance, *you know* is one of the most frequent interjections in oral everyday communication, but so is also the phrase *you know what?* or *what do you know?*. Which of these phrases is a stored sequence in the mind? If the first one is, then the latter larger phrases must allow a compositional process. If the latter two are stored sequences, then they must allow a decompositional process to account for the first phrase.

A similar issue is that chunk-based processing could be seen as akin to deferring recognition of a spoken word until all its phonemes have occurred. Such a mechanism would arguably slow processing. Moreover, strong cues to end-of-sequence may only occur in a few circumscribed contexts. This raises the issue of how word recognition for word sequences would be deferred. Indeed, unlike spoken words, where sublexical components arguably remain highly ambiguous at least in some languages, 'sub-sequence' units in multi-word sequences are words, each linked to distinct semantic representations and form classes. In other words, it seems implausible that "it is time to..." in "it is time to talk" would be analogous to hearing "formul..." (all but the final phoneme of 'formula'), where there are arguably not discrete elements that require actual classification (rather than a distribution of activations/probabilities over possible phonemes or syllables at each position).

Clearly, compositionality cannot be disposed of easily even in the case of frequent multi-word sequences. What could plausibly reconcile the ubiquitous frequency effects for multi-word phrase processing found in the literature while allowing for an essentially compositional system? And, could this change the debate over stored versus computed language? In the next section we propose that prediction and retrodiction processes (cf. Ferreira and Chantavarin 2018; Ferreira and Qiu 2021), here formalized as sensitivity to contextual forward and backward probabilities between words, can account for facilitative effects in language processing for multi-word expressions of the kind empirically found in the literature.

3. A role for prediction and retrodiction in multi-word processing

The evidence and theoretical arguments considered above leave open the crucial question of what determines whether frequent multi-word phrases become stored in (and accessed online from) memory or are compositionally constructed on the fly. In essence we are exploring what determines

whether phrases of various sizes are 'lexically listed'. More specifically we tested whether dynamic probabilistic online processes are informative in answering this question and investigated whether forward and backward transitional probabilities can provide important insights about lexical cohesiveness, which in turn can affect the online processing of multi-word phrases. In reanalysing existing four-word phrases from two published studies, we conducted new corpus analyses (see Method and Results sections) on both the Arnon and Snider (2010) and corresponding developmental Bannard and Matthews (2008) studies and found that the last words in the frequent phrases used in the above studies are also more predictable, both in terms of forward and backward predictability. This, we contend, suggests that predictive and 'postdictive' (or retrodictive) processes may be an important factor determining multi-word storage and processing. Our analyses cannot directly reveal whether participants retrieved multi-word phrases from memory or constructed them online compositionally but they are compatible with the notion that the processing advantage found in the two 'stored sequences studies' may be a consequence of a) pre-activation of the last words in the multi-word sequences (consistent with forward predictability), and/or b) ease of integration of the last word (consistent with backward predictability).

4. Method

4.0.1. Dataset

The dataset under scrutiny contained all 122 experimental stimuli used by Bannard and Matthews (2008) (n = 32) and (Arnon and Snider, 2010) (n = 90). While the two subsets came from separate studies, they were constructed with the same criteria and design in mind, and are thus groupable into a single dataset here. The stimuli were pairs of four-word phrases that differed in the final word. In each pair, the phrases differed in phrase-frequency (high vs.low) but were matched for substring frequency (word, bigram, and trigram): the phrases did not differ in the frequency of the final word, bigram or trigram.

For the Bannard set, the high-frequency repeated 4-word sequences (e.g., *when we go out*) were selected from a naturalistic corpus of about 1.72 million words of maternal child-directed speech. The Arnon set was selected from a 20-million corpus of American English collected from telephone conversations in the Switchboard and Fisher corpora for the Arnon study.

The other half of the dataset was made of sequences matched by the authors with low-frequency sequences on the last word (e.g., *when we go in*), to obtain 61 minimal lexical pairs. Each 4-word sequence had been labelled 'frequent' or 'infrequent', according to the authors' analyses of corpus frequency, and we used such information as Dependent Variable in our analyses. For the Bannard set, the final words of matched sequences were controlled for (a) the frequency of the final word (e.g., *juice* and *noise* were roughly equally frequent), (b) the frequency of the final bigram (e.g., *of juice* and *of noise* were roughly equally frequent), and (c) the length of the final word in syllables. The Arnon set also controlled for trigram frequency. Six additional sequences from the Bannard dataset were labelled 'intermediate frequency' and were not considered in our analysis, because of their insufficient number to form a third category on their own.

4.0.2. Corpus

To calculate new lexical statistics over the existing dataset, we used two corpora. To model child language sequences in the Bannard set, we downloaded all 1-, 3-, and 4-grams of child-directed speech from an online repository of Childes corpora available at http://www.lucid.ac.uk/resources/for-researchers/toolkit/ as part of the Language Researchers' Toolkit project (Chang, 2017). This corpus contains 40,507 1-gram types (9,222,801 tokens), 1,725,122 3-gram types (5,331,077 tokens), and 2,467,181 4-gram types (4,062,022 tokens).

To model adult sequences in the Arnon set, we obtained 1,3, and 4-grams based on the Corpus of Contemporary American English (COCA), one of the largest publicly-available, genre-balanced corpus of English. The data at the time of compilation contained approximately 430 million word tokens.

4.0.3. Measures

From the corpora we obtained three lexical statistics of cohesion for each sequence in the dataset: 1) the frequency of each sequence on logarithmic scale; 2) the forward and 3) backward Surprisal of the last word on each sequence. In psycholinguistics, a hypothesis has gained ground that processing difficulty is proportional to the amount of information conveyed. Surprisal S is an information-theoretic measure that estimates how unexpected a given event is. Conceptually, improbable, i.e. 'surprising' events carry more information than expected ones, so that surprisal is inversely related to probability, through a logarithmic function. In the context of language processing,

if w_1 denotes a multi-word sequence, then the cognitive effort required for processing the next word, w_t, is assumed to be proportional to its surprisal (Hale, 2006):

$$effort(t) \propto surprisal(w_t) = -log(P(w_t \mid w_1, ..., w_{t-1})) \tag{1}$$

where $P(w_t \mid w_1, ..., w_{t-1})$ is the forward probability of w_t given the sentence's previous words $w_1, ..., w_{t-1}$.

For example, the surprisal of one of the sequences in our dataset *when we go out* is simply the sum of the individual items' surprisal:

$$\begin{aligned} S(when\ we\ go\ out) = S(when) + S(we) + S(go) + S(out) = \\ -logP(when \mid < sos >) - logP(we \mid < sos >, when) \\ -logP(go \mid < sos >, when, we) \\ -logP(out \mid < sos >, when, we, go) \end{aligned} \tag{2}$$

where $< sos >$ denotes a start-of-sentence symbol. The summation is relevant psychologically because surprisal is linearly related to reading times, and the reading time of a sequence of words equals the sum of reading times of its parts. Hence, surprisal of a multi-word sequence must equal the sum of surprisals of its parts. In our case, because the high-frequency and low-frequency sequences differed only in the last word, it was sufficient to measure the surprisal at the last word, e.g. comparing

$$-logP(out \mid when, we, go) \tag{3}$$

and

$$-logP(in \mid when, we, go) \tag{4}$$

The measure above is forward surprisal, i.e. as a function of the probability of a word given its previous context. Backward surprisal can also be calculated, based on the backward transitional probability, namely the likelihood of a context preceding a word. It denotes the frequency of the 4-gram sequence relative to all instances of the final word in the sequence. Again using the example above, the relevant comparison of backward surprisal was:

$$-logP(when, we, go \mid out) \tag{5}$$

and

$$-logP(when, we, go \mid in) \tag{6}$$

Forward and backward probabilities were calculated using the corpus n-grams described above.

5. Results

5.0.1. Baseline model

Of the total 122 4-word sequences under scrutiny, 3 from the Arnon and 4 from the Bannard sets were excluded because 4-gram frequencies could not be calculated from the corpora. To first establish that our analyses with our corpora were comparable to original analyses, we assessed whether frequency of 4-gram sequences was a predictor of category assignment. A baseline logistic regression model included the (log)Frequency and Study (Arnon vs Bannard) to predict the category (low frequency vs high frequency sequences, as defined by Bannard and Arnon) of their experimental items (4-gram sequences). In line with the two previous studies, we also found that Frequency was a predictor for both datasets (β = 0.33, CI = -0.37, 1.03, see Table 1 and Figure 1).

5.0.2. Additive model

To assess whether the predictability of the last word of each sequence was informative in distinguishing sequence category, we ran a separate logistic regression adding Forward and Backward surprisal, in addition to (log)Frequency and Study. In this model, Backward surprisal ($\beta = -0.40$, CI = -0.61, -0.19) and Forward surprisal ($\beta = -0.52$, CI = -0.76, -0.27) but not Frequency nor Study were significant predictors in categorising the stimuli, (see Figure 1). The three predictor variables were only weakly to moderately correlated (Forward surprisal and Frequency, r = -0.34, Backward surprisal and Frequency, r = -0.42, Forward surprisal and Backward surprisal, r = 0.18, see Table 2), justifying the choice of including them as linearly independent predictors. Furthermore, a test of multicollinearity

tested negative (the squared root of the Variance Inflation Factor was less than two). Finally, when directly comparing the two regression models, the Additive model dropped the deviance by $265.15 - 230.50 = 34.64$, which was highly significant $p < .001$. Thus, based on these analyses the two categories of stimuli from the Bannard and Arnon datasets were distinguishable by predictability of the last word, more than the frequency of the stimuli. Surprisal estimates based of both forward *and* backward conditional probabilities were predictive of stimulus category, with more surprising sequence endings being categorised as 'low-frequency' items by the logistic regression. These results dovetail with the literature in reading and sentence processing that found that words in more predictable contexts are read more quickly (e.g. Hale, 2006; Frank and Bod, 2011), and suggest that corpus-derived conditional probabilities are a significant predictor of single as well as multiword processing, over and above base frequencies as a covariate.

Table 1: Summary of the logistic regression analyses for variables predicting 4-word sequence category

	Dependent variable:	
	Sequence category	
	Baseline Model	Additive Model
(log) Frequency	0.200^{***} (0.063, 0.337)	-0.023 (-0.189, 0.143)
Study	0.329 (-0.372, 1.031)	0.598 (-0.222, 1.418)
Backward surprisal		-0.402^{***} (-0.613, -0.191)
Forward suprisal		-0.516^{***} (-0.764, -0.268)
Constant	-0.569^{*} (-1.167, 0.029)	5.122^{***} (2.811, 7.434)
Observations	198	198
Log Likelihood	-132.573	-115.252
Akaike Inf. Crit.	271.146	240.505
Note:		*p<0.1; **p<0.05; ***p<0.01

6. Discussion

We conceptually and statistically re-evaluated two well cited empirical studies that manipulated four-word phrases into frequent and infrequent

Table 2: Correlation matrix for variables predicting 4-word sequence category

	(log) Frequency	Forward suprisal	Backward surprisal
(log) Frequency	1		
Forward suprisal	-0.344	1	0.176
Backward surprisal	-0.416	0.176	1

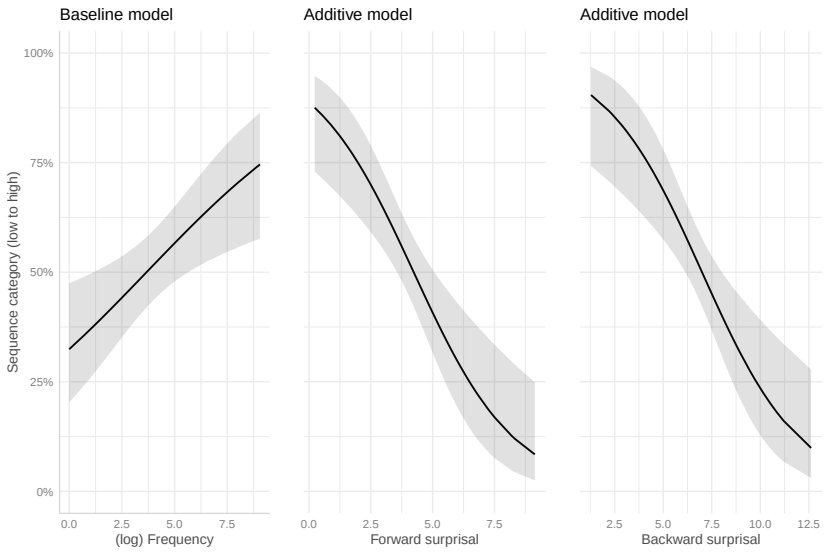

Figure 1: Marginal effects in the Baseline and Additive logistic regressions

categories, and found facilitative processing effects for the frequent phrases. Following these studies, frequency effects for multi-word expressions have been taken as evidence that a larger amount of language than previously acknowledged may be pre-compiled and stored in the mental lexicon rather than being processed on the fly by a real-time processor. In new corpus analyses, we found that the last word in the frequent phrases used in the above studies are also more predictable than in the infrequent phrases, both in terms of forward and backward predictability. This suggests an alternative interpretation of the original studies, namely that multi-word storage effects are prediction and retrodiction effects in disguise. We now discuss the implications of the present results.

6.1. Forward and backward looking

First, our results fit very much with recent accounts that highlight an important role for proactive prediction and integrative 'retrodiction' in language processing and learning (cf. Ferreira and Chantavarin 2018; Ferreira and Qiu 2021; Huettig and Guerra 2019; Huettig and Mani 2016). A large body of psycholinguistic evidence suggests that language users frequently predict upcoming words (e.g., Huettig 2015; Pickering and Gambi 2018, for review). One type of evidence consistent with such views are findings that word-to-word statistical information can constrain interpretation in the forward direction, so information from one word yields predictions about properties of upcoming words. Crucially, in the present study we found also evidence for the importance of probabilistic processing in the backward direction. Accordingly, our results point to a reevaluation of the role of what might be called 'probabilistic retrodiction' in language, which is understudied (or at least currently underappreciated, cf. Ferreira and Chantavarin 2018; Ferreira and Qiu 2021) in the psycholinguistics literature in favour of forward predictive models. In addition, our results suggest that forward and backward predictability are independently informative (and perhaps equally so, as the standardised beta values are of similar magnitude and influence the dependent variable in the same direction) in determining the storage, access, and processing of multi-word phrases. These findings also dovetail with recent evidence that probabilistic integration in the backward direction explains variance in processing modifier–noun collocation combinations like *vast majority* (McConnell and Blumenthal-Dramé, 2019), as well as reading times of naturally occurring sentences read silently (Onnis et al., 2021), and aloud – see Moers et al. (2017), although in the latter study the contributions

of forward and backward probabilities were combined in a single predictor, and could not be disentangled.

6.2. What is retrodiction?

The question of how the past, which has already been observed, can be a random variable that comprehenders model probabilistically, may raise thorny questions of interpretability to some. Many current theoretical treatments conceive of predictive processing as involving an explicit representation of likely future input that is 'compared' to the actual input to compute an error signal. Given such accounts, a model that predicts the past, may perhaps not be considered a reasonable account of probabilistic retrodiction. If we acknowledge that probability theory is just one of several valid levels of describing processing and change the level of description, then the interpretation of the present results is simple. One psychological candidate mechanism is integration, whereby the processing system does not always pre-activate, or predict, upcoming input but intergrates it faster if the preceding context is a good fit, or to put it probabilistically, is more likely to precede it. This fits with experimental evidence that suggests that language input is often fast and sub-optimal and may in a fairly large number of situations 'afford' rather limited forward looking (cf. Huettig and Mani 2016).

6.3. Multi-word processing

How then do forward and backward looking processes affect the processing of multi-word phrases? On the level of the brain, one possibility is that single words are encoded as populations of neurons that can have different levels of activation. Such activation is likely highest when the neurons respond to a perceptual event (such as reading or hearing the word percept itself), or they might encode a perceptual simulation of that event, via spreading of activation with related words. If forward and backward conditional probabilities reflect the degree of potential spreading of activation between words, it is possible to envisage how words in an expression pass recurrent activation back and forth among each other, thus reinforcing each other with different degrees of activation. Higher neuronal activations can lead to faster recognition and thus faster reading or naming times at the behavioural level. Now to understand how a phrase such as *a drink of milk* can be read, named or repeated faster than *a drink of tea*, imagine a population of interconnected neurons that functions as a distributed and dynamic (over time) representation for *a drink of ....* At time step 1 the population code can spread

activation to various words that might continue the sequence, and quicker activations are expected for words that have a higher forward probability (*milk* versus *tea, alcohol, water, soda*, etc.). At timestep 2, *milk* or *tea* are read or heard and thus their percepts send bottom-up activations that add up to the pre-activations that were spread at timestep 1. Because the forward probability of *milk* is higher than *tea*, neuronal preactivation was higher for *milk* and the word can be recognised faster than *tea*.

This can be taken as the neural instantiation of the effect of forward probability on reading the last word on the 4-word phrases contemplated in this study, and is consistent with recent accounts that explain prediction in terms of neural pattern completion (Falandays et al., 2021). But how would backward probability influence processing times? Because the backward probability of *a drink of ...* is higher given *milk* than given *tea*, the perceptual activation of *milk* can send stronger feedback signals back and forth to a *a drink of ...* which reinforce each other, ultimately producing higher neuronal activation patterns for the sequence *a drink of milk* than for the sequence *a drink of tea*. We point out here that behaviourally such a neuronal state of affairs would translate into the stored sequences effects found in the literature, but crucially without the need for the sequence to be 'unanalyzed' and stored as a single mental representation. This is because the underlying neuronal structure of the lexicon can still be instantiated as a network of more or less loosely connected population codes for word representations that spread activation to each other in a web-like fashion. The strength of activation that flows back and forth from these words determines how fast these words are processed as a sequence, and is proportional to word-to-word probabilistic properties such as forward and backward probabilities, frequencies, and numerous potential other factors not considered here, such as semantic relations, phonological similarity, and grammatical dependencies (cf. Ferreira and Qiu 2021).

We stress here however that our results should not be taken as ruling out that some multi-word phrases can be stored and retrieved as a whole. We do interpret our findings however as suggesting that there is most likely a strong limit to what kind of sequences end up stored as multi-word sequences and will be retrievable whole versus being compositionally constructed online. We believe that the present results are most compatible with some form of a dual-route process, in which compositional construction of multi-word sequences is akin to a default process but leaving open the possibility for storage and retrieval of multi-word units.

6.4. Future work and conclusion

Further research is required to explore the circumstances that increase the likelihood of storage and (preferential) access of multi-word sequences. Similarly, another important task for future research will be to investigate the exact mechanisms of how predictive and retrodictive processes determine the extent to which frequent multi-word phrases are compositionally constructed on the fly. For example, it may be possible to assess the independent contribution of forward and backward surprisal on different real-time processing tasks, such as self-paced reading, phrase repetition, phrase recognition, and phrase naming tasks, by selectively manipulating the informativeness of each cue (high or low), while maintaining constant the sequence overall frequency. It is possible to select from a large database of language such as Google Books multi-word sequences that are matched in forward surprisal but differ in backward surprisal, and vice versa. Based on our regression analyses, we predict facilitatory effects of processing (faster reading times, more accurate repetitions, and faster recognition) for both types of stimuli.

Electrophysiological studies may also turn out to be a fruitful avenue for further work. For example, when considering neural activity, the N400 ERP component has been studied extensively and taken as a measure of expectation violation, including probabilistic expectations that are measurable in terms of conditional probabilities between elements. Because the N400 is sensitive to different degrees of probabilistic violations, it is a candidate neural signature for both forward and backward probabilistic processing. Thus, one would predict that a stronger N400 ERP component is correlated with higher levels of multi-word surprisals in both the forward and backward direction, lending support for a common neural mechanism.

Another direction for future work could be to explore the effect of stored multi-word sequences on (word) cohort processing in speech processing (cf. Allopenna et al. 1998). If a multi-word sequence is processed as a chunk, reduced cohort competition should be observed for words in the sequence other than the first word (similar to reduced activation of 'bone' in trombone' or 'ate' in 'agitate' in spoken word recognition).

Finally, it is important to mention that the focus of the present study has been on whether people learn and process multi-word phrases as lexical units rather than as sequential combinations of individual words. In this type of research, the items under scrutiny are typically fragments of sentences that occur within phrases and are all syntactically cohesive, such as *when we go out, a lot of noise, I have to pay*, etc. Perhaps for this reason, such work

has mostly ignored any hierarchical syntactic analysis of multi-word units. Further work thus could also usefully 'scale up' to make more contact with contemporary hierarchical syntactic compositional parsing approaches (cf. Ferreira and Qiu 2021).

6.5. Acknowledgments

We like to thank Jim Magnuson and an anonymous reviewer for their useful comments on a previous version of this paper.

References

Allopenna, P.D., Magnuson, J.S., Tanenhaus, M.K., 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of memory and language 38, 419–439.

Arnon, I., Snider, N., 2010. More than words: Frequency effects for multi-word phrases. Journal of memory and language 62, 67–82.

Bannard, C., Matthews, D., 2008. Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. Psychological science 19, 241–248.

Bod, R., 2006. Exemplar-based syntax: How to get productivity from examples. The linguistic review 23, 291–320.

Chang, F., 2017. The lucid language researcher's toolkit [computer software].

Elman, J.L., 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. Cognitive science 33, 547–582.

Falandays, J.B., Nguyen, B., Spivey, M.J., 2021. Is prediction nothing more than multi-scale pattern completion of the future? Brain Research , 147578.

Ferreira, F., Chantavarin, S., 2018. Integration and prediction in language processing: A synthesis of old and new. Current directions in psychological science 27, 443–448.

Ferreira, F., Qiu, Z., 2021. Predicting syntacting structure. Brain Research in press.

Frank, S.L., Bod, R., 2011. Insensitivity of the human sentence-processing system to hierarchical structure. Psychological science 22, 829–834.

Goldberg, A.E., 2006. Constructions at work: The nature of generalization in language. Oxford University Press on Demand.

Hale, J., 2006. Uncertainty about the rest of the sentence. Cognitive science 30, 643–672.

Huettig, F., 2015. Four central questions about prediction in language processing. Brain research 1626, 118–135.

Huettig, F., Guerra, E., 2019. Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. Brain Research 1706, 196–208.

Huettig, F., Mani, N., 2016. Is prediction necessary to understand language? probably not. Language, Cognition and Neuroscience 31, 19–31.

Jackendoff, R., 2002. Foundations of language: Brain, meaning, grammar, evolution. Oxford University Press, USA.

Kolodny, O., Lotem, A., Edelman, S., 2015. Learning a generative probabilistic grammar of experience: A process-level model of language acquisition. Cognitive Science 39, 227–267.

McConnell, K., Blumenthal-Dramé, A., 2019. Effects of task and corpus-derived association scores on the online processing of collocations. Corpus Linguistics and Linguistic Theory .

Moers, C., Meyer, A., Janse, E., 2017. Effects of word frequency and transitional probability on word reading durations of younger and older speakers. Language and Speech 60, 289–317.

Nosofsky, R.M., 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: learning, memory, and cognition 14, 700.

Onnis, L., Lim, A., Cheung, S., Huettig, F., 2021. What does it mean to say the mind is inherently forward looking? exploring 'probabilistic retrodiction' in language processing. Manuscript under revision .

Pfänder, S., Behrens, H., 2016. Experience counts: An introduction to frequency effects in language, in: Experience counts: Frequency effects in language. De Gruyter, pp. 1–20.

Pickering, M.J., Gambi, C., 2018. Predicting while comprehending language: A theory and review. Psychological bulletin 144, 1002.

Pinker, S., 2015. Words and Rules: The Ingredients Of Language. Hachette UK.

Pinker, S., Ullman, M.T., 2002. The past and future of the past tense. Trends in cognitive sciences 6, 456–463.

Snider, N., Arnon, I., 2012. A unified lexicon and grammar? compositional and non-compositional phrases in the lexicon, in: Frequency effects in language representation. Mouton de Gruyter Berlin, pp. 127–163.

Tomasello, M., 2000. The item-based nature of children's early syntactic development. Trends in cognitive sciences 4, 156–163.

Ullman, M.T., 2004. Contributions of memory circuits to language: The declarative/procedural model. Cognition 92, 231–270.

Ullman, M.T., 2016. The declarative/procedural model: A neurobiological model of language learning, knowledge, and use, in: Neurobiology of language. Elsevier, pp. 953–968.

Zipf, G.K., 1949. Human behavior and the principle of least effort. Addison-Wesley press.