CrossMark

English Language and Linguistics 20.3: 421–437. © Cambridge University Press 2016
doi:10.1017/S1360674316000320

Formulaicity as a determinant of processing efficiency: investigating clause ordering in complex sentences

DANIEL WIECHMANN
University of Amsterdam

and

ELMA KERZ
RWTH Aachen University

(Received 1 December 2015; revised 27 June 2016)

Prior research has identified a large number of processing-related factors that affect language users' choices between two (or more) competing linguistic variants. However, one factor that up until now has not been taken into account is the reliance on formulaic language. This is surprising given a growing body of evidence demonstrating that recurrent multiword units – formulaic sequences and low-scope constructional patterns – play a key role in language learning and processing. Building on insights from research on formulaic language and usage-based constructionist models of language, we take a first step towards determining whether and to what extent formulaicity affects choices language users make in encoding the preverbal message. Focusing on clause ordering in English complex sentences with adverbial clauses (ACs), we investigate whether the choice between a preposed and a postposed AC is affected by the degree of formulaicity of the clause, as measured by (1) the amount of formulaic sequences in the AC and (2) the degree of entrenchment of an AC when treated as a low-scope pattern. Based on a reanalysis of the data in Diessel (2008), we found that the likelihood of an AC being produced in the cognitively more demanding sentence-initial position increased with higher degrees of formulaicity for both measures investigated. We conclude that reliance on formulaic language can be conceived of as a 'support strategy' compensating for increased processing effort.

1 Introduction

1.1 Processing-based accounts of clause ordering in English

Language users often have to choose between two (or more) competing variants that encode roughly the same information. The choices, which are commonly understood to involve non-conscious decision-making processes (Jaeger 2010: 25), have been variously referred to as 'constructional alternation', 'constructional choices', 'grammatical variation', 'structural alternation' and 'syntactic variation' – these terms will be used interchangeably here. The large body of research into choice phenomena in English includes investigations of the dative alternation (e.g. Bresnan *et al.* 2007), the genitive alternation (e.g. Rosenbach 2002; Hinrichs & Szmrecsanyi 2007), preposition placement (e.g. Gries 2003; Hoffmann 2011), the active/passive alternation (e.g. Weiner & Labov 1983), the locative alternation (e.g. Goldberg 2006;

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

Iwata 2008), synthetic/analytic comparatives (e.g. Mondorf 2009; Cheung & Zhang, this issue) as well as numerous types of reduction phenomena, such as relativizer omission (e.g. Jaeger & Wasow 2008), reflexive pronoun omission (Eitelmann, this issue) or complementizer omission (e.g. Race & MacDonald 2003; Jaeger 2010; Kataari, this issue; Rohdenburg, this issue), full vs contracted forms (e.g. Frank & Jaeger 2008), and optional argument and adjunct omission (e.g. Brown & Dell 1987; Resnik 1996). This research, typically associated with either discourse-pragmatic and processing-based accounts, has identified a large set of factors that influence choice phenomena (see Wasow 2002; Bresnan *et al.* 2007; Jaeger 2010; Wiechmann & Kerz 2013). In discourse-pragmatic accounts, the choice between the competing variants is seen to depend on which of the two is in accordance with the principles of information structure (see e.g. Givón 1995). In processing-based accounts, motivations for these alternations are divided into speaker-based or hearer-based ones: for example, in speaker-motivated accessibility-based accounts speakers are thought to produce cognitively more accessible elements earlier in the sentence, thereby maximizing fluent incremental production (see e.g. Branigan *et al.* 2008 for an overview; see also Levy & Jaeger 2007 and Jaeger 2010 for more recent speaker-based accounts based on the principle of Uniform Information Density, which states that speakers choose the variant that allows them to distribute information uniformly across the linguistic signal). In hearer-based accounts, a speaker chooses a certain variant either because it reduces processing costs for the hearer or it helps to get the intended message across in an unambiguous manner (see Gibson 2000; Hawkins 2004; Temperley 2007). Another idea put forward in the literature relating choice phenomena to processing effort is that of 'support strategies', where choices of linguistic variants can compensate for an increased processing effort (see Mondorf 2009; Mondorf & Pérez-Guerra, this issue). We will turn to this idea and its relevance for the purposes of the present study below.

We now deal with the processing-based accounts of the choice phenomenon investigated in the present study, namely the phenomenon of clause ordering in English complex sentences with adverbial clauses (AC). English permits adverbial clauses to either follow or precede their associated main clause, as exemplified in (1) and (2) below.

(1) You might be forced to play the half-volley, **when you are coming into the net**.
(example from the British component of *International Corpus of English*)

(2) **When you are coming into the net**, you might be forced to play the half-volley.
(variant of attested usage)

However, the two variants differ in terms of processing load: the production of a complex sentence with a preposed AC, like the one in (2), requires that the biclausal structure has been planned at the beginning of the utterance, whereas a postposed AC, like the one in (1), can in principle be planned at a later time. Producing a preposed AC may thus increase working memory load and hence increase the processing difficulty (see Hawkins 1994, 2004; Gibson 1998, 2000). The most comprehensive study of

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

clause ordering in complex sentences with ACs is Diessel (2008). Based on data from the British component of the *International Corpus of English* (ICE-GB), this study investigated the effects of four factors on clause ordering in complex sentences with temporal ACs, illustrated in examples (1) and (2) above: (i) the semantics of the AC, (ii) its syntactic complexity, (iii) its length, and (iv) conceptual order (i.e. iconicity of sequence). Diessel (2008) found that preposing was more likely when the AC was relatively short and the order of clauses corresponded to the temporal order of the events described by the two clauses.[1] Diessel (2008: 25) interprets his findings as consistent with processing-based accounts as both length and conceptual order contribute to the overall processing demand of complex sentences. The present study investigates the potential effect of a further factor in clause ordering of complex sentences with ACs, namely the effect of the degree of formulaicity of the AC, and examines whether language users use formulaicity as a support strategy to compensate for the increased processing efforts resulting from placing the AC in the cognitively more demanding environment, i.e. when the AC is placed in sentence-initial position. To this end, we build on and attempt to synthesize insights from two research traditions that have up until now proceeded largely in parallel, namely formulaic language research and research on usage-based constructionist models of language.

1.2 Formulaicity from a usage-based constructionist perspective

In its earlier phases, formulaic language research has focused on idiomatic expressions, i.e. recurring sequences that display some kind of semantic opacity or syntactic irregularity. Such patterns, e.g. *to cut a Gordian knot* or *run-of-the-mill*, are commonly treated as non-compositional units that are holistically stored in the mental lexicon. However, as aptly observed by Wray (2000: 466), these types of patterns constitute a rather small subset of formulaic language. In order to approach the phenomenon more comprehensively

> it is necessary to allow for the possibility that word sequences may be formulaic even though they do not need to be, that is, even though they are semantically transparent and syntactically regular.

Motivated by an increasing awareness of the pervasiveness of recurrent patterns in language use and their crucial role in language learning and processing (see e.g. Schmitt 2004; Ellis *et al.* 2008; Wray 2008; Corrigan *et al.* 2009a, 2009b; Arnon & Snider 2010; Shaoul & Westbury 2011), more recent formulaic language research has widened its scope so as to include a wide range of co-occurrence phenomena, which has given rise to a wide range of terms used to describe such phenomena (see

[1] Diessel (2008) also found that – for certain types of temporal clauses – clause order was affected by the semantics of the AC, i.e. clause order depended on whether the meaning of the AC was (i) purely temporal, (ii) temporal with an implicit conditional meaning, or (iii) temporal with an implicit causal or purposive meaning (see Diessel 2008: 14).

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

Wray 2000: 465), encompassing chunks, multiword units, prefabricated routines and patterns, ready-made expressions, lexical bundles, multi-word sequences, n-grams, formulaic sequences, etc.

Some definitions of these terms allow for flexibility in intervening words with open slots, some refer to arbitrarily long strings of words, some place the focus on figurative meaning and some cover aspects of semantic compositionality or non-compositionality. In the present study we adopt the term 'formulaic sequence' defined as

> a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (Wray & Perkins 2000: 1)

This definition highlights two important aspects: the first aspect concerns the representational issue of formulaic sequences, i.e. whether such sequences are holistically stored and retrieved (see Wiechmann *et al.* 2013 for an overview of computational models of linguistic knowledge). The second and more relevant aspect in the context of the present study is that this definition allows the inclusion of not only concrete lexical material but also other meaning elements, and also the inclusion of both optional and flexible slots, bringing the notion of 'formulaic sequence' closer to the notion of 'construction' in constructionist approaches to language. Like research into formulaicity, constructionist theory, too, focused on semantically and syntactically opaque patterns in the early phases of its development (see Goldberg 1995; Kay & Fillmore 1999). One of the earliest definitions of a construction is Goldberg (1995: 4):

> C is a CONSTRUCTION iff$_{def}$ C is a form-meaning pair $<F_i, S_i>$ such that some aspect of F_i or some aspect of S_i is not strictly predictable from C's component parts or from other previously established constructions (Goldberg 1995: 4)

Constructions cover units that range from morphemes and words via fully lexically filled patterns, such as idioms (e.g. *kick the bucket*), to partially lexically filled constructions (e.g. *pull* [PERSON]*'s leg*), referred to here as 'low-scope patterns', to fully schematic constructional patterns (e.g. the English ditransitive construction, i.e. NP V NP NP). The traditional compartmentalization between syntax and lexicon has thus been abandoned by constructionist approaches, and has been replaced with a continuum between a schematic (abstract) and a substantial (lexically specific) pole, the so-called syntax–lexicon continuum. In later work, the definition of constructions was modified so as to also encompass fully semantically predictable patterns, provided that they occur frequently enough (see e.g. Goldberg 2006). This connects to the idea in Wray & Perkins' (2000: 1) definition given above and the considerations of holistic storage and retrieval, which in constructionist accounts is a consequence of processes of entrenchment or automatization (see Langacker 1987; see also the phenomenon of chunking in Bybee 2010). Both the degree of formulaicity of a sequence and the degree

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

of entrenchment of a construction are related to frequency of occurrence, and a wide range of operationalizations of entrenchment – ranging from raw frequency counts to likelihood measures to asymptotic and exact hypothesis tests – has been suggested in the literature (see e.g. Evert 2004; Wiechmann 2008).

Turning to the role of formulaic sequence in language learning, usage-based models have highlighted the item-based and frequency-based nature of the development of grammatical knowledge: children are shown to be rather conservative learners, whose early productions mimic the particular sequences of lexical items they have frequently encountered in their input. Over time, the positions, or slots, in these lexical sequences are slowly and gradually generalized, leading to the emergence of more schematic, grammatical knowledge (see MacWhinney 1998; Tomasello 2003; Diessel 2004; Dabrowska & Lieven 2005; Abbot-Smith & Tomasello 2006; Bannard & Matthews 2008; Dabrowska *et al.* 2009; Matthews & Bannard 2010; Kidd *et al.* 2010, *inter alia*). This type of learning trajectory, which proceeds from formulas to low-scope patterns to fully schematized abstract constructional language knowledge, is referred to as item-based or exemplar-based learning. Two of the key studies on the role of formulaic sequences in child language processing are Bannard & Matthews (2008) and Matthews & Bannard (2010). Adopting a usage-based language acquisition stance, Bannard & Matthews (2008) identified frequently occurring formulaic sequences in child-directed speech, e.g. *sit in your chair*, and matched them to infrequent sequences, e.g. *sit in your truck*. They tested children's ability to produce these sentences in a phrase-repetition task and found that two- and three-year-olds were significantly more likely to repeat frequent sequences correctly than they were to repeat infrequent sequences correctly. They also reported that the three-year-olds were significantly faster at repeating the first three words of an item if they formed part of a formulaic sequence (e.g. they were quicker at saying *sit in your* when the following word was *chair* than when it was *truck*). Matthews & Bannard (2010) looked at the child-directed input again to investigate when children would generalize over lexical patterns to a low-scope pattern, e.g. *sit in your X*, where 'X' is a generalization over the words experienced in that position. They used a similar experimental paradigm to look at the effects of the statistical properties of the final word's slot in the four-word sequences, namely its statistical variability (measured in terms of slot entropy) and semantic density. With regard to the former, they tested if sequences that appear with a large number of different words in the child's input are processed differently from sequences that appear with only a small number of different words in the final position of the pattern. With regard to the latter, they hypothesized that the processing of sequences that have semantically similar completions, like *back in the* __, differed from that of sequences that have semantically diverse completions, e.g. *a piece of* __. They found that children were more fluent when there was high variability in the types of words to appear in the final word's slot and also when there was high semantic coherence among the words that filled that slot in the child's input.

In usage-based models, language learning and language processing are typically considered to be two sides of the same coin, since they both rely on mechanisms

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

of pattern finding (see also statistical learning; e.g. Christiansen & Chater 2008). This is a continuous process which does not end at some discrete point of time in ontogenetic development, but instead takes place across the lifespan (the learning-as-processing assumption; e.g. Chang *et al.* 2006). From that perspective it is not surprising that adult language users, too, are sensitive to distributional information at many grain-sizes (MacDonald *et al.* 1994; Trueswell *et al.* 1994; Mitchell *et al.* 1995). There is a bulk of research demonstrating sensitivity to formulaic sequences in adult language processing (Ellis *et al.* 2008; Arnon & Snider 2010; Tremblay & Tucker 2011; see also Shaoul & Westbury 2011 for an overview). For example, using a variation of the experimental design of Bannard & Matthews (2008), Arnon & Snider (2010) found that adults – like children – process frequent 4-word phrases faster than infrequent ones, suggesting that the sensitivity for statistical regularities is retained in adulthood. Reali & Christiansen (2007) investigated the effect of frequency on the processing difficulty of different types of relative clause constructions. Based on an analysis of the *American National Corpus*, they first showed that pronominal object relative clauses are significantly more frequent than pronominal subject relative clauses when the embedded pronoun is personal (*I*, *we*, *me*, *us*, *she*, *he*, *they*, *her*, *him*) rather than impersonal (*it*). Using four self-paced reading experiments, they showed that differences in pronominal object/subject relative clause processing mirrored the pattern of distribution revealed by the corpus analysis, demonstrating adults' sensitivity to frequencies of low-scope patterns. Dabrowska *et al.* (2009) found that spontaneously produced complex constructional types, such as questions involving long-distance dependencies, are of extremely stereotypical nature, i.e. they are highly formulaic: the matrix clause subject is commonly *you*, the verb *say* or *think*, the auxiliary is nearly always *do* and the matrix clause rarely contains any additional material. Consistent with the predictions of the usage-based approach, they concluded that language users' knowledge of such constructions is hence best captured in terms of relatively specific low-scope patterns rather than in terms of highly productive general rules: the lion's share of questions with long-distance dependencies are instances of two lexically specific templates, WH *do you think* S-GAP? and WH *did you say* S-GAP?, where 'WH' is an interrogative *wh*-word and 'S-GAP' is a subordinate clause with a missing constituent. The majority of the remaining question types were minimal variations on these templates that contained a different matrix subject or a different verb or a different auxiliary or an additional element like an adverb or complementizer; only 4 per cent of the instances deviated from the prototype in more than one respect, suggesting that low-scope patterns posited in usage-based models of grammar play an important role in language processing (see Langacker 2000; Verhagen 2005; Dabrowska 2008).

While the prominent role of formulaic sequences in language learning as well as their facilitatory effects on language processing have been amply demonstrated, to our knowledge there is no study that demonstrates the effects of formulaicity on constructional choices in language production.

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

1.3 Aims and scope of the present study

As described above, prior research has identified a number of processing-related factors affecting the language user's choice of otherwise equivalent variants in language production. This has been demonstrated for various types of choice phenomena in English, including the choice phenomenon investigated here, i.e. clause ordering in complex sentences.

Bringing together insights from formulaic language research and usage-based constructionist models, which have highlighted the prominent role of formulaic sequences and low-scope patterns in language learning and processing, the aim of the present study is to investigate the role of formulaicity in constructional choice phenomena.

Focusing on complex sentences with temporal ACs, we seek to determine whether and to what extent clause ordering is affected by the degree of formulaicity of the AC as measured by (i) the amount of formulaic sequences in the AC, and (ii) the degree of entrenchment of an AC when treated as a low-scope pattern. If language users use formulaicity to mitigate processing demands, then we would expect to observe higher degrees of formulaicity of ACs in the cognitively more demanding sentence-initial position. Finding such an effect would thus constitute evidence that language users use formulaicity as a support strategy to compensate for the increased processing efforts.

2 Method

2.1 Data

The present study is based on a reanalysis of the data used in Diessel (2008) on clause ordering in complex sentences with temporal ACs. The data set is a random sample of 600 occurrences of such constructions from the *International Corpus of English* (ICE-GB), which contained from each modality (spoken/written) 100 clauses headed by the adverbial subordinator *when*, 50 by *after*, 50 by *before*, 50 by *once*, and 50 by *until*. The sample was restricted to finite ACs – disregarding participial and gerund constructions. ACs that (i) occurred within the main clause, i.e. in medial positions, or that (ii), following Hengeveld (1989), are related to the main clause at the speech act level were removed. The ACs in the data set were already described with respect to four variables: (i) CONCEPTUAL ORDER (iconicity), (ii) SYNTACTIC COMPLEXITY, (iii) MEANING and (iv) LENGTH (see section 1.1). CONCEPTUAL ORDER was expressed as a three-level variable, distinguishing ACs denoting a prior event from ACs denoting a posterior or simultaneously occurring event. SYNTACTIC COMPLEXITY was expressed as a two-level factor: an AC was considered simple if it consisted of a single clause only and was considered complex when it contained at least one additional subordinate clause. The five-level factor MEANING distinguished (a) purely temporal meanings from temporal meanings with an implicit (b) conditional, (c) concessive, (d) causal or (e)

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

purposive meaning (see Diessel 2008: 477 for details). The fourth variable, LENGTH, was coded as a continuous variable that expressed the proportion of words of the complex sentence that belonged to the AC. For the purpose of investigating the effects of formulaicity, two additional variables were introduced: (v) N-GRAM FREQUENCY and (vi) PATTERN FREQUENCY. The variable N-GRAM FREQUENCY captures the total number of formulaic sequences appearing in the AC. To operationalize the notion of a 'formulaic sequence', we used a corpus-analytic measure based on frequency counts of 2-grams, 3-grams and 4-grams that occurred at least 1,000 times per 100 million words in the *British National Corpus* of English (Burnard 1995). The total number of formulaic sequences in an AC were then centred on the mean, i.e. we subtracted the mean of all data points from each individual data point. The variable PATTERN FREQUENCY expresses the degree of entrenchment of an AC when treated as a low-scope pattern. The first step in the assessment of the degree of entrenchment involved describing the AC with regard to three factors: (a) the use of subordinating conjunction of an AC (*when*, *before*, *after*, *once*, *until*), (b) the morphosyntactic realization of the AC subject (lexical vs pronominal subject), and (c) the transitivity of the AC (intransitive vs transitive clause). The subordinating conjunction signals the semantic relationship between the events described by the main clause and the adverbial clause and has been shown in prior research to be associated with preferred relative positions (see Diessel 2005). The type of morphosyntactic realization of the subject affects the information-structural properties of a clause level construction: pronominal subjects mark their discourse referent as 'given' (see Prince 1981) or 'accessible' (Ariel 1990), meaning that the corresponding clause can serve purposes of bridging, i.e. it can serve the function of smoothening information-structural chains (see Verstraete 2004; Wiechmann & Kerz 2013 and references therein). The transitivity of an AC is viewed here as providing a coarse-grained classification of event types, as the number of core syntactic arguments in a clause indicates the conceptualization of the event described by the clause: transitive clauses typically express an asymmetrical relation between two participants, which is initiated and directed by the first participant, while intransitive clauses typically denote single-participant events, in which some state or action is predicated (see Hopper & Thompson 1980; Croft 1990; DeLancey 1987; Langacker 1990). Cross-tabulation of these three factors gave rise to twenty different possible low-scope patterns. To illustrate this, the example in (3) contains an adverbial clause that instantiates the low scope pattern {SUBORDINATOR == *when*, SUBJECT == pronominal, TRANSITIVITY == transitive}, whereas the example in (4) instantiates the low-scope pattern {SUBORDINATOR == *before*, SUBJECT == lexical, TRANSITIVITY == intransitive}.

(3) *When you get a tax rebate* you get the money back after about a year.
(4) He drives right through the cross hairs just *before the missile is released*.
(examples from data set used in the present study)

The degree of entrenchment of an AC when treated as a low-scope pattern was measured by assessing the frequency of the corresponding low-scope pattern in our

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

Table 1. *Variables investigated in this study*

	Feature labels and scale		Values distinguished
Outcome variable	POSITION	2-level factor	position of the adverbial clause relative to the main clause (preposed/ postposed)
Control variable	MODALITY	2-level factor	spoken versus written language
Predictors			
Predictors from Diessel (2008)	MEANING	5-lvl factor	purely temporal, temporal with implicit conditional, concessive, causal or purposive meaning
	SYNTACTIC COMPLEXITY	2-level factor	simple versus complex (adverbial clause contains additional subordinate clause)
	ICONICITY	3-level factor	adverbial clause denotes a prior, posterior or simultaneously occurring event
	LENGTH	continuous	proportion of lexical items of complex sentence that belong to adverbial clause
Predictors related to formulaicity	N-GRAM FREQUENCY	continuous	centred total amount of 2-grams, 3-grams and 4-grams in adverbial clause
	PATTERN FREQUENCY	continuous	centred frequency of pattern defined by (I) choice of subordinator, (II) morphosyntactic realization of adverbial clause subject, (III) transitivity of adverbial clause

data, which also was centred on the mean. Table 1 presents an overview of the full set of variables investigated in this study.

2.2 Statistical modelling

To test for the 'support strategy' status of formulaicity, in the sense explicated in this issue (see Mondorf & Pérez-Guerra, this issue) and to determine whether the likelihood of the production of a cognitively more demanding variant – i.e. the

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

production of an AC in sentence-initial position – is affected by formulaicity, we fitted statistical models to the data in which the positioning of an AC was regressed on n-gram frequency and pattern frequency, while controlling the effects of other co-determinants of clause ordering reported in prior research, as well as the effects of modality, i.e. potential differences between spoken vs written language (see table 1). Since the relationship of formulaicity and clause positioning may be non-linear, the data were analysed using Generalized Additive Models (henceforth 'GAMs'; see Hastie & Tibshirani 1990; Wood 2011). GAMs extend the standard multiple linear regression model by allowing non-linear functions of each of the variables. The basic form of a GAM is given in (5) and its logit form, which was needed for the dichotomous outcome variable, clause position, is given in (6). The usefulness of GAMs lies in their ability to simultaneously fit a number of nonlinear functions (f_1-f_p) of a number of variables ($x_1 - x_p$), while retaining the additivity of linear models.[2]

$$(5)\ y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \epsilon_i$$

$$(6)\ \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \epsilon_i$$

The model was fitted using a smoothing spline with a maximum of five degrees of freedom, which determine the degree of flexibility of the fitted functions. Analysis of variance (ANOVA) was used to decide whether the function of a given variable should be linear or non-linear and to determine adequate degrees of freedom for the smoothing spline used in the GAM (see Hastie & Tibshirani 1990). All computations were performed using the R statistical software system (R Core Team 2015), as well as functions offered by the libraries GAM 1.2 (Hastie 2015), and Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation (mgcv) (Wood 2011).

3 Results

We found that all variables except SYNTACTIC COMPLEXITY were significant predictors of clause ordering in complex sentences with temporal ACs. Model comparisons based on likelihood ratio tests showed that the inclusion of both formulaicity indicators resulted in a significantly better model fit than any model without these variables, and that the likelihood of preposing increases with growing levels of formulaicity for both measures, n-gram frequency or pattern frequency ($\text{MODEL}_{\text{Diessel (2008)}}$ vs $\text{MODEL}_{\text{Diessel (2008)} + \text{pattern frequency}}$: $\chi(1) = 7.51$, $p < 0.01$, $\text{MODEL}_{\text{Diessel (2008)} + \text{pattern frequency}}$ vs $\text{MODEL}_{\text{Diessel (2008)} + \text{pattern frequency} + \text{ngram frequency}}$: $\chi(1.55) = 6.13$, $p < 0.05$). Table 2 shows the results obtained from the GAM with the variables investigated in Diessel (2008) and the two indicators of formulaicity.

[2] For categorical variables, the model estimates piecewise constant functions (one constant for each of the levels of the factors).

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

Table 2. *Results from the GAM that included all variables from Diessel (2008) and both indicators of formulaicity (negative estimates indicate a preference for sentence final positioning of the AC)*

	Estimate	SE	z	P(>z)	
Parametric coefficients					
Intercept	− 1.63	0.89	− 1.83	0.07	.
syn complex = yes	− 0.26	0.35	− 0.74	0.46	
modality = written	− 0.66	0.22	− 2.98	0.00	**
meaning = concessive	− 37.79	274.00	0.00	1.00	
meaning = conditional	2.01	0.81	2.49	0.01	*
meaning = purpose	0.22	1.31	0.17	0.87	
meaning = temporal	1.28	0.79	1.62	0.11	
inconicity = prior	1.86	0.31	5.92	0.00	***
inconicity = simultaneous	0.01	0.46	0.02	0.99	
length	− 2.64	0.74	− 3.58	0.00	***
Approximate significance of smooth terms					
	edf	Ref.df	Chi	P(>Chi)	
s(Pattern Frequency)	1	1	7.51	<0.01	**
s(Ngram Frequency)	1.59	1.97	6.13	<0.05	*
R-sq (adjusted)	0.21				
Deviance explained	19.90%				
UBRE	0.024				
N	526				

Figure 1 shows the component smooth functions illustrating the effects of formulaicity. We found that the likelihood of preposing increases with increasing degrees of formulaicity of an AC for both indicators investigated here, i.e. n-gram frequency and pattern frequency. For both measures, the difference between spoken and written language was not statistically significant.

4 Discussion

The aim of the present study was to add to the existing research on processing-based accounts of constructional choice phenomena in English by investigating the role of formulaicity, which – even though it has been shown to play a prominent role in both language learning and processing – has up until now not been taken into account. Focusing on clause ordering in English complex sentences with temporal ACs, we investigated whether and to what extent clause ordering is affected by the degree of formulaicity of the AC. Drawing on the idea of support strategies (Mondorf 2009; Mondorf & Pérez-Guerra, this issue), we expected that – if formulaicity is used to decrease the processing cost of a construction in cognitively more demanding environments – higher degrees of formulaicity would be observed in ACs in

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

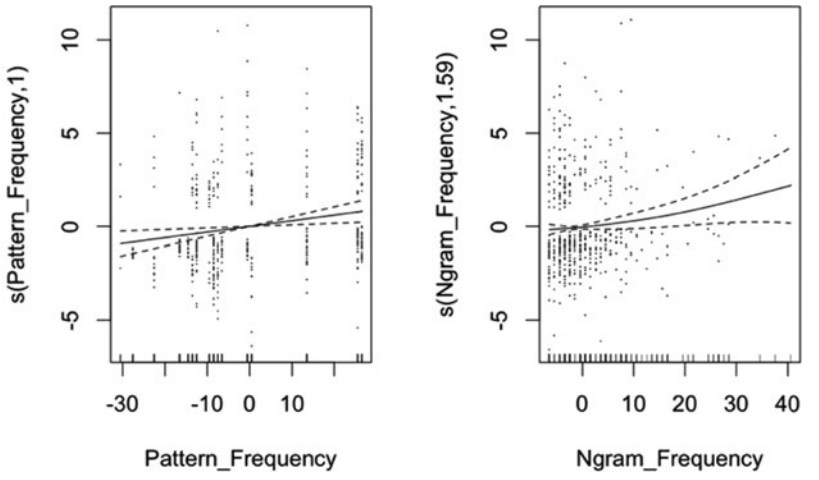


Figure 1. Effects of formulaicity on likelihood of preposing the AC, as expressed by pattern frequency and n-gram frequency

sentence-initial position, as this requires that a language user has planned the biclausal structure at the beginning of the sentence. To determine the degree of formulaicity of an AC, two measures were used: (i) the amount of formulaic sequences in the AC and (ii) the degree of entrenchment of an AC when treated as a low-scope pattern. With respect to the first measure, we expected to observe a larger amount of high-frequency 2-grams, 3-grams and 4-grams in ACs produced in sentence-initial position, compared to the amount of such sequences in ACs produced in sentence-final position. With respect to the second measure, we expected to observe higher degrees of entrenchment of the low-scope patterns capturing the use of sentence-initial ACs, compared to degrees of entrenchment of those capturing their use in sentence-final position.

In line with our expectations, we found that the inclusion of both indicators of formulaicity led to a significant increase in model fit over a model that included only the variables used in Diessel (2008). These findings are consistent with a growing body of evidence demonstrating that formulaic sequences and low-scope patterns play a key role in both child and adult language learning and processing (e.g. Reali & Christiansen 2007; Ellis *et al.* 2008; Matthews & Bannard 2008; Dabrowska *et al.* 2009; Bannard & Matthews 2010; Arnon & Snider 2010; Trembley & Tucker 2011; for reviews, see Bannard & Lieven 2012; Conklin & Schmitt 2012; Ellis 2012). This evidence comes from a great number of psycholinguistic studies including studies using acceptability judgments, reading speed studies, repetition and recall studies, and eye-movement studies (see Shaoul & Westbury 2011 for an overview). In all these studies formulaic sequences and frequent low-scope patterns were easier to process, suggesting that children and adults are sensitive to the statistics of formulaic sequences.

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

The findings reported in the present study contribute to the existing literature on processing-based accounts of choice phenomena in English by showing that aspects of formulaicity figure in language users' choices between two competing variants and that the reliance of formulaic sequences and low-scope patterns qualifies as a support strategy (Mondorf & Pérez-Guerra, this issue). The present study has provided first evidence for the phenomenon of clause ordering in complex sentences, showing that the cognitively more demanding variants, ACs in sentence-initial position, tend to be more formulaic than their sentence-final counterparts. Future work is needed to assess to what extent formulaicity as a determinant of processing efficiency is at work in other choice phenomena (see section 1.1 for an overview). This work will also need to address the informativeness of measures of formulaicity proposed in the literature to operationalize the construct of a formulaic sequence as well as that of the degree of entrenchment of a low-scope pattern (see e.g. Evert 2004 for an in-depth discussion of a wide range of mathematical formulae that interpret co-occurrence frequency data; see e.g. Hale 2001, Levy & Jaeger 2007, Frank 2013 for discussions of relevant information theoretic measures, viz. surprisal and entropy; see e.g. Wiechmann 2014 for a proposal to employ association rule learning techniques to express degrees of entrenchment of low-scope patterns and more abstract constructions).

5 Conclusion

The current study took a first step towards determining the effects of formulaicity on language users' choices between competing constructional variants in language production. Based on a reanalysis of data used in Diessel (2008), the study investigated whether clause ordering in English complex sentences with temporal ACs is influenced by degrees of formulaicity of the AC. Bringing together insights from formulaic language research and research on usage-based constructionist models of language, we proposed two measures of formulaicity capturing effects at different levels, i.e. at the level of formulaic sequences and at the level of low-scope patterns. We found that the likelihood of the production of an AC in the cognitively more demanding sentence-initial position increased with increasing degrees of formulaicity for both measures investigated here. We conclude that reliance on formulaic language at both these levels qualifies as a 'support strategy' compensating for increased processing effort.

Authors' addresses:

Institute for Logic, Language and Computation
University of Amsterdam
Spuistraat 210
1012 VT Amsterdam
Netherlands
d.wiechmann@uva.nl

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

Department of English Linguistics
RWTH Aachen University
Kármánstr. 17/19
52062 Aachen
Germany
kerz@anglistik.rwth-aachen.de

References

Abbot-Smith, Kirsten & Michael Tomasello. 2006. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *Linguistic Review* 23(3), 275–90.
Ariel, Mira. 1990. *Accessing noun phrase antecedents*. London: Routledge.
Arnon, Inbal & Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62(1), 67–82.
Bannard, Colin & Elena Lieven. 2012. Formulaic language in language acquisition. *Annual Review of Applied Linguistics* 32, 3–16.
Bannard, Colin & Danielle Matthews. 2008. Stored word sequences in language learning. *Psychological Science* 19, 241–8.
Branigan, Holly P., Martin J. Pickering & Mikihiro Tanaka. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua* 118, 172–89.
Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Kraemer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.
Brown, Paula & Garry Dell. 1987. Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology* 19(4), 441–72.
Burnard, Lou. 1995. *Users' reference guide for the British National Corpus*, version 1.0. Oxford: Oxford University Computing Services.
Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.
Chang, Franklin, Gary Dell & Kathryn Bock. 2006. Becoming syntactic. *Psychological Review* 113(2), 234–72.
Christiansen, Morten & Nick Chater. 2008. Language as shaped by the brain. *Behavioral and Brain Sciences* 31(5), 489–509.
Conklin, Katy & Norbert Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics* 32, 45–61.
Corrigan, Roberta, Edith Moravcsik, Hamid Quali & Kathleen Wheatley. 2009a. *Formulaic language*, vol. 1: *Distribution and historical change*. Amsterdam: John Benjamins.
Corrigan, Roberta, Edith Moravcsik, Hamid Quali & Kathleen Wheatley. 2009b. *Formulaic language*, vol. 2: *Acquisition, loss, psychological reality, and functional explanations*. Amsterdam: John Benjamins.
Croft, William. 1990. Possible verbs and the structure of events. In Savas L. Tsohatzidis (ed.), *Meanings and prototypes: Studies in linguistic categorization*, 48–73. London: Routledge.
Dabrowska, Ewa. 2008. Rules or schemas? Evidence from Polish. *Language and Cognitive Processes* 19, 225–71.
Dabrowska, Ewa & Elena Lieven. 2005. Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics* 16(3), 437–74
Dabrowska, Ewa, Caroline Rowland & Anna Theakston. 2009. The acquisition of questions with long-distance dependencies. *Cognitive Linguistics* 20(3), 571–97.
DeLancey, Scott. 1987. Transitivity in grammar and cognition. In Russel Tomlin (ed.), *Coherence and grounding in discourse*, 53–68. Amsterdam and Philadelphia: John Benjamins.

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

Diessel, Holger. 2004. *The acquisition of complex sentences*. Cambridge: Cambridge University Press.
Diessel, Holger. 2005. Competing motivations for the ordering of main and adverbial clauses. *Linguistics* 43, 449–70.
Diessel, Holger. 2008. Iconicity of sequence: A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive Linguistics* 19, 457–82.
Ellis, Nick. 2012. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Language* 32, 17–44.
Ellis, Nick, Rita Simpson-Vlach & Carson Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL. *TESOL Quarterly* 42(3), 375–96.
Evert, Stefan. 2004. The statistics of word co-occurrences: Word pairs and collocations. PhD thesis, University of Stuttgart.
Frank, Austin & T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *The 30th annual meeting of the Cognitive Science Society*, 939–44. Washington, DC.
Frank, Stefan L. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science* 5, 475–94.
Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68, 1–76.
Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita & Wayne O'Neil (eds.), *Image, language, brain*, 95–126. Cambridge, MA: MIT Press.
Givón, Talmy. 1995. *Functionalism and grammar*. Amsterdam: John Benjamins.
Goldberg, Adele. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization of language*. Oxford: Oxford University Press.
Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. London and New York: Continuum.
Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.
Hastie, Trevor. 2015. *gam: Generalized Additive Models*. R package version 1.12. https://CRAN.R-project.org/package=gam
Hastie, Trevor & Robert Tibshirani. 1990. *Generalized additive models*. London: Chapman & Hall CRC.
Hawkins, John. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
Hawkins, John. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
Hengeveld, Kees. 1989. Layers and operators in functional grammar. *Journal of Linguistics* 25, 127–57.
Hinrichs, Lars & Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11(3), 437–74.
Hoffmann, Thomas. 2011. *Preposition placement in English: A usage-based approach*. Cambridge: Cambridge University Press.
Hopper, Paul J. & Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56, 252–99.

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

Iwata, Seizi. 2008. *Locative alternation: A lexical-constructional approach*. Amsterdam: John Benjamins.
Jaeger, T. Florian 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61, 23–62.
Jaeger, T. Florian & Tom Wasow. 2008. Processing as a source of accessibility effects on variation. *Proceedings of the 31st Annual Meeting of the Berkeley Linguistic Society*, 169–80. Ann Arbor, MI: Sheridan.
Kay, Paul & Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language* 75, 1–33.
Kidd, Evan, Elena V.M. Lieven & Michael Tomasello. 2010. Lexical frequency and exemplar-based learning effects in language acquisition. *Language Sciences* 32, 132–42.
Langacker, Roland W. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*. Stanford, CA: Stanford University Press.
Langacker, Ronald W. 1990. Settings, participants, and grammatical relations. In Savas L. Tsohatzidis (ed.), *Meanings and prototypes: Studies in linguistic categorization*, 213–38. London: Routledge.
Langacker, Ronald W. 2000. A dynamic usage-based model. In Michael Barlow & Susanne Kemmer (eds.), *Usage-based models of language*, 1–63. Stanford, CA: CSLI Publications.
Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schlökopf, John Platt & Thomas Hoffman (eds.), *Advances in neural information processing systems*, 849–56. Cambridge, MA: MIT Press.
MacDonald, Maryellen C., Neal J. Pearlmutter & Mark S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101, 676–703.
MacWhinney, Brian. 1998. Models of the emergence of language. *Annual Review of Psychology* 101, 676–703.
Matthews, Danielle & Colin Bannard. 2010. Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive Science* 34, 465–88.
Mitchell, Don C., Fernando Cuetos, Martin Corley & Mark Brysbaert. 1995. Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research* 24, 469–88.
Mondorf, Britta. 2009. *More support for* more-*support: The role of processing constraints on the choice between synthetic and analytic comparative forms*. Amsterdam: John Benjamins.
Prince, Ellen. 1981. Toward a taxonomy of given/new information. In Peter Cole (ed.), *Radical pragmatics*, 223–55. New York: Academic Press.
R Core Team. 2015. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. www.R-project.org.
Race, David. S. & Maryellen C. MacDonald. 2003. The use of 'that' in the production and comprehension of object relative clauses. In *Proceedings of the 26th annual meeting of the Cognitive Science Society*, 946–51. Chicago: Cognitive Science Society.
Reali, Florencia & Morten Christiansen. 2007. Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language* 53, 1–23.
Resnik, Philip. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61, 127–59.
Rosenbach, Anette. 2002. *Genitive variation in English: Conceptual factors in synchronic and diachronic studies*. Berlin: Walter de Gruyter.
Shaoul, Cyrus & Chris Westbury. 2011. Formulaic sequences: Do they exist and do they matter? *The Mental Lexicon* 6(1), 171–96.
Schmitt, Norbert. 2004. *Formulaic sequences*. Amsterdam: John Benjamins.

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press

Temperley, David. 2007. Minimization of dependency length in written English. *Cognition* 105(2), 300–33.
Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Harvard: Harvard University Press.
Tremblay, Antoine & Benjamin V. Tucker. 2011. The effect of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, 6, 302–24.
Trueswell, John C., Michael K. Tanenhaus & Susan M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language* 33, 285–318.
Verhagen, Arie. 2005. *Constructions of intersubjectivity: Discourse, syntax and cognition*. Oxford: Oxford University Press.
Verstraete, Jean-Christophe. 2004. Initial and final position of adverbial clauses in English: The constructional basis of the discursive and syntactic differences. *Linguistics* 42, 819–53.
Wasow, Tom. 2002. *Postverbal behavior*. Stanford, CA: CSLI Publications.
Weiner, Judith E. & William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19, 29–58.
Wiechmann, Daniel. 2008. On the computation of collostruction strength. *Corpus Linguistics and Linguistic Theory* 4(2), 253–90.
Wiechmann, Daniel. 2014. *Understanding relative clauses: A usage-based view on the processing of complex constructions*. Berlin: Mouton de Gruyter.
Wiechmann, Daniel & Elma Kerz. 2013. The positioning of concessive adverbial clauses in English: Assessing the importance of discourse-pragmatic and processing-based constraints. *English Language & Linguistics* 17(1), 1–23.
Wiechmann, Daniel, Elma Kerz, Neal Snider & T. Florian Jaeger (eds.). 2013. Parsimony and redundancy in models of language. *Language and Speech* 56(3), 257–64.
Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society* (B) 73(1), 3–36.
Wray, Alison. 2000. Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics* 21(4), 463–89.
Wray, Alison. 2008. *Formulaic language: pushing the boundaries*. Oxford: Oxford University Press.
Wray, Alison & Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20, 1–28.

https://doi.org/10.1017/S1360674316000320 Published online by Cambridge University Press