

The Use of Psycholinguistic Formulaic Language in the Speech of Higher Level Japanese Speakers of English

Stephen F. Cutler

Cardiff University

doi: <http://dx.doi.org/10.7820/vli.v06.1.Cutler>

Abstract

A recent study by Cordier (2013) suggests that psycholinguistic formulaic sequences (multiword units that present a processing advantage to the individual speaker) may be more prevalent in L2 speakers than previously thought. The current study adopts the same identification process to explore the use of psycholinguistic formulaic sequences in the speech of Japanese Speakers of English (JSE).

Eight adult JSE at intermediate or advanced levels of English each performed two speaking tasks: a structured interview and a narration task. Formulaic sequences were identified on the basis of hierarchical conditions applied in strict order. The first condition was fluency and the second condition checked for holisticity (using given diagnostic criteria). For each sample, two measures of formulaicity were calculated: FS% (the percentage of syllables that were part of a formulaic sequence) and ANR (the average number of formulaic syllables per run).

The mean formulaicity of the samples (FS%=34.6%, ANR=1.64) suggests that psycholinguistic formulaic sequences, as defined and identified here, may be a significant feature in the speech of intermediate/advanced JSE. The study also confirms the sensitivity of the results to task, with significantly more formulaic sequences used in the interview task than in the narration. Overall, the identification process was found to be a useful and systematic way of identifying formulaic sequences, but some further refinements of the diagnostic criteria and measures used are also suggested.

1. Background

1.1 Psycholinguistic Formulaic Sequences and L2 Speakers

Formulaic sequences may be defined as prefabricated multiword strings that behave as a single lexical unit. Such sequences are thought to be a significant and ubiquitous feature of native speaker language (Ellis, 2012; Nattinger & DeCarico, 1992; Pawley & Syder, 1983) and to play a key role in facilitating fluency and automaticity in speech (e.g., Eijzenberg, 2000; Towell, Hawkins, & Bazergui, 1996; Wray, 2002). The central claim is that the sequence as a single holistic lexical item is processed more quickly during speech production than if processing the same sequence online on a word-by-word basis. For this reason, the acquisition and use of an appropriate stock of formulaic sequences would seem to be a useful goal for

L2 learners wishing to develop their fluency. However, the existing research on use of formulaic sequences in L2 speakers suggests that such usage is limited and inconsistent at best. For example, Paquot and Granger (2012) have reviewed the use of formulaic sequences in L2 and found that, even in more advanced speakers, it is marked by the underuse of referential collocations, multiword verb phrases, and idiomatic usage, and by the overuse of some meta-discursive expressions. This research entailed the identification of formulaic sequences in corpora of L2 language using standardized lists or frequency-based methods. There have also been some explorations of individual usage. For example, in a case study on the speech of an intermediate-level Japanese speaker of English, Wood (2009) found that about 12% of her speech consisted of formulaic sequences.

Overall, research into L2 usage of formulaic sequences is limited and has been hampered by a lack of consistency in the definition of formulaic sequences and how they are identified. In particular, the focus has tended to be on sequences that are considered to be formulaic “in the language” (such as idioms and high-frequency multiword units). Wray (2008) makes the distinction between such externally defined sequences and those which may be “psycholinguistic” units in the lexicon of the individual speaker. A number of researchers (e.g., Dahlmann, 2009; Erman, 2007) have shown that these are not necessarily the same, particularly for L2 speakers. For example, an L2 speaker may know of a particular idiom (which is formulaic in the language) but not be able to use it smoothly. At the same time, a nonidiomatic expression (such as “It’s a real problem”) may become psycholinguistically formulaic for that speaker through frequent repetition. Such formulaic sequences acquired by “fusion” (Peters, 1983; Schmitt & Carter, 2004) may be particularly important for L2 speakers as they represent language that is especially useful and relevant to the individual. Research based on frequency measures or standardized lists within corpora will however tend to miss these (unless it is a corpus of individual usage).

Cordier (2013) defines “psycholinguistic formulaic sequences” as multiword units that present a processing advantage to the individual speaker—either because they are stored holistically or because they are processed automatically as a unit. This definition extends a widely used definition by Wray (2002, p. 9) and facilitates the identification of formulaic sequences (as holistically processed items) on the basis of the spoken output. While there is no direct way to measure the storage or processing of lexical items in the speaker’s mind, the study of the temporal features of speech output can give indication of the nature of language processing (Temple, 2000). One key temporal feature that has been widely used for this purpose is fluency (e.g., Lin, 2010; Towell et al., 1993; Wray, 2002). In particular, the absence of disfluency markers (such as pauses, hesitation, and repetition) was used as a criterion for formulaicity in studies by Erman (2007) and Dahlmann (2009), and also in algorithmic approaches such as those of Brooke, Tsang, Hirst, and Shein (2014). A further approach to identification is the use of diagnostic criteria (e.g., Wood, 2009; Wray, 2008). In this approach, a number of different criteria for formulaicity (such as “there is something grammatically unusual about this word string”) are listed and the satisfaction of one or more of these is taken to indicate that a sequence is likely to be formulaic.

Combining the fluency and diagnostic criteria approaches to identification, a recent study by Cordier (2013) has suggested that psycholinguistic formulaic sequences

may be more prevalent in the speech of intermediate and advanced L2 speakers than previously thought. She analyzed the speech of five advanced British L2 French speakers who had each undertaken a set of different speaking tasks consisting of interviews, discussions, and story narrations. Formulaic sequences were identified by applying the fluency and diagnostic criteria on a hierarchical basis (following Hickey, 1993) meaning that conditions were applied in a strict order. For a sequence to be declared formulaic, it first had to satisfy the necessary condition of fluency and then also had to satisfy at least one diagnostic criterion to indicate that it showed signs of being a holistic unit. The main results using this methodology were that an average 27.7% of her participants' speech was formulaic. In addition, she found significant differences in observed formulaicity across different tasks, with the story-telling task producing fewer formulaic sequences than the interview or discussion tasks.

1.2 Current Study

The current study used the same identification process and hierarchical criteria to investigate the use of psycholinguistic formulaic sequences in the speech of Japanese speakers of English (JSE). Its aim was to estimate the amount and type of formulaic speech used by a particular group of intermediate and advanced JSE and to check how this compared with the previous research. In undertaking a study of this nature, it is important to recognize that any count of formulaic sequences in an individual's speech depends on how they have been defined and the measurement process used. The practical and theoretical issues associated with investigating formulaicity in the speech of L2 speakers will therefore also be discussed.

The three main research questions are as follows:

RQ1 To what extent do psycholinguistic formulaic sequences feature in the speech of these intermediate/advanced JSE, and how does this compare with results from previous research?

RQ2 How does the nature of the task affect the number of formulaic sequences used?

RQ3 What types of formulaic sequences are used by the speakers and how do these contribute to overall formulaicity in these speakers?

An effective method of identifying formulaic sequences in the speech of L2 learners is important if we are to monitor their acquisition and usage. This study, being the first to apply these hierarchical criteria to JSE, provides an opportunity for testing the methodology as well as giving insight into the prevalence of formulaic sequences by this group of speakers.

2. Method

2.1 Participants

The participants were eight JSE, all of whom were volunteer office workers recruited from companies in Japan that were known to the researcher. The participants were chosen on the basis of availability and to provide a mix of background (in terms

of experience and opportunities to use English) and proficiency levels. There were seven females and one male and their ages ranged from 32 to 55. Four of the participants were from the same company, and three of the participants had similar jobs (associated with book-keeping and accountancy). To provide a point of reference, two native speakers from the United Kingdom also undertook the identical process. Both were working adults with occupations unrelated to teaching English or linguistics.

2.2 Procedure

The participants each undertook two speaking tasks: a structured interview about their work lasting 4–5 minutes, and a story narration based on a picture sequence (around 3–4 minutes). For the story, they had a choice of three picture sequences and were given 2 minutes to prepare. Participants were told the nature and timings of the tasks but not the focus of the research. Informed consent was obtained and they were assured about the anonymity of their contributions. All tasks were recorded and transcribed, with pauses and other relevant disfluency marked. Formulaic sequences were then identified according to a set of hierarchical conditions, following the methodology of Cordier (2013). These conditions were applied in three stages to provide a progressive filtering of the transcribed speech.

2.2.1 Necessary: phonological coherence

The first necessary condition was that of phonological coherence, here operationalized as fluent pronunciation. This has been used as a validation measure in the identification process before (e.g., Dahlmann, 2009; Erman, 2007; Raupach, 1984) but not as an initial necessary condition in a hierarchy of criteria. Signs of disfluency were defined to be:

- (1) unfilled pauses > 0.25 seconds
- (2) filled pauses (e.g., *er*, *umm*, *ah*)
- (3) syllable lengthening > 0.4 seconds
- (4) repetition or repair/retracing

The 0.25 seconds cut-off for unfilled pauses follows a standard used frequently in fluency research (e.g., Kormos & Denes, 2004; Lennon, 2000). Filled pauses were taken as nonwords not containing semantic information. For example, lexical fillers (e.g., *you know*, *yeah*) were not taken as filled pauses since they have a function and may themselves be examples of formulaic sequences. The identification of syllable lengthening follows Dahlmann (2009) and was taken to indicate the end of a run. These disfluency indicators were used to segment the speech stream into fluent runs. For example:

SACHI: it's // funny because he // I'm working in the office // and it // it's just he and me // so // when he went on business overseas // I just...

2.2.2 Necessary: At least one typical condition showing a holistic dimension

Fluent runs can potentially be quite long stretches of speech and are not necessarily formulaic in themselves. Indeed there may be several formulaic

sequences along with individual words within a fluent run. Therefore, a further way of identifying the formulaic sequences from within the runs was required. The second necessary criterion defined by Cordier (2013) was that there should be at least one typical condition showing a holistic dimension. The diagnostic criteria used here were adapted from those used by Wray (2008) and Wood (2009) and are as follows:

- (1) Grammatical or semantic irregularity: The meaning of the sequence is not given by its parts, or the grammar of the sequence is not that typically used to express the meaning. Examples from the current study included: “they lived happily ever after,” “bits and pieces,” “can’t handle.”
- (2) Regular sequences with semantic or functional unity: These are typically grammatical units, common collocations, proper names, or other sequences with a clear holistic mapping of form to meaning or function. Examples from the current study included: “in charge of,” “of course,” “on the other hand,” “typical day,” “Toshima Ward.”
- (3) Sequences likely to have been learned or used as a whole by the speaker: This was based in the diagnostic criteria from Wray (2008, p. 116): “based on direct evidence or my intuition, there is a greater than chance level probability that this speaker will have used this precise formulation before in communication with other people”. Examples from the current study included expressions from the speaker’s work experience (e.g., “total administration time,” “TOEIC essay contest”) or ones that they were likely to have learned before (“on the other hand”).

It should be noted that the above criteria are by no means mutually exclusive, and a sequence may satisfy more than one criteria (e.g., “on the other hand” above). This is not surprising since there are a number of potential causal or theoretical links between the criteria. For example, most irregular sequences known to a speaker are likely to have been learned or experienced as a whole. However, evidence of holistcity only requires the satisfaction of one criterion. So, for the purposes of this procedure, no special significance is attached to sequences satisfying multiple criteria.

2.2.3 *Graded: frequency*

A further graded condition used was that of intra-speaker frequency (i.e., does the speaker use the same term repeatedly). In a small speech sample, it is not possible or desirable to use the repetition of an expression as a necessary criterion. However, when expressions are repeated by a speaker, it adds to the likelihood that they are formulaic (assuming the other conditions are also satisfied). For example, one participant said “I’m very surprised” on three different occasions (even when narrating the past).

2.3 *Measures*

Two main measures of “formulaicity” were used. For comparative purposes these were identical to the ones used by Cordier (2013):

- (1) FS% (Percentage of formulaic syllables): the number of syllables in the speech sample that were part of a formulaic sequence divided by the total number of syllables in that sample.
- (2) ANR (Average number of formulaic syllables per fluent run): the number of syllables that were part of a formulaic sequence divided by the number of fluent runs in the speech sample.

The FS% measure gives an overall sense of how much of speech is part of a formulaic sequence, while ANR gives a sense of how they divide up the speech stream. In addition to the formulaicity measures, some standard temporal measures of speech fluency were calculated for each sample in order to explore how formulaicity may vary with fluency. These were the Speech Rate (SR) in syllables per minute, and Mean Length of Runs (MLR) which measures the average length in syllables of a fluent run between disfluency markers (e.g., Kormos & Denes, 2004; Lennon, 2000).

3. Results

Overall, 4,798 words (6,340 syllables) were spoken by the eight participants over the two tasks and 663 formulaic sequence tokens were identified (449 types). These contained 1,685 words (2,285 syllables). The average number of words (syllables) per formulaic sequence was 2.54 (3.56). There were 214 repetitions (22.2%) overall, with 67 tokens (40 types) being repetitions across two or more participants. The most repeated sequences were “for example” (12 tokens across 5 participants), “you know” (11 tokens / 2 participants), and “I think” (9 tokens / 6 participants).

3.1 Types of Formulaic Sequence Used

To explore the different types of formulaic sequence that participants used, sequences were categorized according to a broad typology developed by Cordier (2013). This was chosen to provide a direct comparison with the previous study. In this typology, “Referential sequences” are defined as those predominantly used to refer to entities such as objects, places, times, or ideas. “Meta-discursive expressions” are sequences used to structure, comment on, or engage with the discourse or message, and “Sentence builders” (from Nattinger & DeCarrico, 1992) are the fixed parts of patterns used to build sentences and phrases. The relative distribution of sequences across each category types is given in Table 1 along with examples from the study for each category and subcategory.

3.2 Formulaic Sequence Usage by Task

In order to explore differences in the usage of formulaic sequences across the two tasks, mean values of each formulaicity measure across the participants were calculated. Table 2 shows these values (along with the range for each) for each task and in total. Comparing the two tasks, the results show that more formulaic sequences were used in the first task (the interview

Table 1. Distribution of Formulaic Sequences by Category

Category	Subcategory and examples	No. (%)
Referential	Verb phrase – <i>have to deal with</i>	486 (74)
	Noun phrase – <i>book stores</i>	
	Time/place complements – <i>last year</i>	
	Adverbials – <i>on behalf of</i>	
	Whole clause – <i>they lived happily ever after</i>	
Meta-discursive	Hedges – <i>some kind of</i>	100 (15)
	Fillers – <i>you know</i>	
	Asides – <i>what do I do?</i>	
	Discourse structure – <i>for example</i>	
Sentence builders	<i>I think ___</i>	77 (12)
	<i>I'm not good at ___</i>	
	<i>It's nothing to do with ___</i>	

Table 2. Mean Values (and Ranges) for Both Formulaicity Measures

	Task 1 (Work interview)	Task 2 (Picture story)	Total
FS%	38.2% (33.2–48.1)	31.0% (26.0–38.4)	34.6% (29.6–40.3)
ANR	1.89 (1.03–2.79)	1.39 (0.53–2.64)	1.64 (0.82–2.63)

about their job) than in the second picture narration task. Using a paired *t*-test (two-tailed), these differences were found to be significant ($t=3.14$, $p=0.016$ and $t=3.62$, $p=0.009$) for both of the formulaicity measures (FS% and ANR). For the combined samples, the mean FS% was 34.6% and mean ANR was 1.64. These mean figures are substantially higher than those found by Cordier (2013) whose five advanced French learners had mean FS% = 27.7% (range 22.1–31.0) and mean ANR = 1.50 (range 0.83 – 1.90) over the five tasks they undertook.

3.3 Formulaic Sequence Usage by Participant

A summary of the quantitative measures of formulaic sequence usage and fluency for each participant are given in Table 3, arranged in order of fluency (SR). Note: participants have been given pseudonyms.

As can be seen from the data, formulaicity as measured by ANR (the average number of formulaic syllables per fluent run) increases consistently in line with fluency (SR). In particular, the two participants (Yayoi and Yoko) who had considerable experience (2 years or more) of living overseas also had the highest fluency and ANR scores. On the contrary, the FS% measure does not show a clear pattern with respect to fluency. For example, the participant Wataru has the highest FS% score (40.3%) but was one of the less fluent speakers (SR=97.0) on the tasks. The two native speakers who did the same tasks and followed the same procedure had considerably higher usage of formulaic sequences than all of the participants (FS%=46.4 and 48.1%, ANR= 3.74 and 4.81) and they were also more fluent (SR=182.0 and 195.7). This provides a good validation of the procedure.

Table 3. Summary Fluency and Formulaicity of Participants Over Both Tasks

Participant	Sex/Age	TOEIC*	FS%	ANR	SR (syll/min)	MLR (syll)
Junko	F-40+	650	30.9	0.81	70.9	2.54
Eri	F-50+	735	29.6	0.84	83.6	2.82
Wataru	M-40+	–	40.3	1.44	97.0	3.50
Sachi	F-40+	865	36.0	1.78	115.7	4.96
Kanae	F-30+	940	35.6	1.58	123.4	4.44
Mami	F-30+	–	33.8	1.81	127.3	5.34
Yayoi	F-40+	975	31.9	2.21	148.3	6.80
Yoko	F-40+	960	38.5	2.63	175.9	6.85

MLR, Mean Length of Runs; TOEIC, Test of English for International Communication.

4. Discussion

4.1 Use of Psycholinguistic Formulaic Sequences

Insofar as they can be reliably measured on the basis of the criteria used here, the FS% figures suggest that psycholinguistic formulaic sequences may be a significant part (e.g., 30–40%) of the speech of the JSE participating in the study. The sequences used were mainly referential (verb phrases, noun phrases, time/place complements), accounting for 74% of all sequences. Within this category, there were few repetitions between or within the individual participant samples, and (as in the previous research) there were few examples of grammatically or functionally irregular sequences found. Meta-discursive and sentence building sequences accounted for a smaller proportion of the sequences overall (15 and 12% respectively), but the majority of repeated expressions (e.g., “I think,” “for example,” “you know”) were from these two categories. The distribution of sequences by category and the mostly standard nature of these matches what Cordier (2013) found with her advanced French learners. Overall, the picture of psycholinguistic formulaic sequence usage that emerges is that of the speakers using a breadth of canonical (transparent and grammatical) referential sequences, each being used only once or twice with almost no overlap across participants. These are then supplemented by a number of useful meta-discursive or sentence building expressions which tend to be repeated more, particularly by the participants with higher degrees of formulaicity in their speech.

Regarding the two different tasks that the participants undertook, there was a significant difference in the formulaicity of samples in them. This was true for both formulaicity measures FS% and ANR, with the interview task producing more sequences than the story-telling in each case. This supports the finding of previous research. For example Cordier (2013) found significant differences between all the task types used, with the more interactive interview and discussion tasks yielding more formulaic sequence usage than the narrative task. In the current study, this may be thought to reflect the familiarity of the topics as much as the tasks themselves. In the work interview task, participants tended to use expressions specifically related to their work and experience (e.g., “procedures for foreigners,” “put the cheque in,” “test administration,” “month end” etc.)

which they have likely used frequently before. In the story narration however, the content was not so familiar to the participants and there were likely to be fewer referential sequences easily available to them. On the other hand, when narrating in general, there are potential opportunities to use common sequences for organizing discourse (e.g., expressions for sequencing time and events such as “last year” or “after that”) that the participants could have usefully employed. However, apart from a few examples (e.g., “the next day,” “ten years later”), these were not used extensively by most speakers in this study.

While the distribution of sequences by task, category, and regularity is similar to that found in Cordier’s study, the formulaicity figures in the current study (for intermediate/advanced JSE) are higher than those found for her advanced British speakers of French. Despite the obvious difference that the texts were in different languages in the two studies, the size and direction of the difference in the FS% scores is perhaps surprising. A possible contributory factor may have been a small difference in the pause cut-off length used (0.25 seconds compared to the 0.2 seconds used by Cordier). However, a follow-up analysis on a sample of the sequences identified as formulaic in the study found that none would have been rejected even if a 0.2 seconds cut-off was applied. A further possibility is that, due to the essentially probabilistic and contextual nature of diagnostic criteria, there may be systematic differences in applying the criteria in the second stage of the identification process. This point is explored further in the next section.

4.2 Identification Challenges

Although a consistent and well-defined process was used, the actual application of the method highlighted particular challenges inherent in identification arising from the nature of formulaic sequences themselves and the necessarily interpretative nature of diagnostic criteria. Three particular challenges were illustrated in the study.

4.2.1 Degree of “fixedness” within the sequence

Formulaic sequences may be either fixed or constructed as frames with slots for variables (Wray, 2002). In addition, they may be subject to expansion (e.g., adding an intensifier within the sequence) or nesting (placing one sequence in the variable slot of another). Deciding which of these options is applicable in individual cases can be challenging, and use of the conditions and criteria may not always be able to resolve this. Such decisions are important however since they may affect which words within the string are taken to be part of the formulaic sequence, thereby affecting the quantitative measures of formulaicity. The following example from the study illustrates this challenge.

- (a) YAYOI: it’s partially the subcontractor’s job to train proctors

The expression in (a) was delivered fluently by the participant and therefore satisfies the first condition for being a formulaic sequence. For the second (holistic) condition, either criterion 2b (“has functional or semantic unity”) or 2c (“has been used in the same form to convey the same meaning”) may be applicable. However, they may potentially be applied at different levels of abstraction.

For example, it is possible that the whole expression is formulaic as this is a work-related topic which has clearly been discussed before. On the contrary, it could be that the frame “it’s someone’s job to do something” is formulaic for this speaker, with the (familiar) variables slotted in appropriately and the qualifier “partially” added as an (optional) expansion.

4.4.2 *Dynamic nature of formulaicity*

The study also provided examples illustrating the potentially dynamic and context-based nature of formulaicity in the individual speaker (e.g., Ellis, 2012). For example, Junko in her interview initially appeared to construct the phrase “PR unit” (as the English translation of her department name) and then subsequently used it in a formulaic way.

(b) JUNKO: My job is a PR- (1) unit? (.) I am in PR unit. [...] I think (...) PR unit is very conservative

The phrase “PR unit” does have a semantic unity (Criterion 2b) and is repeated (Criterion 3). So, the two fluent cases of the phrase in the example are taken to be formulaic sequences in the current procedure. However, the evidence of earlier disfluency of the sequence also seems important. For example, here it seems to indicate that the sequence is newly formed and, as such, may only be temporarily available in a holistic form. Other potential indicators of such “temporary formulaicity” may include mixtures of fluent and nonfluent usage of a sequence, or the repetition of a formulaic expression taken from the interviewer’s question. Indeed, examples of both indicators were observed in the current study. The extent to which this kind of contextual information should be applied will depend on the needs of the research and how one views the status of newly formed or temporary formulaic sequences. However, since such decision will affect the count of formulaic sequences observed, it is important for any identification process to be explicit about how it deals with these cases.

4.2.3 *Use of “multiword” as a defining feature*

In most approaches, formulaic sequences are taken to be explicitly “multiword” sequences operating in a unitary way. In such cases, the word is by implication a defining feature of the formulaic sequence. However, as Wray (2014) argues, the concept of the word is not always clear, due to the existence of contractions, polywords, compound nouns, hyphenated words, and so on. While explicit clarifications can be made at the definitional stage (e.g., in this study, contractions, polywords, and hyphenated words are all taken to be multiple words), there were examples from the study that reveal the slightly arbitrary nature of using the word as a defining feature for identification. For example, “test takers” and “a lot of” were included as formulaic but not the single words “examinees” or “many,” even though on definitional criteria they are essentially equivalent. This highlights a challenge in applying a multiword criteria as a definitional feature of formulaic sequences, and is another potential source of difference in the identification process,

4.3 The Formulaicity Measures

Two variables, ANR and FS%, were used in this study to provide a measure of the “formulaicity” of the participants’ speech samples, and the results show a different pattern across the participants for each. ANR (the average number of formulaic syllables per fluent run) seems to have a close association with fluency, with ANR values increasing in line with increasing SR. However, for FS% (the proportion of syllables that were part of a formulaic sequence), there is not such a clear pattern. For example, one participant, Wataru, had a high value for FS% even though he spoke quite hesitantly (as shown by his fluency measures). At the same time, one of the most fluent speakers, Yayoi, had a comparatively low FS% over her two samples. One way to interpret this is to acknowledge that different measures indicate different aspects of performance and processing. For example, researchers (e.g., Towell et al., 1996) have argued that fluency as measured by MLR (i.e., a greater ability to formulate runs) may be due to greater proceduralization in processing (e.g., in the formulator in Levelt’s model of speech production, 1993) and that such proceduralization is facilitated by the use of formulaic sequences. However, how such usage is measured is also important and the results here suggest a possible differentiation of the roles of FS% and ANR.

A case such as that of Wataru, who uses a comparatively high number of formulaic sequences but with high number of disfluent gaps between them (as indicated by low ANR), demonstrates that the proportion of syllables that are formulaic (i.e., FS%) is not necessarily a useful measure of formulaicity to associate with aspects of speech processing such as proceduralization. The FS% figure represents the proportion of speech that is part of a formulaic sequence, but it does not indicate the number and length of sequences or how they fit together into fluent runs (for which ANR may be more appropriate). What this highlights is that although the FS% variable may have intuitive appeal as an apparent measure of how formulaic a speech sample is, it may not be the most appropriate measure for this purpose.

5. Conclusion

This study shows that psycholinguistic formulaic sequences, defined as fluent, semantically or functionally coherent multiword sequences, may be a significant feature in the speech of intermediate/advanced JSE. The results of this first study to use these particular identification criteria on such speakers broadly agree with the main findings of the previous research using the same method, and give some further insight into the prevalence of psycholinguistic formulaic sequences in L2 speakers as well as the practical challenges of identifying these. It also adds further weight to the finding that formulaic sequence usage is sensitive to the kind of task that is used to elicit speech. Overall, the study demonstrates how a systematic hierarchical procedure can be used to identify formulaic sequences in a useful way. In particular, the use of disfluency as an initial criterion provided a clearly quantifiable starting point for identification that can be consistently applied. Examples of sequences used by participants also highlighted some theoretical and definitional aspects of formulaic sequences that will be helpful in making the diagnostic criteria more robust and in interpreting the meaning of formulaicity measures such as FS% and ANR.

At the same time, there are some clear limitations to the study. In particular, this was a small study with specific group of learners which therefore has limited generalizability on its own. In addition, undertaking the procedure highlighted a number of the inherent challenges in identifying formulaic sequences in spoken output. These centered on the dynamic and graded nature of formulaicity and the interpretative nature of diagnostic criteria. Two recommendations for making the process more robust therefore can be proposed. First, ensure that there are explicit, theoretically justified “rules” to cover ambiguous cases (such as when there is a mix of fluent and disfluent examples of the same sequences or when there are multiple interpretations). These help in further standardizing the process. It is also particularly important to use contextual information from the task and from the individual’s speech sample as a whole, and to specify how to apply it. However, even with such refinements, it should be recognized that the diagnostic criteria are based on likelihoods and are not always strictly quantifiable on the evidence available. So, a second important recommendation is to utilize multiple judges to make the diagnostic assessments and to have explicit rules and procedures to deal with disputed cases when pooling the results.

Overall, the study supports the suggestion that the use of psycholinguistic formulaic sequences (as measured by ANR for example) is associated with fluency. An observation from the study was that a principal area of difference in formulaic sequence usage between participants with higher and lower ANR (and fluency) was in the use of meta-discursive and sentence starter expressions and their repetition. In particular, higher fluency participants tended to use (and repeat) a greater number of general discursive expressions (sequencers, hedges, and fillers) and longer types of sentence building patterns. This suggests that a useful focus, even for the higher level JSE in this study, may be to support them in becoming fluent in the production of a prioritized set of such formulaic sequences, in order to enhance their output delivery.

References

- Brooke, J., Tsang, V., Hirst, G., & Shein, F. (2014). Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 753–761), Dublin, Ireland, August 23–29 2014.
- Cordier, C. (2013). *The presence, nature and role of formulaic sequences in English advanced learners of French: A longitudinal study* (Unpublished doctoral thesis). Newcastle University, Newcastle, UK.
- Dahlmann, I. (2009). *Towards a multi-word unit inventory of spoken discourse* (Unpublished doctoral thesis), University of Nottingham, Nottingham, UK.
- Ellis, N.C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, 32, 17–44. doi:10.1017/S0267190512000025
- Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics*, 12(1), 25–55. doi:10.1075/ijcl.12.1.04erm

- Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (ed.), *Perspectives on fluency* (pp. 25–42). Ann Arbor, MI: University of Michigan Press.
- Hickey, T. (1993). Identifying formulas in first language acquisition. *Journal of Child Language*, 20, 27–41. doi:10.1017/S0305000900009107
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164. doi:10.1016/j.system.2004.01.001
- Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 43–60). Ann Arbor, MI: The University of Michigan Press.
- Levelt, W.J. (1993). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Lin, P.M.S. (2010). The phonology of formulaic sequences: A review. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 174–193). London, UK: Continuum.
- Nattinger, J.R. & DeCarrico, J.S. (1992). *Lexical phrases and language teaching*. Oxford, UK: Oxford University Press.
- Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130–149. doi:10.1017/S0267190512000098
- Pawley, A., & Syder, F.H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C. Richards & R.W. Schmidt (Eds.), *Language and communication* (pp. 191–225). London, UK: Longman.
- Peters, A.M. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.
- Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp.1–22). Amsterdam, Netherlands: John Benjamins.
- Temple, L. (2000). Second language learner speech production. *Studia Linguistica*, 54(2), 288–297. doi:10.1111/1467-9582.00068
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–119. doi:10.1093/applin/17.1.84
- Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics/Revue Canadienne de Linguistique Appliquée*, 12(1), 39–57.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press
- Wray, A. (2008). *Formulaic language: pushing the boundaries*. Oxford, UK: Oxford University Press.
- Wray, A. (2014). Why are we so sure we know what a word is? In J. Taylor (Ed.), *The Oxford handbook of the word*. Oxford, UK: Oxford University Press.