Usage to grammar or usage and grammar? Comparing the development of a learner-external and a learner-internal formula with L2 learners' propositional language (use) across proficiency levels: a corpus-based pilot study

by Thomas Antony Hammond

British Council's Master's Dissertation Awards 2021

Commendation

www.teachingenglish.org.uk/publications/elt-masters-dissertations

Usage *to* grammar or usage *and* grammar? Comparing the development of a learner-external and a learner-internal formula with L2 learners' propositional language (use) across proficiency levels: a corpus-based pilot study.

Table of contents

Table of figures

I) Introduction

This dissertation presents a pilot study which combines concepts from recent formalist research regarding the role of formulaic language (FL) in second language acquisition (SLA), namely those of Myles & Cordier (2017) and Bardovi-Harlig & Stringer (2017)[1]. The role of FL in SLA has been the subject of continuous debate; on one side researchers such as Wong-Filmore (1976), Eskildsen & Cadierno (2007) and Ellis (2012) place FL at the centre of their usage-based sequence of acquisition, which posits that learners extract and therefore acquire syntax from target, model formulas. Alternatively, researchers such as Hanania & Gradman (1977), Krashen & Scarcella (1978) and Bohn (1986) argue that holistic processing advantages[2] to lower-level learners are the sole beneficiaries of FL, as essentially syntax develops independently. Myles & Cordier (2017) state that a common issue with research on both sides of the debate, however, is that they often fail to systematise the concept of

[1] The impact of these studies has not gone unnoticed in the literature; both of which have been deemed essential for 'advancing our understanding of the topic in important ways' (Wulff 2019: 20).

[2] See chapter VII at the end of this paper for a glossary of such terms.

formulaicity, which results in a lack of direction and therefore limits the implicational domain of their findings. Following Wray (2008), the authors pose a distinction between what is formulaic in a given language (*linguistic clusters*) and what is formulaic in the learners' mind (*processing units).* Bardovi-Harlig & Stringer (2017) offer new, empirical evidence to bear on the nature of the former's role in SLA; through examining the development of conventional expressions, the authors' findings suggest that these do not take the form of idealized target models as catalysts for acquisition, but are instead 'reflective of autonomous syntactic development' in line with the learners' interlanguage grammar (63). Adopting the aforementioned dichotomy of formulaicity, the present study identifies both a processing unit (*I don't know*) and a linguistic cluster (*how much is it*) in a learner corpus of English and compares the development of these across proficiency levels alongside learners' use and accuracy of their compositional syntactic properties in the learners' corresponding propositional language. It is only through such a comparison that transparencies between the production of each type of formula and the learners' grammatical competence can be appropriately identified, and implications regarding the role of FL in SLA can be addressed more comprehensively.

The study presents results which support Myles & Cordier's (2017) distinction of the two types of formula as separate phenomena; benefits of a more 'holistic' storage and faster processing seem to be reserved for the processing unit only, whilst the development of the linguistic cluster manifests similarly to the conventional expressions in Bardovi-Harlig & Stringer (2017), showing accuracy of fixed lexical categories and difficulties with functional ones which reflect the learners' interlanguage grammar. The data shows that whilst the processing unit can be used as a memorised routine and allows lower-level learners to engage in more advanced linguistic performance, this is not reflective of their generative competence, and the two concepts should therefore be viewed as separate. It is the latter of these concepts which seems to be responsible for the production and modification of both types of formulas at higher levels, that is, we see the formulas integrated into propositional language but only when generative competence is appropriately developed. It is also posited that generative competence could be responsible for how learners conceptualise the formulaicity of the language around them, and it is suggested that learners could actually benefit from early analysis of formulas, rather than holistic production. The implications drawn from the results present an alternative to the usage-based sequence of acquisition; rather than formulas driving grammar it would seem that the production of the formulas themselves for these particular learners under investigation are somewhat constrained by respective grammatical competence. Independent development of such therefore seems to be what determines the learners' use and accuracy of syntax, rather than the extraction of patterns from fluent productions of an unedited, target formula.

This dissertation will therefore proceed as the following. Chapter II reviews the relevant literature and summarises key points from the two papers on which this investigation is built, chapter III outlines the research question and hypotheses that can be predicted from such, and chapter IV describes the methodology and theoretical framework adopted for the analysis along with the respective limitations of this initial 'pilot' investigation. Chapter V presents the results and discussion in light of the relevant hypotheses, and chapter VI looks at any wider implications of these with respect to the role of FL in the second language acquisition process. Conclusions are finally offered in chapter VII.

II) Literature Review

1) Defining and identifying formulaic language: traditional perspectives

The attention FL has been given in recent years is reflective of the abundance of descriptional and definitional terms in the literature that have been used to refer to its characterisation. Wray (2002) notes 60 terms used to describe aspects of formulaicity, some of them involving a measure of 'conceptual duplication' but many of them referring to different phenomena depending on the approach/stance being taken[3]. Traditional definitions embrace the notion of 'a multimorphemic unit memorised and recalled as a whole, rather than generated from individual items based on linguistic rules' (Myles et al. 1998: 325), which encapsulates the fact that FL is a phenomenon realized through a process of pragmatic inferencing, involving semantic and phonetic reduction until the contribution of the individual component parts are redundant, and the expression is instead conceptualized as a 'whole' (Wray, 2008). The most widely cited and traditional definition is that of Wray (2002), stated below:

'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' (p. 9).

[3] For example, Myles & Cordier (2017) note that 'chunk' is often used in psycholinguistic research whereas 'clusters' is favoured in corpus-linguistics (p. 5).

Arguably as problematic as defining FL is its identification in a language; Ellis (2012) poses the determination of criteria which should be used to correctly identify FL as *frequency, association* and *native norms*. That strings of language which occur often should be considered formulaic is not as straightforward as it seems (Biber, Conrad & Cortes, 2004), as the fact that a formula is above a certain *frequency* threshold does not necessarily imply either psycholinguistic salience or coherence (Ellis 2012). *Association* can often be measured statistically, for example, the Mutual Information (MI) test is a common tool used in corpus linguistics which assesses the degree to which the words in a phrase occur together more frequently than would be expected by chance (Oakes, 1998), where a higher MI score indicates a stronger association between the words. Like with frequency, statistical measures of association come with their disadvantages (Evert 2005); the MI score for example can give too much prominence to rare combinations of words which are in fact not very frequent in real numbers, and conversely, some nouns have such a high overall frequency that they occur in combination with many keywords without having any 'mutual' relation to them at all[4] (Lindquist, 2009). The term *native norms* refers to FL reflecting native-like selection and native-like fluency (Pawley & Syder 1983), a notion which in itself is hard to define and systematize amongst the abundance of contexts and environments where a language is used.

The picture that emerges from the above then is a rather opaque and ambiguous one, with such a wide-ranging and encompassing conceptualization of 'formulaicity' leading to a lack of clarity and unification in both definition and identification alike. Despite these

[4] Lindquist (2009) uses the British National Corpus (BNC) as an example, where *'time'* is the most frequent noun with 18,330 occurrences, meaning it is likely to occur close to many words solely by chance (p. 76).

inconsistencies, it is somewhat consensual that a mastery of FL is what makes up a very large portion of a learner's ability to succeed in a second language (Fillmore 1979), which has given rise to a large body of research investigating its role in the SLA process. Traditionally, FL has been associated (and central within) usage-based approaches to language acquisition, which are now given a brief overview in the following section.

2) Usage *to* grammar: the role of formulaic language in usage-based proposals of (second) language acquisition

Usage Based Language (UBL) approaches differ from the traditional dichotomy of syntax and lexicon and propose a mental lexicon in which 'abstract grammatical patterns and the lexical instantiations of those patters are jointly included, and which may consist of many different levels of schematic abstraction' (Tummers, Heylen, & Geeraerts, 2005: 228-229). Such models of both L1 and L2 acquisition pose a popular learning sequence as being from 'formulaic phrase, to limited-scope slot-and-frame pattern, to fully productive schematic pattern' (Ellis, 2012: 18), effectively an extension of one of the foundations of Emergent Grammar, that 'structure is not an overarching set of abstract principles, but more a question of a spreading of systematicity from individual words, phrases, and small sets' (Hopper 1987: 143). On this account, grammar is essentially acquired from the statistical abstraction of patterns of form-meaning correspondence which correlate to their usage experience, which means that FL serves as the main input for the acquisition of syntax.

Although this learning sequence is extendable to both L1 and L2 acquisition, this paper will make reference only to the latter, as this is ultimately what concerns the present investigation.

The 1970's saw some of the first investigations into the proposed usage-based sequence of L2 acquisition, notable studies include Hakuta (1974) and Wong- Fillmore (1976) who argue that L2 learners (particularly children) start with prefabricated patterns which they break down into compositional parts to extract rules governing their L2 before creative language use ensues. More recently, Eskildsen & Cadierno (2007) proposed how one instantiation of the formula *'I don't know'* was the basis for initial stages of language use before this was expanded on and generalised across to other verbs and pronouns. Mellow's (2008) longitudinal case study of Ana saw how she also initially used a limited number of complex constructions that were connected to a particular set of verbs, before she gradually extended these across an increasing range of constructions.

Whilst the idea that the linguistic input learners receive undoubtably has a substantial influence on their second language acquisition (Ellis & Wulff 2015), and indeed the more often certain items co-occur, the more 'entrenched' that construction becomes in the learners' mind (Divjak & Caldwell-Harris 2015) all of the studies cited above -bar that of Eskildsen & Cadierno (2007)- have investigated younger (child) learners at the earliest stages of development, which leaves little implication of the role FL plays in adult SLA and throughout the acquisition process as a whole.

3) Usage *and* grammar: formulaic language and grammar as independent processes

On the other side of the debate, several researchers have argued that syntax develops independently of FL, labelling the two as separate phenomena existing as alternative communication strategies, with FL predominantly used as a short-term production tactic to fulfil pragmatic/discourse functions. Hanania & Gradman (1977) noted that at the start of their study, Fatmah (a 19- year old Arabic speaker living in the U.S) produced utterances that seemed to consist predominantly of memorized items that are commonly used in certain social contexts with children. The authors stated that these were merely strings of sounds that she had managed to appropriate in particular discourse situations; '*thank you'* and '*do you like'* for example, were conceptualized as single units, but there was no evidence to show that Fatmah recognized the individual component parts within these structures, nor could she use these words in new combinatory constructions. Krashen & Scarcella (1978), in their review of the literature on studies investigating the role of FL in SLA up to that point, concluded that most research supported the position that although prefabricated routines *may* evolve into patterns, the creative construction process develops independently alongside these. Bohn (1986) continued in a similar vein when looking at a younger naturalistic L2 learner, showing how his 8-year-old subject Heiko at early stages of development used modal auxiliaries i.e. '*would you like'* as a short term production tactic only; that is, this construction was not used to acquire the auxiliary '*would*', and no such learning strategy was detected throughout his development. Granger (1998) was also in line with this concept, branding such quick and advanced production tactics as the sole beneficiaries of FL, as 'there does not seem to be a direct line from prefabs to creative language' (p. 157). More recently, Myles (2004) and Wray & Fitzpatrick (2008) also accept the use of FL as a memorization effort to accomplish communicative needs; both studies look at how this can lead to overrepresentation of linguistic knowledge in L2 learners, as memorized repertoires of lexical strings are often used to suit conventional communicative situations which are associated with certain strings of language.

It would seem then that the role of FL in the usage-based sequence of acquisition is not as 'putative' (Ellis 2012: 29) as is often claimed in the literature. Indeed, when giving an essentially usage-based review of previous studies documenting the use of FL in language acquisition, Wulff (2019) concludes by acknowledging that 'FL may play a lesser role in L2 acquisition compared to L1 acquisition; formulaic sequences likely matter more to some learners than others and are therefore best seen as an optional rather than a strictly required route for L2 acquisition' (p. 30).

What can be said of studies on both sides of the debate is that they often come under scrutiny from their over-generalisation of claims due to FL being used as such a broad term (as introduced in section 1). For example, Bohn (1986) claims that Wong- Fillmore's (1976) formulaic frame structure of 'can + PRN + VP' to explain the child's derivation of the utterances *'can you give me one of those'*, *'can we take 'em home now'* and *'can I read this one'* is an instance of her extending the notion of 'formulaic frame structure' so far that it becomes vacuous. Bohn states that on the basis that utterances with one lexical item in common in one position are considered formulaic (like the examples above), then formulaicity 'is nothing short of a pervasive phenomenon in the speech of learners and of competent speakers' (1986: 191). Similarly, a large amount of terms used by studies on the other side of the debate differ in their terminology when seemingly wanting to review and refer to the same phenomenon i.e. *'prefabs'*, *'routines'*, *'constructions'*, which results in a similar lack of clarity and direction.

It is clear then, that in order to better understand the role of FL in the L2 acquisition process, a more precise and systematic understanding of its definition and identification must be in place. A proposal of such from very recent formalist work attempts this, details of which are now given in the following section.

4) Recent developments in formalist approaches to formulaic language and its role in SLA: learner-external and learner-internal formulas

4.1) Problems with the traditional definition

Myles & Cordier (2017) take issue with Wray's (2002) traditional definition; the claim that there is no 'generation or analysis by the language grammar' yet the fact a sequence can be 'discontinuous' (p. 9) is somewhat contradictory, as if a sequence is discontinuous, i.e. a frame with slots for insertion of variable items, 'it is difficult to conceive that no grammatical processing is taking place at all' (Myles & Cordier 2017: 19). The authors also take issue with the fact that the traditional definition has become somewhat of an 'umbrella term' for FL (Weinert, 2010), a notion that Wray herself deems as problematic, as often conclusions are drawn from studies about FL in general when in fact the approaches taken only deal with a certain type of formulaicity (Wray 2012). For example, collocations and idiomatic expressions used by both native speakers and L2 learners refer to what is formulaic in a *given* *language*, whereas sequences of language which are stored and processed holistically by a learner refer to what is formulaic within an *individual's conceptualisation* (Myles & Cordier 2017). Clearly these are different phenomena, yet researchers tend to use the term 'FL' collectively to refer to both instances.

Taking inspiration from Wray's (2008) distinction of the above dichotomy of formulaicity as *speaker-external* and *speaker-internal*, Myles and Cordier (2017) state that although overlap can be expected between what is formulaic in a given speaker and what is formulaic in the language around the speaker, the two types of formulaicity are 'nonetheless different phenomena and must be investigated as such' (p. 5), as the former refers to an internal cognitive process and the other to an external linguistic phenomenon. In light of this, the authors offer new terminology to distinguish between the two constructs; learner-external FL are retermed as *linguistic clusters* (LC), and learner-internal FL as *processing units* (PU). Their definitions are given below.

> a) *linguistic clusters*: multimorphemic clusters which are either semantically or syntactically irregular, or whose frequent co-occurrence gives them a privileged status in a given language as a conventional way of expressing something.
>
> b) *processing units*: a multiword semantic/functional unit that presents a processing advantage for a given speaker, either because it is stored whole in their lexicon or because it is highly automatised.

(Myles & Cordier 2017: 12)

As is reflected in the above definitions, identification of processing units would seem of greater complexity due to their individualistic and intrinsic nature. The next section presents the authors' proposed criteria for the identification of such.

4.2) Identification of processing units (learner-internal FL)

The definition of a PU posed in (b) is more cautious than Wray's (2002) of FL, in that the emphasis is on the processing *advantage* rather than complete holistic storage. This decision was made largely to adhere to methodological soundness, as whilst 'it is not possible to reliably prove holistic storage, it is less problematic to demonstrate the faster and easier processing of certain sequences of words in relation to others' (Myles & Cordier 2017: 10). Such a definition also fits better with the idea of formulaicity as a graded phenomenon rather than a categorical one, a popular notion with many researchers (see for example Coulmas 1994; Ellis 2012; Wulff 2019). The authors also state that a 'crucial' element of PU identification is that in learner productions, these are seemingly far more advanced than their propositional language, a concept which has not gone unrecognized previously (see for example Myles 2004; Wray & Fitzpatrick 2008).

Necessary Criterion- Phonological Coherence

Myles & Cordier (2017) state that the primary criterion needed for identification of PU's is *phonological coherence*. This criterion is 'primary' in that any additional criteria must only be applied on the subset of candidate PU's that show phonological coherence (p. 19), rendering the identification method they propose a hierarchical one. The importance of such derives from the fact that 'utterance fluency' (Segalowitz 2010), which is based on the temporal and phonetic variables of speech, gives the best indication of a learners' underlying cognitive processes of language production (Rehbein 1987). Since PU's are essentially an internal cognitive process, it follows that their pronunciation should be compatible and representative of this (i.e. phonologically salient). In other words, something cannot be considered formulaic if not pronounced fluently, as this implies that online grammatical processing is taking place.

Additional criterion- Semantic/Functional Unity

Semantic and/or functional unity can refer to a wide range of sequences that achieve a set function in their usage, including time expressions *'last year'*, '*at the moment'*, sequences to introduce one's opinion '*I think that'*, as well as semantically irregular sequences '*it's raining cats and dogs*' and sequences which have their unity in their function of fillers '*I don't know'* (Myles & Cordier 2017). The authors also state that even sequences that are not grammatically unified, in that they cannot be classified as a constituent (Borjars & Burridge 2013), can express semantic and/or functional unity i.e. '*out of the'*, '*because of the*' and '*a sort of*'. Such sequences are traditionally associated with frequency-based approaches in corpus linguistics (Ebeling & Hasselgard 2015) and have been labelled with a variety of terminology previously in the literature, such as *'incomplete phrases'* Lindquist (2009), and ‘*lexical bundles’* (Gries 2008). The authors justification for inclusion of phrases which lack grammatical unity is that many of these do carry a holistic quality because in their entirety they can be mapped onto one functional goal; for example, the phrase *‘I think that’* is composed of [NP+ VP + C] yet its unified function is to ‘introduce one’s opinion’ (Myles & Cordier 2017: 20).

Reinforcing Criteria- Frequency

As PU’s are learner-specific, frequency counts can only be taken into consideration in the productions of the individual learner(s) investigated, that is, ‘the fact that a sequence is frequent in other corpora is no guarantee that it will be part of a particular learners formulalect’ (Myles & Cordier 2017: 21). As is implicit with their definition of PU’s, frequency is to be considered a graded criterion, so that the more frequent a phrase is within the same learner’s production, the more reliable its status as a PU. Following (Ejzenberg 2000), the authors also accept inter-learner frequency, that is, frequency across a homogenous set of learners, where homogeneity refers to proficiency, background and educational experience.

4.2) Learner- internal approach to formulaicity in SLA

It is proposed that PU’s are the type of FL that should be investigated in L2 learners when it comes to understanding the role FL plays in L2 acquisition and the idea that such sequences act as seeds for the development of syntax, as it is essentially these types of formulas that learners conceptualise as a unit, and that consequently present a processing advantage. Such sequences may or may not be formulaic in the language a speaker is learning[5], but this is somewhat irrelevant, as 'what we are investigating is not how L2 learners appropriate or not externally defined FL (LC's), but how chunking processes operate in L2 learning' (Myles & Cordier 2017: 8). Support for such can also be found in the pragmatics literature; Kecskes (2019) states that 'from the perspective of intercultural pragmatics only those linguistic units should be considered formulaic that have some psychological saliency for the language user' (p. 145). Indeed, studies that have exclusively investigated the role and development of LC's in L2 acquisition have delivered mixed results. Schmitt and colleagues have been instrumental in research into the access and use of LC's by both native-speakers (NS) and L2 learners in recent years, including collocations, idioms and lexical bundles. Schmitt, Grandage and Adolphs (2004) looked at corpus-derived sequences and found that their results suggested only a minority of the target sequences were stored holistically by L2 learners, and Siyanova-Chanturia, Conklin, & Schmitt (2011) used eye-tracking to investigate the online processing of idioms, to find that these presented a processing advantage to NSs only, not L2 learners. Building on from such, Conklin & Schmitt (2012) when examining the access and use of idioms and collocations found that whilst native speakers of a language benefited from processing these holistically, L2 learners processed the strings in a similar fashion to propositional language, i.e. word to word. They concluded by stating that the collocations and idioms were only processed quicker if the phrases were known, and that for

[5] The authors note that native speakers (NS) have usually automatized the formulaicity in the language around them, but this cannot be assumed for L2 learners (p. 25).

L2 learners, when the FL is idiomatic, the figurative meanings are actually more difficult to process than nonidiomatic language.

It therefore seems unlikely that LC’s present processing advantages for L2 learners which further emphasises their distinction from PU’s; in the case of idioms and collocations, the above studies show that learners often have more difficulty with these than propositional language, which weakens any implications that learners are using these types of FL as seeds to extract syntax. Indeed, when applying their PU criteria (outlined in 4.2) to a large corpus of advanced L2 learners of French, Myles & Cordier (2017) noted that irregular or highly idiomatic sequences represented a very small minority of the units identified, with most of the strings identified as PU’s being grammatically regular. The *development* of PU’s and LC’s should therefore be investigated as independent phenomena, if we are to better understand their overarching role in L2 acquisition.

The most comprehensive investigation of the development of LC’s in the L2 acquisition process comes from very recent empirical data; Bardovi-Harlig & Stringer (2017) examined learner production of various conventional expressions across proficiency levels. Details of such are now given in the following section.

5) The development of a learner- external type of formulaicity (conventional expressions): Bardovi-Harlig & Stringer (2017)

5.1) The experiment

Bardovi-Harlig & Stringer (2017) selected a number of conventional expressions and tested a variety of learners at different levels of proficiency who were attempting to produce the same target in the same communicative context. The conventional expressions were selected by a data driven process; field observations of spontaneous conversations in the community where the study was conducted allowed for creation of scenarios for oral conversation simulation, and then two pilot studies assured that NSs responses showed a single favourite conventional expression in such scenarios. These were operationalized as 'native-speaker use greater than 50%' (Bardovi Harlig & Stringer 2017: 69).

In total, there were 271 L2 learners of English who participated in three different cross-sectional studies, ranging from low-intermediate to low-advanced levels of proficiency (levels 3-6), representing a range of L1 backgrounds including Central-Asian and Indo-European languages. The first task was a time-pressured oral production one, where respondents listened and responded to 32 scenarios over individual headsets whilst simultaneously reading them on a screen. These data were supplemented by data from an experiment used previously (see Bardovi-Harlig & Vellenga 2012), which looked at the effect of instruction on the acquisition of two sets of conventional expressions, consisting of the same oral conversation simulation for both the pre-test and post-test[6]. The final set of data was taken from an untimed, self-paced task which involved aural recognition and self-

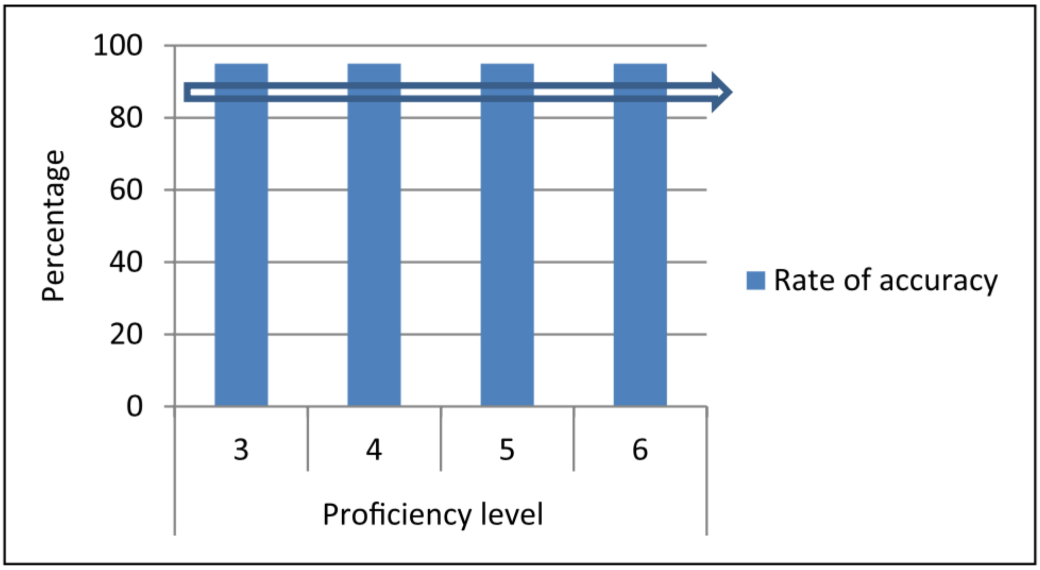
[6] In this second study, 36 students of 11 L1s participated, providing 1,152 responses for each of the pre and post- test (Bardovi-Harlig & Stringer 2017).

assessment (Bardovi-Harlig 2014). Here, learners heard 20 expressions and chose from options which best represented their knowledge of the expression, and then participated in a ‘written elicited imitation that revealed their interlanguage structure for the expression presented’ (Bardovi-Harlig & Stringer 2017: 71).

5.2) Results

The results from all tasks showed that many expressions are ‘learned early, score high for accuracy and do not substantially change with proficiency’ (Bardovi- Harlig & Stringer 2017: 73), expressions which we could reasonably perceive as candidates for PU’s for these individual learners (Myles & Cordier 2017). These include *‘nice to meet you’*, *‘you too’* and *‘thank you’* which show the trajectory below.

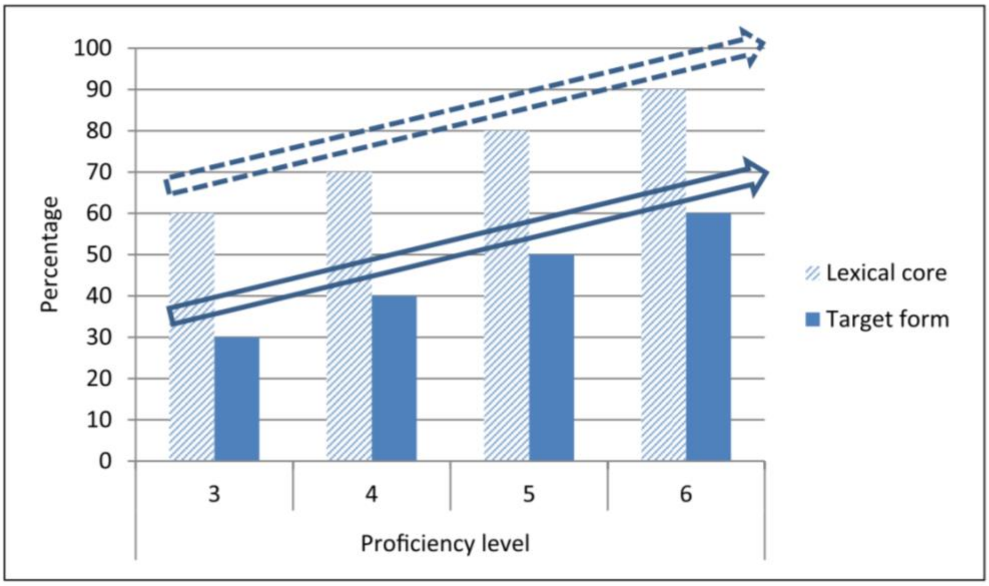
(1)



(p. 70)

However, many conventional expressions composed of three or more morphemes showed gradual development across proficiency levels towards accurate production, which included *'sorry I'm late'*, *'that'd be great'* and *'I really appreciate it'*. More specifically, this involved production in appropriate contexts showing 'gradual acquisition of a lexical core of a formula that is not fully grammatically specified and is filled in by the learner's interlanguage grammar' (Bardovi-Harlig & Stringer 2017: 79), represented in figure 2 below.

(2)



(p. 79)

For example, interlanguage forms for *'I really appreciate it'* included *'I appreciate'*, *'I'm appreciate'*, '*I appreciate for you'* and '*I will appreciate it to you'*. (Bardovi-Harlig & Stringer 2017: 77). The authors state that sociopragmatic knowledge and pragmalinguistic knowledge are at play here, as the former allows learners to recognise the context as appropriate for the targeted conventional expressions in the task[7], and the latter determines their linguistic resources available for the realization of the speech act (Bardovi-Harlig & Stringer 2017). Such linguistic resources are dependent on their interlanguage grammar, which is what is left to fill in the 'fuzzy' functional slots around the lexical core, meaning that, ultimately, the conventional expressions show transparency to the interlanguage

[7] This is also dependent on if the learner was familiar with the target expressions (p. 73).

grammar throughout the acquisition process. Regarding the overarching role of FL in the L2 acquisition process, these results entail that the formulas of this kind (LC's) in this particular instance 'do not take the form of idealized models of grammatical well-formedness acting as catalysts for acquisition, but are reflective instead of autonomous syntactic development' (Bardovi-Harlig & Stringer 2017: 63).

5.3) Limitations of the experiment and implications for the present study

The above experiment is essential for our understanding of what Myles & Cordier (2017) would define as learner-external (LC) formulas and their development in the L2 learning process, which does not correspond to that proposed in usage-based sequences of acquisition previously outlined in section 2, and instead supports the acquisition of syntax as independent from formulas as presented in section 3. However, any implications drawn from this study surrounding the overarching role of FL in SLA can be based only on LC's, which, as we saw in section (4.2), are not the type of formulas that should be perceived (if any) as candidates from which syntax is inferred, these are instead suggested to be PU's in taking a learner-internal approach (Myles & Cordier 2017). The authors themselves are aware of such a limitation; 'we recognize that conventional expressions and acquisitional formulas are clearly distinct phenomena, and our data only bring evidence to bear on the nature of the former' (Bardovi-Harlig & Stringer 2017: 84), (where 'acquisitional formulas' correspond to PU's). Whilst they do note that certain acquisitional formulas are produced early and accurately, and do not change with proficiency, no attention is given to these and they are instead referred to in passing.

Further, the authors claim that certain conventional expressions show ‘transparency to the learners’ interlanguage grammar’ (p. 63), yet the study tested learner productions of these expressions *in isolation*, meaning that there was no productive evidence of the learners’ propositional language that could be compared with to see if their interlanguage productions are indeed in line with their overall grammatical competence. Also, using learners from only a subset of proficiency levels subsequently limits any general claims wanting to be made regarding the role of conventional expressions across the acquisition processes as a whole. A more comprehensive insight could be given through using a wider range of proficiency levels, i.e. beginner through to upper-advanced.

III) Research Question

In light of the above, this paper aims to conduct a corpus-based pilot study investigating the role of FL in the acquisition of syntax for adult L2 learners of English, building on from the limitations of Bardovi-Harlig & Stringer (2017) presented above, and assuming the dichotomy of formulaicity proposed by Myles & Cordier (2017). It will do so by identifying both a PU and a LC in a learner corpus, and comparing the development of these in selected learner production files across proficiency levels alongside associated syntactic properties in their propositional language. Hypotheses that can be made ahead of the pilot study based on the conclusions drawn from the literature review are below in (a) and (b).

1) Predictions/hypotheses

(a) The *processing unit* (learner-internal formula) is produced accurately by all selected learners and across all proficiency groups under analysis (by virtue of showing phonological coherence). The PU remains the same in form across all proficiency levels (no interlanguage variations are found), and shows the trajectory presented in Fig. (1), whilst the propositional language complexity increases across proficiency levels. When the PU is used, this is to achieve a set function in the discourse upon a socio-pragmatic contextual cue and is evidence of a faster processing strategy used by the learner. Therefore, in the lowest proficiency group under analysis, the same syntactic complexity of the PU cannot be found to be used, or indeed used accurately, in the propositional language structures of the same speaker, and instead errors are found with these.

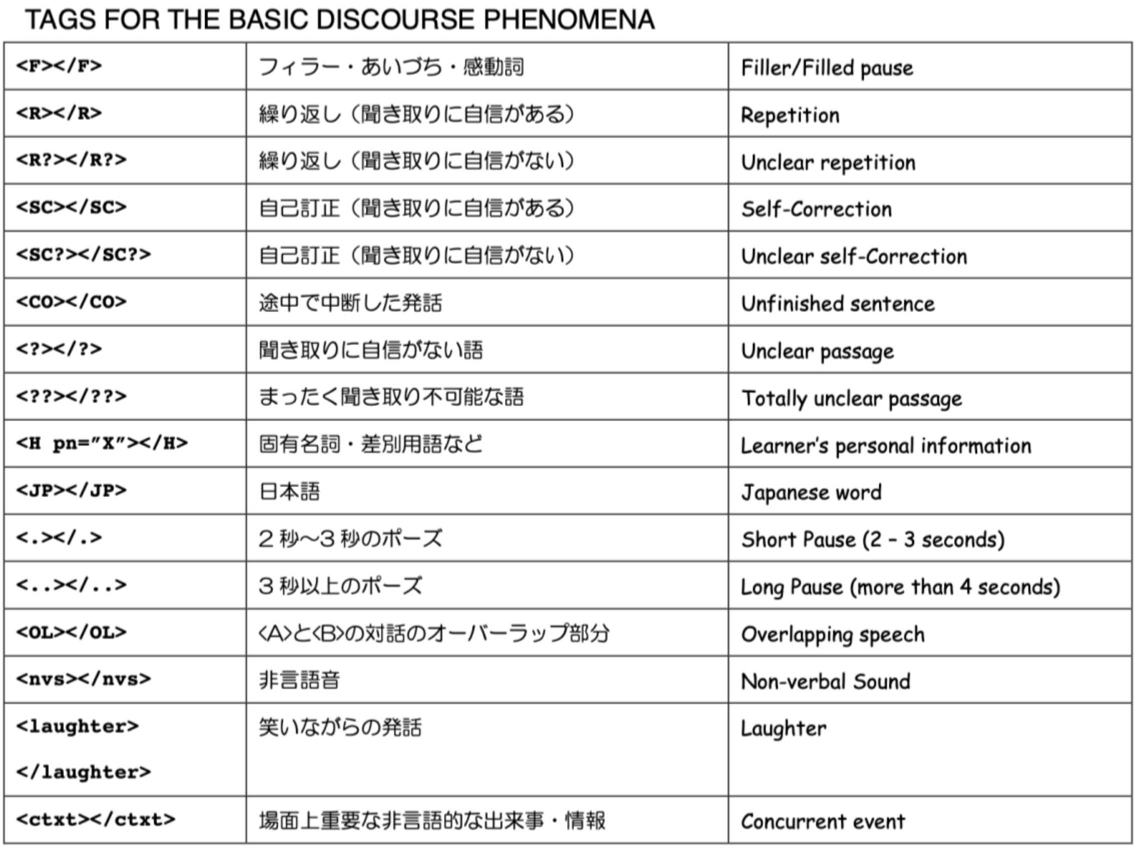
(b) The *linguistic cluster* (learner-external formula) will follow the pattern presented by the data in Bardovi-Harlig & Stringer (2017). Interlanguage variations of the LC are therefore identified, and when these occur, they show commonalities with errors made in propositional language, reflective of autonomous syntactic development and transparency to the learners' grammatical competence. Interlangauge variations of the LC will include fixed lexical elements (a lexical core) and inaccuracy with functional categories, which will improve across proficiency levels towards a target structure, in line with propositional language development following the trajectory in Fig. (2).

IV) Methodology

1) The JLICT Corpus

The corpus used as the source of production data in this study is the Japanese Learner English (JLE) Corpus constructed by the National Institute of Information and Communications Technology in 2004, which contains transcripts of 1,281 audio-recorded speech samples resulting in a makeup of 1.2 million words. The speech samples are of English oral proficiency interview tests based on the Standard Speaking Test (SST), which is a collaboration between the American Council on the Teaching of Foreign Languages (ACTFL) and the ALC Press (a Japanese language learning and publishing company). The SST distinguishes 9 proficiency levels based on the criteria of text type, accuracy, pronunciation, fluency and overall task and function, and takes the form of a 15 minute conversation between a test candidate and interviewer, who uses various techniques and picture prompts to 'stimulate natural conversation to the maximum extent possible in a testing situation' (Tono et al. 2001: 2). This element of the test type is advantageous for the present study, as more natural conversation gives for a more accurate a representation of the learners' propositional language. The interview tasks are also consistent throughout the proficiency levels; all interviews follow a five-stage format consisting of warm up, picture prompt, role play, picture sequence prompt and wind-down, meaning it is possible to identify common functional language which is used in the same contexts throughout the whole corpus (see section 2.2 of this chapter for how this aspect of the corpus is advantageous for identification of the LC). Further, the corpus contains annotations (tags) of relevant prosodic and discourse phenomena, such as *long/short pauses, speaker-repetition/self-correction,* which are instrumental in the recognition of FL that is phonologically coherent, a criterion of central importance to the identification of PU's (presented in section II: 4.2). The relevant tags are given below in (3).

(3)



TAGS FOR THE BASIC DISCOURSE PHENOMENA

<F></F>	フィラー・あいづち・感動詞	Filler/Filled pause
<R></R>	繰り返し（聞き取りに自信がある）	Repetition
<R?></R?>	繰り返し（聞き取りに自信がない）	Unclear repetition
<SC></SC>	自己訂正（聞き取りに自信がある）	Self-Correction
<SC?></SC?>	自己訂正（聞き取りに自信がない）	Unclear self-Correction
<CO></CO>	途中で中断した発話	Unfinished sentence
<?></?>	聞き取りに自信がない語	Unclear passage
<??></??>	まったく聞き取り不可能な語	Totally unclear passage
<H pn="X"></H>	固有名詞・差別用語など	Learner's personal information
<JP></JP>	日本語	Japanese word
<.></.>	２秒～３秒のポーズ	Short Pause (2 - 3 seconds)
<..></..>	３秒以上のポーズ	Long Pause (more than 4 seconds)
<OL></OL>	〈A〉と〈B〉の対話のオーバーラップ部分	Overlapping speech
<nvs></nvs>	非言語音	Non-verbal Sound
<laughter> </laughter>	笑いながらの発話	Laughter
<ctxt></ctxt>	場面上重要な非言語的な出来事・情報	Concurrent event

(JLEC 2004)

Another advantage of using the JLEC is that, essentially for this study, it includes each learner's proficiency level based on the SST scoring method. This means that spoken language production ranging from absolute beginner through to proficient levels of competence can be readily accessed and compared, making it possible to analyse and compare FL and the characteristics of propositional language at each developmental stage.

2) The pilot study: limitations of time and scope

Initally, the plan was to identify 6 formulas (3 PU's, 3 LC's) through manually examining each transcript in the corpus, and making a note of sequences of words which seemed to be used often by the learners, as the corpus is made up of individual data files only and does not come as part of an interface which can be manipulated as a whole. As can be expected, this was a rather time-consuming process that suffered from a lack of direction and principle, as strings of words were being chosen largely through intuition as what I deemed 'formulaic', suffering from the same deficit of FL identification as presented in section II: 1. Such issues led me to narrow down the scope of the investigation; the decision was made to take a sample approach (Biber 1993; Leech 2007), whereby 3 broadly distinct proficiency levels would be chosen and 5 learners would be selected from each of these to compare both their use of the FL under investigation alongside their propositional language. The decision was also made to search for two formulas only, one that could be classified as a PU and one as a LC. This leads to a total of 30 transcripts under analysis (15 learners who use the PU, and 15 who use the LC), which is rather a reduced data sample. This consequently infers that the study is somewhat *corpus-informed* (McEnery & Hardie 2012), in that it cannot adhere to 'total accountability' as not all the corpus is used to address every part of the research question (Leech 1992: 112). Whilst the whole corpus is used to identify the two formulas (see sections 2.1 and 2.2 of this chapter), it is specific learners chosen as part of a data selection sample whose propositional language will be analysed alongside these. In Bardovi-Harlig & Stringer (2017), the intentions of their study was not to cast generalisations but rather to present empirical evidence as an alternative to usage-based models of acquisition for those *specific learners* in their analysis. The present study's intentions can be viewed in a similar light, as an instance where 'we may seek in a corpus a specific example which, in itself, falsifies a hypothesis- thereby making the totality of the data in some sense irrelevant' (McEnery &

Hardie 2012: 16). It is hoped however that the sample chosen for this analysis, when the degree of homogeneity is considered across the corpus[8], can adhere to at least some level of representativeness as for its results to be indicative of significant implications if the study were to be carried out across the whole corpus.

Whilst corpus studies benefit from the generalisability that comes with analysis of large data sets, there are some factors which actually credit the use of such a small-scale, data selected sample. Examination of the syntactic complexity of each leaner's propositional language involves far more labour-intensive and detailed 'qualitative' analysis, for which the use of smaller amounts of data is possible (McEnery & Hardie 2012) and indeed commendable, considering the given the time-frame (for this dissertation project). In this sense, the corpus is being used similarly to that of many Critical Discourse Analysis (CDA) oriented research, where the corpus is seen as a repository of examples (Flowerdew 1997) from which a small amount of data is analysed taking into account not just the text, but the context in which it was produced and interpreted.

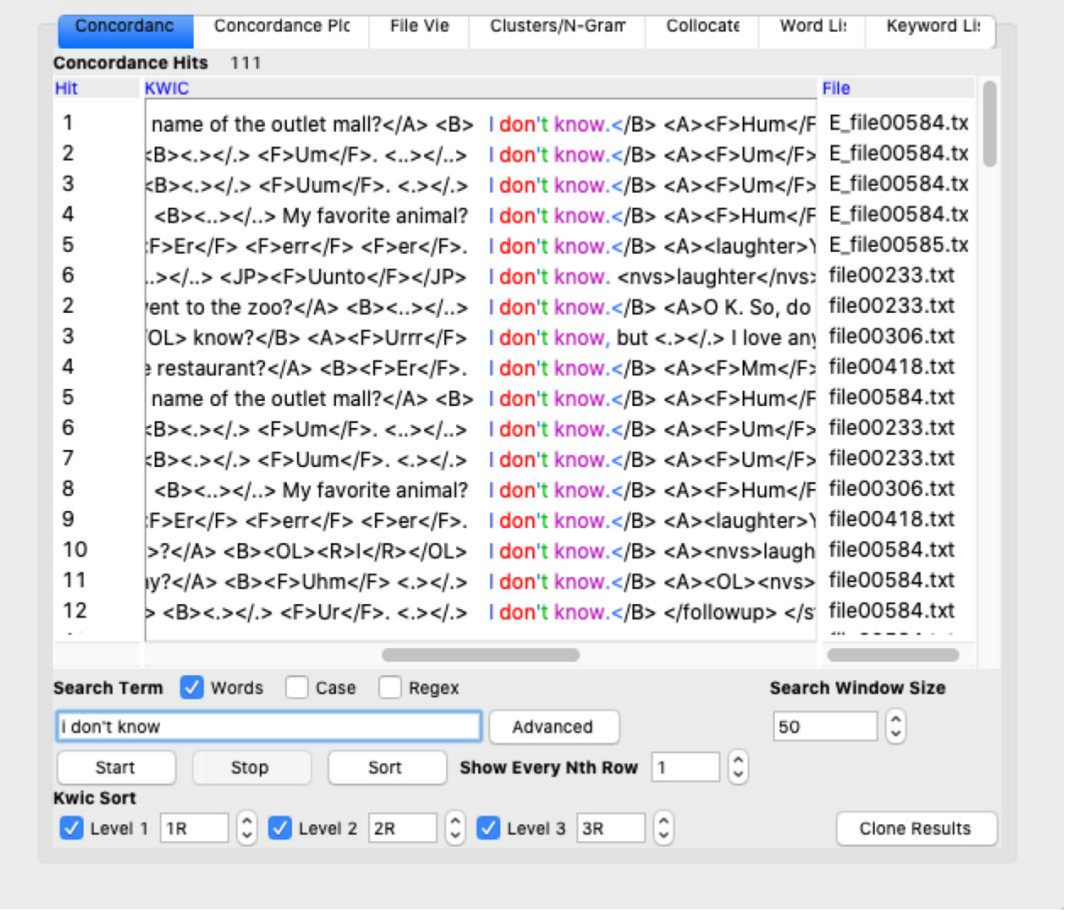
The limitations borne from time and scope presented above are why this investigation is being presented as a pilot study, with its main purpose restricted to presenting results from individual learners which are encouraging of insightful implications that could be drawn upon through following up the experiment on a large-scale basis. The sections below now document the identification process for both the *processing unit* and *linguistic cluster* alike.

[8] All learners are from the same L1 background (*Japanese*), of the same age category (*adult*) and are participating in the same context/interview task (*SST Proficiency test*).

3) Processing unit- ‘I don’t know’

Myles & Cordier (2017) state that sequences which have their semantic unity in their function of fillers such as ‘*I don’t know’* are often good candidates for PU’s in individual learners. Moreover, their particular example of *‘I don’t know’* is made up of three words, a combination which is encouraging of faster psycholinguistic processing as ‘sequences of two to five words are the most salient ones in natural language, yielding phraseologically interesting units’ (Ebeling & Hasselgard 2015: 209). In light of such, this formula seemed a good place to start. Instead of manually searching through each learner file for instances of *‘I don’t know’* as mentioned as an initial strategy in the previous section, all the corpus files were uploaded to the AntConc software (Anthony 2014), which acts as a central interface and allows you to search for concordance lines of multiword phrases and shows these in a Key Word In Context (KWIC) view. A screenshot of such a view is given below for exemplary purposes.

(4)



Concordance Hits 111

Hit	KWIC		File
1	name of the outlet mall?</A> <B>	I don't know.</B> <A><F>Hum</F	E_file00584.tx
2	:B><.></.> <F>Um</F>. <..></..>	I don't know.</B> <A><F>Um</F>	E_file00584.tx
3	:B><.></.> <F>Uum</F>. <.></.>	I don't know.</B> <A><F>Um</F>	E_file00584.tx
4	<B><..></..> My favorite animal?	I don't know.</B> <A><F>Hum</F	E_file00584.tx
5	:F>Er</F> <F>err</F> <F>er</F>.	I don't know.</B> <A><laughter>\	E_file00585.tx
6	.></..> <JP><F>Uunto</F></JP>	I don't know. <nvs>laughter</nvs:	file00233.txt
2	ent to the zoo?</A> <B><..></..>	I don't know.</B> <A>O K. So, do	file00233.txt
3	OL> know?</B> <A><F>Urrr</F>	I don't know, but <.></.> I love any	file00306.txt
4	restaurant?</A> <B><F>Er</F>.	I don't know.</B> <A><F>Mm</F:	file00418.txt
5	name of the outlet mall?</A> <B>	I don't know.</B> <A><F>Hum</F	file00584.txt
6	:B><.></.> <F>Um</F>. <..></..>	I don't know.</B> <A><F>Um</F>	file00233.txt
7	:B><.></.> <F>Uum</F>. <.></.>	I don't know.</B> <A><F>Um</F>	file00233.txt
8	<B><..></..> My favorite animal?	I don't know.</B> <A><F>Hum</F	file00306.txt
9	:F>Er</F> <F>err</F> <F>er</F>.	I don't know.</B> <A><laughter>\	file00418.txt
10	>?</A> <B><OL><R>I</R></OL>	I don't know.</B> <A><nvs>laugh	file00584.txt
11	y?</A> <B><F>Uhm</F> <.></.>	I don't know.</B> <A><OL><nvs>	file00584.txt
12	> <B><.></.> <F>Ur</F>. <.></.>	I don't know.</B> </followup> </s	file00584.txt

Search Term ☑ Words □ Case □ Regex

I don't know

Search Window Size 50

Show Every Nth Row 1

Kwic Sort

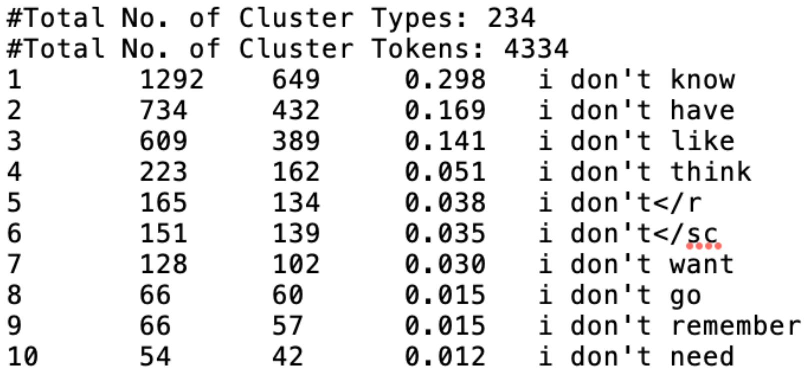
☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

The screenshot in (4) shows how, at least for these learners, *'I don't know'* is used as a filler, as the phrase appears in isolation as opposed to part of subordinate constructions, and often between gaps of speech (note that <F></F> indicates filler pause, <.></.> indicates a short pause of 2 seconds and <..></..> a longer one of up to 4 seconds- see fig 3). It also shows that there are no pauses, repairs or speaker repetition and correction within the production of *'I don't know'*, and indeed this was a theme that ran throughout the majority of the learners in the corpus. On this basis, we could therefore infer that *'I don't know'* is pronounced with phonological coherence. Note here, however, that another limitation of both the corpus and the pilot study should be addressed. The shortest pauses indicated in the corpus are those of 2-3 seconds (fig 3), yet Myles & Cordier suggest that for true phonological coherence, a 'fluent run' should be ideally operationalized as 'a multiword sequence pronounced without filled or unfilled pauses longer than 0.2 second' (2017: 19). The current pilot study can therefore only assume phonological coherence based on the discourse/prosodic information given in the corpus and use this in combination with the other criteria (semantic functional unity, frequency) in classification of a PU. Such a limitation is another reason why the authors' own example of a good candidate for a PU -'*I don't know'*- was originally selected for analysis. A study that builds from this pilot one should not be so assuming however; the corresponding audio files of the learner transcripts could be loaded into the Praat software[9], which can precisely measure the prosodic duration of the formula produced by the learners, to thus check the 0.2 second criterion of phonological coherence more systematically. Issues of time and scope did not permit this stage of the identification process for this pilot study however, so this note of caution and future direction will have to suffice.

The final (reinforcing) criterion of frequency, in this case also inter-learner frequency, can be checked through the N-grams feature in AntConc, which allows for identification of the preferred learner constructions after *'I don't'* across the whole corpus in terms of *frequency*, *range* and *probability*. This can be seen below in (5).

(5)

[9] Praat is a computer program created by Boersma & Weenik (2005), with which you can analyse, synthesise and manipulate speech using audio files.



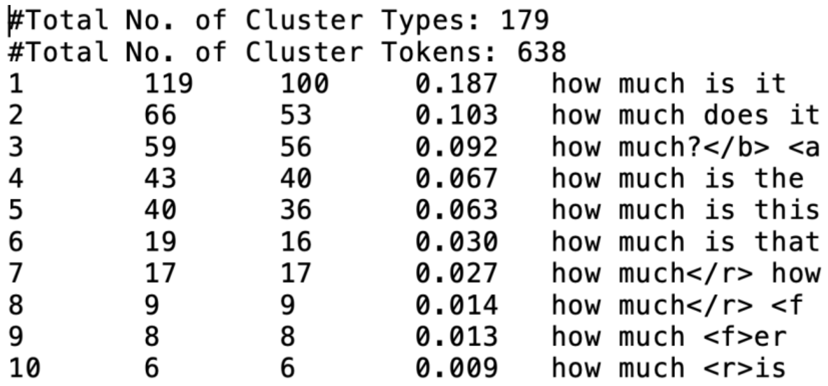
```
#Total No. of Cluster Types: 234
#Total No. of Cluster Tokens: 4334
1       1292    649     0.298   i don't know
2       734     432     0.169   i don't have
3       609     389     0.141   i don't like
4       223     162     0.051   i don't think
5       165     134     0.038   i don't</r
6       151     139     0.035   i don't</sc
7       128     102     0.030   i don't want
8       66      60      0.015   i don't go
9       66      57      0.015   i don't remember
10      54      42      0.012   i don't need
```

This clarification of frequency across the corpus, along with its unity as a functional filler and fluent (uninterrupted) pronunciation, renders- for the sake of this pilot study- *'I don't know'* as the PU which will be under investigation. Note that this methodology of identification is essentially a combination of the phraseological approach (top down) and frequency-based one (bottom up) (Ebeling & Hasselgard 2015: 209), as it involves a predefined sequence of words (*I don't know*) that has been tested against corpus tools (AntConc) to give confirmation of its formulaicity to the individual learners of the corpus. This combination of approaches in learner corpus research has been labelled as 'combining the best of two worlds' and a 'step in the right direction' (Granger & Paquot 2008: 45), so it is this approach which is also used to identify the LC. Details are such are now documented in the next section.

4) Linguistic cluster- 'how much is it'

The LC’s selected for analysis in Bardovi-Harlig & Stringer (2017) were driven by field observations of spontaneous conversations in the community (see section II: 5.1). This essentially means that frequently used language attached to specific discourse scenarios was their starting point for identification of a linguistic unit, a methodology which is commendable, especially when it is considered that many communicative situations often have FL attached to them (Schmitt 2010). As mentioned in section 1 of this chapter, the JLEC corpus makes use of many of the same task-type scenarios across the proficiency levels, one of which is a ‘travel’ role-play involving the learners at one stage having to enquire about the price of the journey/ticket. Initial skims of the learner files who partake in this particular interview task indicates that upon this socio-pragmatic cue, learners are using the phrase *‘how much’* followed by some variational construction such as ‘*does it cost*’ ‘*is it*’ ‘*is this’* etc. The N-grams feature on AntConc allows for identification of the preferred construction after ‘*how much*’, again in terms of frequency, range and probability. This is shown below in (6).

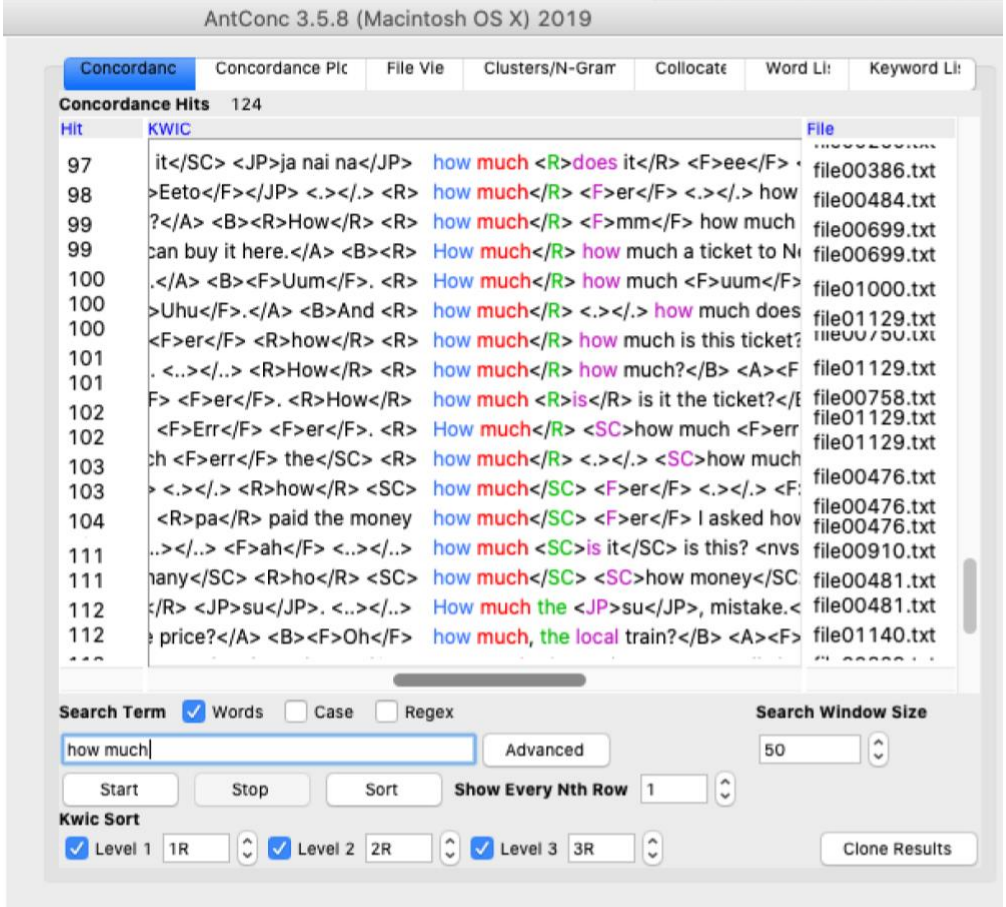
(6)



```
#Total No. of Cluster Types: 179
#Total No. of Cluster Tokens: 638
1       119     100     0.187   how much is it
2       66      53      0.103   how much does it
3       59      56      0.092   how much?</b> <a
4       43      40      0.067   how much is the
5       40      36      0.063   how much is this
6       19      16      0.030   how much is that
7       17      17      0.027   how much</r> how
8       9       9       0.014   how much</r> <f
9       8       8       0.013   how much <f>er
10      6       6       0.009   how much <r>is
```

Figure (6) shows that the variant used most by the learners in this corpus to ask the cost of something is '*how much is it'*. Section II: 5.1 also stated how Bardovi-Harlig & Stringer (2017) refined their selection of conventional expressions through two pilot studies which corresponded them with native speaker selection of a preferred variant. This stage was able to be replicated in the present study through use of a referential corpus, that is, 'a corpus designed to be representative of a language in order to provide comprehensive information about the language' (Cheng 2011: 217). The most widely recognised representative corpus is the 100 million-word British National Corpus (BNC), which is said to give the clearest picture of the English language as it contains a large amount of both written and spoken language across various genres, registers and contexts. When searching for the phrase *'how much is it'*, in the BNC, this returned 66 hits (0.59 per million) compared with '*how much does it cost*' with 24 hits (0.21 per million) and '*how much is this*' with only 13 hits (0.12 per million). In light of the above, '*how much is it'* would seem an appropriate LC to analyse in the transcripts of those speakers who participated in the 'travel' task. What is more- unlike '*I don't know'*- when looking at the concordance lines of '*how much is it'* in the KWIC view, many learners do not pronounce this string coherently, with evidence of pauses, learner repetition and correction internal to the production of the formula. An example screenshot is below in (7).

(7)



Concordance Hits 124

```
Hit  KWIC                                                                                File
97   it</SC> <JP>ja nai na</JP>   how much <R>does it</R> <F>ee</F>               file00386.txt
98   >Eeto</F></JP> <.></.> <R>   how much</R> <F>er</F> <.></.> how             file00484.txt
99   ?</A> <B><R>How</R> <R>      how much</R> <F>mm</F> how much               file00699.txt
99   can buy it here.</A> <B><R>  How much</R> how much a ticket to N            file00699.txt
100  .</A> <B><F>Uum</F>. <R>     How much</R> how much <F>uum</F>              file01000.txt
100  >Uhu</F>.</A> <B>And <R>     how much</R> <.></.> how much does            file01129.txt
100  <F>er</F> <R>how</R> <R>     how much</R> how much is this ticket?         file00758.txt
101  . <..></..> <R>How</R> <R>   how much</R> how much?</B> <A><F              file01129.txt
101  F> <F>er</F>. <R>How</R>     how much <R>is</R> is it the ticket?</        file00758.txt
102   <F>Err</F> <F>er</F>. <R>   How much</R> <SC>how much <F>err               file01129.txt
102  ch <F>err</F> the</SC> <R>   how much</R> <.></.> <SC>how much             file01129.txt
103  <.></.> <R>how</R> <SC>      how much</SC> <F>er</F> <.></.> <F:           file00476.txt
103   <R>pa</R> paid the money    how much</SC> <F>er</F> I asked hov            file00476.txt
104  ..></..> <F>ah</F> <..></..> how much <SC>is it</SC> is this? <nvs          file00476.txt
111  nany</SC> <R>ho</R> <SC>     how much</SC> <SC>how money</SC:              file00910.txt
111  </R> <JP>su</JP>. <..></..>  How much the <JP>su</JP>, mistake.<           file00481.txt
112  price?</A> <B><F>Oh</F>      how much, the local train?</B> <A><F>          file00481.txt
112                                                                              file01140.txt
```

Search Term ☑ Words □ Case □ Regex

how much

Search Window Size 50

Show Every Nth Row 1

Kwic Sort ☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

Such a lack of phonological coherence confirms that, certainly for these learners, '*how much is it'* is not perceived as a PU, giving further support for its categorization instead as a LC.

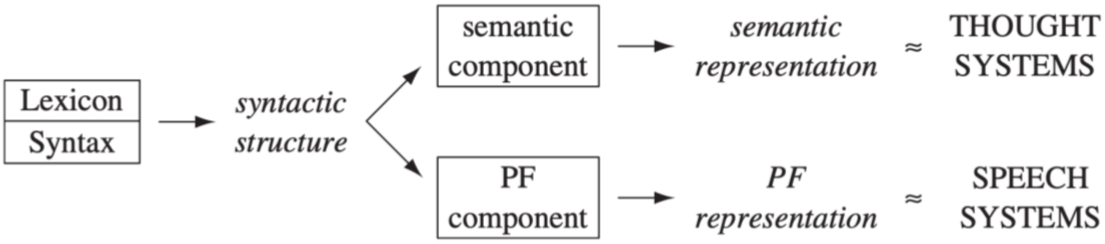
Now that both types of formula have been identified, section 3 outlines the theoretical framework that will be assumed when examining their syntactic structure and associated properties, which is largely based on Radford (2009).

5) Theoretical framework and syntactic properties of selected formulas

This investigation assumes the theoretical framework of U(niversal) G(rammar) in its 'current incarnation', Minimalism (see for example Chomsky 1995), as this would seem to be able to account for both developmental stages of interlanguage grammars and the final-state ones of native speakers (Slabakova 2008: 85). Unlike usage-based models (see section II: 2) the basis of this framework sees the organisation of 'grammar' as a dichotomy of lexicon and syntax, the latter of which serves as input into both the semantic component[10] (mapping syntactic structure into a corresponding semantic representation) and the PF component (mapping syntactic structure into a phonetic form) (Radford 2009). This core organisation of the framework is demonstrated in the diagram below.

[10] This corresponds to Chomsky's (1981) more traditional Logical Form/ LF representation.

(8)



(Radford 2009: 14)

At the heart of syntactic structure is the operation merge, whereby phrases and sentences are built in a bottom-up fashion by merger operations (Chomsky 1982), each of which combines a pair of constituents together to form a larger one (Radford 2009). This section therefore presents syntactic structure in binary-branching tree diagrams which contain information about hierarchical structure[11] (i.e. containment/constituent structure relations) only (Yang 1999).

Note that the following does not contain great detail regarding the principles surrounding the derivation of the structures, but rather acts as a description of the structure and properties

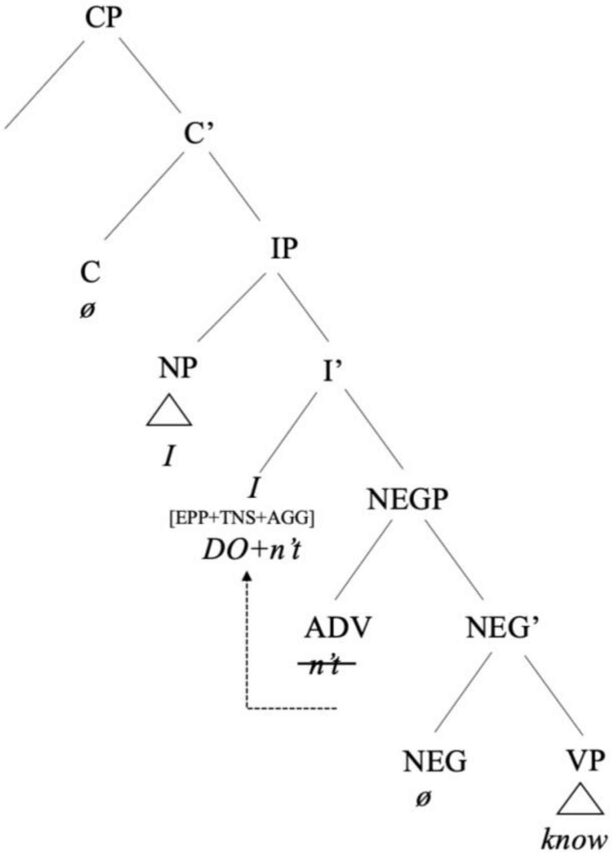
[11] As opposed to linear structure (word order) which is essentially redundant as ‘it can be predicted from hierarchical structure by simple word-order rules’ (Radford 2009: 46).

associated with the formulas under analysis, which can in turn be identifiable in the learners' propositional language. This is the basis on which the learners' generative competence can be measured, and by virtue of this, whether or not their production of the FL under investigation is indeed transparent with individual grammatical competence. A structural/syntactic description of the FL is also fundamental when wanting to draw on any implications regarding generalisation from these structures, that is, whether or not the learners are extracting- and therefore acquiring- syntactic structure from these.

5.1) 'I don't know'- structure and proprieties

The assumed syntactic structure for the processing unit *'I don't know'* is given in the tree in (9).

(9)



The derivation is as follows. The VP '*know'* is formed and then merged with a null NEG head (ø) to form the NEG-bar constituent, merging with the negative clitic '*n't'* which serves as its specifier, forming the NEGP '*n't ø know'* (Radford 2009). On this view, the structure in (8) assumes the negative particle *not* as the specifier of NEGP, not VP (see for example Rizzi 1990; Chomsky 1995). This analysis is given plausibility through historical perspectives[12]

[12] In several earlier varieties of English, sentences containing *not* also contained the negative particle *ne*, where *ne* would serve as the head NEG constituent of NEGP with *not* as its specifier (Radford 2009).

and the lack of accountability which comes with the traditional analysis of *not* in specVP, in examples such as '*he may not be coming tonight'*, where *not* instead appears to occupy a position between the modal auxiliary *may* and aspectual auxiliary *be* (Radford 2009).

The NEGP is then merged with I which contains a Tense [TNS] feature (affix) attracting the negative clitic *'n't'* to attach to it, leaving behind and deleting its original occurrence in spec-NEGP. I also contains an Extended Projection Principle [EPP] feature (Chomsky 1981) requiring it to have a specifier that is a nominal expression i.e an *extended projection* into an IP containing a (syntactic) subject; in our case, the PRN *'I'*. This EPP feature works in conjunction with an Agreement [AGG] feature, so that the syntactic subject must agree with I in person/number, resulting in the syntactic structure *'I+ [TNS]Af + n't + ~~n't~~ + ø + know'*. As the TNS affix feature on I is stranded, in that it cannot find a verbal host to attach to because NEG is neither overt nor verbal, when this structure is sent to the PF component it is spelled out instead as 'an appropriately inflected form of the dummy/expletive auxiliary DO' (Radford 2009: 168), a property commonly referred to as DO-support (O'Grady 2011). This means that DO + *[TNS]Af* + *n't* is consequently spelt out as *don't*, with the combinatory EPP, TNS and AGG features outlined above contributing to the resulting PF spell-out as '*I don't know*'[13]. Note that (9) also assumes the null complementiser analysis (Radford 2009), which determines that all finite clauses are CP's headed by either a null or overt complementiser (in this case it is the former).

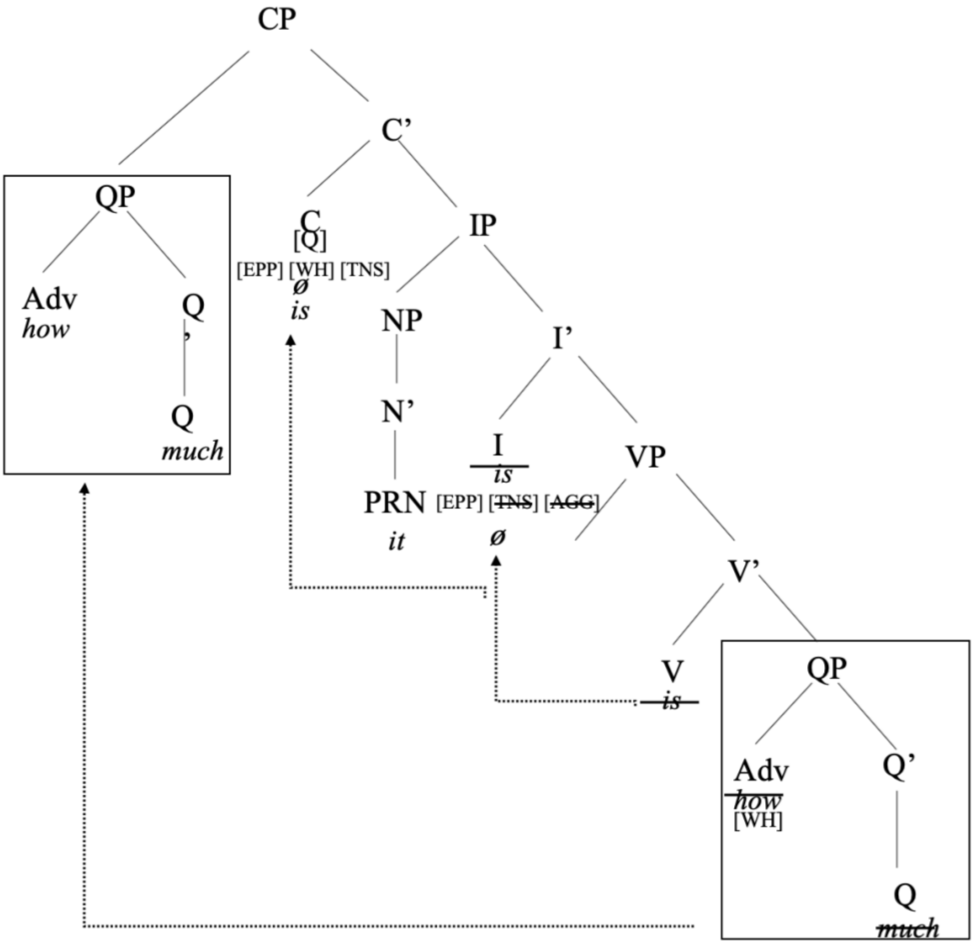
[13] In this sense, *'n't'* is essentially a PF enclitic, as it attaches to the end of an immediately preceding auxiliary host (inserted DO) in the PF component (Radford 2009).

In light of the above, the syntactic properties that will be under analysis for those selected learners who use '*I don't know*' across proficiency levels are DO-support, cliticization, category NEGP and the features [EPP], [TNS] and [AGG]. As stated in section 3 of this chapter, the analysis will also look at how these are generalised, that is, all of the properties outlined above used in combination outside of the processing unit in a novel propositional language structure.

5.2) 'How much is it'- structure and properties

The derivation of *'how much is it'* is shown in the corresponding tree structure below in (10).

(10)



'How much' is a quantifier phrase (QP) made up of the quantifier '*much'* with '*how*' as its specifier, which carries a [WH] feature by virtue of being a wh-word/expression[14]. The wh-word at this stage of derivation is *in-situ*, in that the QP is a complement of the head V *'is'*.

[14] A wh-expression is an expression containing an interrogative word beginning with wh- i.e. *what, which, who* etc, but also encompasses *'how'* where its behaviour is syntactically similar (Radford et al. 1999).

As *'is'* is an auxiliary verb, and there is no modal auxiliary in the sentence, it is raised to I, a derivation known as Verb Raising (O'Grady 2011). Like in (9), I contains an [EPP] feature requiring it to have a nominal syntactic subject as its specifier, in this case the PRN '*it*', deriving the structure '*it + is + ~~is~~ [EPP] [~~TNS~~][~~AGG~~]+ how [WH]+much',* which is headed by a CP in light of the null complementiser analysis. In root interrogatives/main clause questions, the null feature on C carries a [WH] feature requiring the clause to contain a wh-expression. Like I, C also has an [EPP] feature, requiring C to have a specifier matching the [WH] feature (i.e. a wh-word) in order for the resulting structure to be interpreted as a question[15] (Radford et al. 1999). Such features trigger wh-movement, that is, movement of the *smallest accessible constituent* containing a [WH] feature- in accordance with the convergence principle- to spec CP, meaning that the wh-word is now *ex-situ*. The convergence principle accounts for why the whole QP '*how much*' is moved to spec CP and not just the wh-word *'how'* in isolation, a process that is sometimes referred to as pied-piping (Ross 1967). This leads to the structure '*how [WH]+much + it + is + ~~is~~ [EPP] [~~TNS~~][~~AGG~~]+ ~~how~~ [WH]+~~much~~*, where the original occurrences of '*is*' and *'how much'* are left as null copies, in line with the copy theory of movement (Chomsky 1995). As our LC is a main clause/root question, C additionally carries a [TNS] feature which attracts the auxiliary in I to move to C, attaching to a null affixal interrogative complementiser (Radford 2009), a process known as inversion. This derives the final structure of *how [WH]+much+ is + it + ~~is~~ + ~~is~~ [EPP] [~~TNS~~][~~AGG~~]+ ~~how~~ [WH]+~~much,~~* ultimately spelt out as '*how much is it'*.

In light of the above description, the syntactic properties that will be under analysis for those selected learners who use the LC *'how much is it'* are wh-movement (including

[15] In accordance with the 'Interrogative Condition' (see Radford et al. 1999; Radford 2009).

convergence/pied piping within this), inversion and features [EPP], [TNS] and [AGG]. As with the PU, the analysis will also look at the generalisation of the properties outlined above used in combination outside the LC in novel propositional language structures. Note that the examination of wh-movement in the learners' propositional language structures is not limited to main clause/root interrogatives; it will look also at other manifestations of wh-movement, namely relative clauses (including free-relatives and 'that' relatives) and interrogative complement clauses. This is in the hope of giving a more systematic interpretation of the learners' competence/acquisition of the property as a whole.

Now that both the PU and LC have been identified along with their associated syntactic properties, the next section documents the specific data extraction procedure.

6) Data extraction

For the PU, learner levels *2*, *5* and *8* were loaded into the Antconc software independently, and *'I don't know'* was searched for throughout each level as a whole. This allows you to see all instances in each level where the formula is produced with phonological coherence (see 2.1 of this chapter), providing verification that, for the sake of this investigation, '*I don't know'* is indeed used as a PU for these particular learners. Out of these learners, 5 were picked form each proficiency group at random for analysis of corresponding syntactic properties in their propositional language (see 5.1 of this chapter). This resulted in the analysis of 15 learners in total who used the PU.

Levels 2, 5 and 8 were selected in the hope that comparison of the PU and propositional language structures across a broad range of proficiency levels would give for a more comprehensive insight into their relationship and development, building on a limitation of the methodology in Bardovi-Harlig & Stringer (2017) presented in section II 5.3.

For the LC, learner levels *3*, *5* and *7* were loaded into the Antonc software independently, as these were the levels which contained the 'travel' task where learners had to enquire about the price of a journey/ticket, hence priming the discourse/pragmatic contextual cue '*how much is it*' (see section 4 of this chapter). The proficiency levels chosen for analysis of the PU were not constrained in this way (i.e. by task-type), as '*I don't know*' is predominantly used as a functional filler throughout the corpus and is a common feature of spoken language. This meant that many learner productions from all proficiency levels were available for analysis of *'I don't know'*, which allowed for a slightly wider range of comparison. However, the intermediate level production for both types of formulas is consistent (level 5), and the range of proficiency levels for the LC analysis is still sufficient enough to presume a significant difference between developmental levels. In each of these levels, of the learners who produced some variation of *'how much is it'*[16] in this appropriate discourse context, 5 were again chosen at random for analysis of their corresponding propositional language, resulting in 15 learners in total.

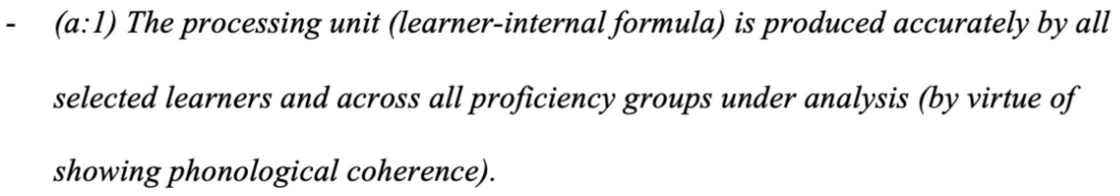
[16] Either accurately or inaccurately, as phonological coherence is not a fundamental characteristic of linguistic clusters (see section II: 5).

The results and relevant discussion of both analyses in light of the hypotheses stated in section III: 1(a) and (b) are now given below in chapter VI, before a discussion of any wider implications that can be drawn from such in chapter VII.

V Results and discussion

1) Processing unit: 'I don't know'

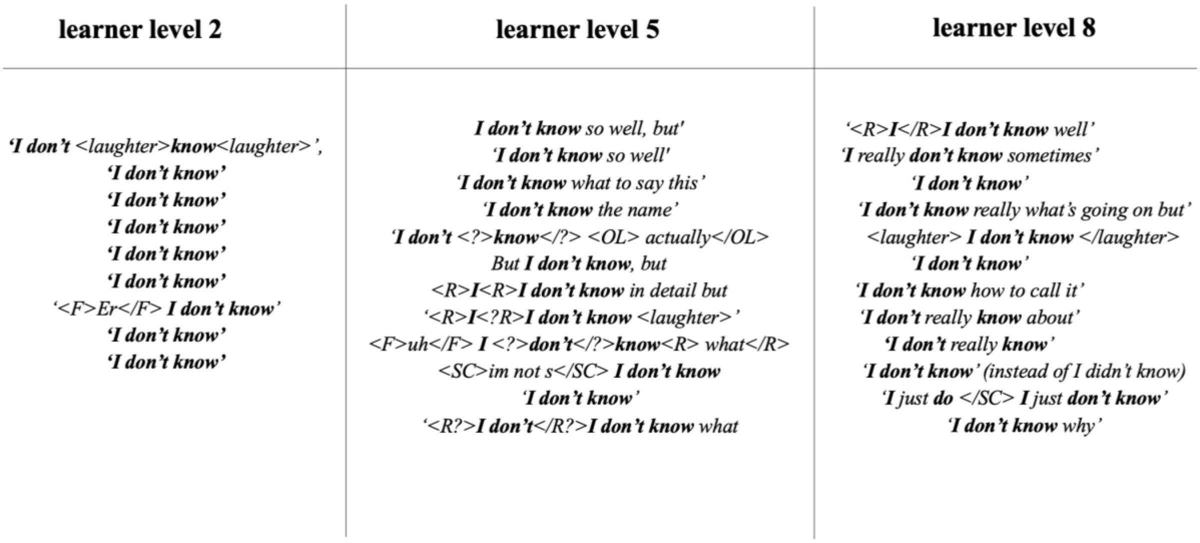
The hypothesis presented in (a) can be split into 2 parts (*a:1, a:2*). The first part is stated (rather, repeated) below:



- *(a:1) The processing unit (learner-internal formula) is produced accurately by all selected learners and across all proficiency groups under analysis (by virtue of showing phonological coherence).*

The table in (11) shows the productions of *'I don't know'* by the 15 learners under analysis.

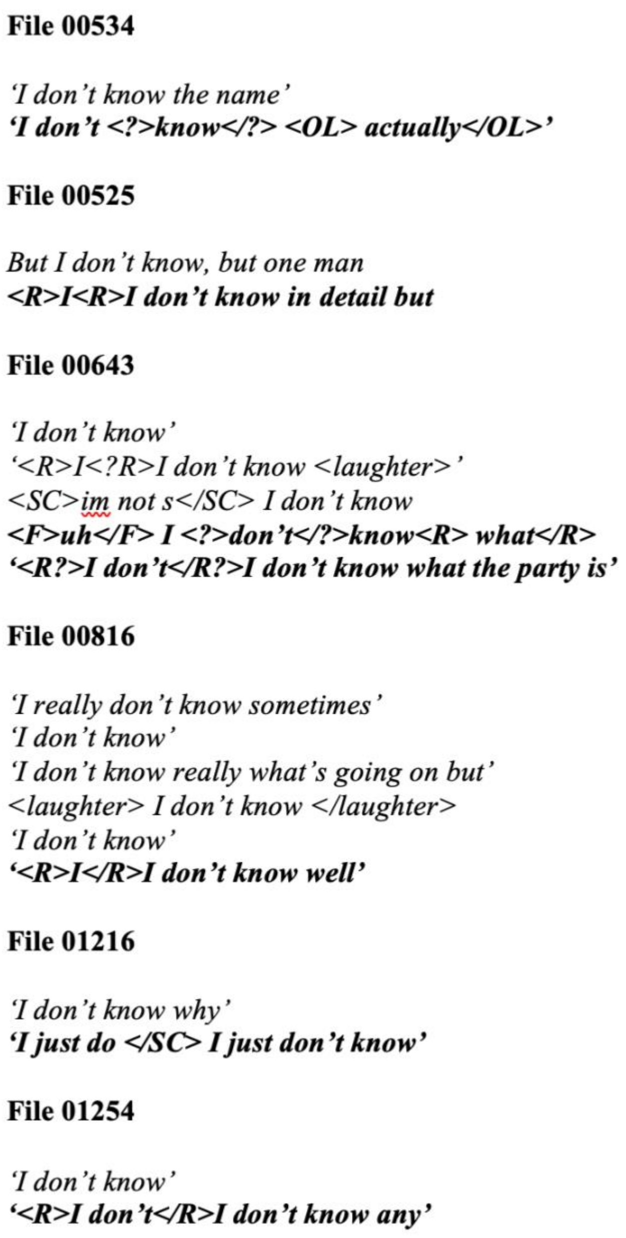
(11)



learner level 2	learner level 5	learner level 8
*‘**I don’t** <laughter>**know**<laughter>’,* ***‘I don’t know’*** ***‘I don’t know’*** ***‘I don’t know’*** ***‘I don’t know’*** ***‘I don’t know’*** *‘<F>Er</F> **I don’t know**’* ***‘I don’t know’*** ***‘I don’t know’***	***I don’t know** so well, but'* *‘**I don’t know** so well'* *‘**I don’t know** what to say this’* *‘**I don’t know** the name’* *‘**I don’t** <?>**know**</?> <OL> actually</OL>* *But **I don’t know**, but* *<R>**I**<R>**I don’t know** in detail but* *‘<R>**I**<?R>**I don’t know** <laughter>’* *<F>uh</F> **I** <?>**don’t**</?>**know**<R> what</R>* *<SC>im not s</SC> **I don’t know*** *‘**I don’t know**’* *‘<R?>**I don’t**</R?>**I don’t know** what*	*‘<R>**I**</R>**I don’t know** well’* *‘**I** really **don’t know** sometimes’* *‘**I don’t know**’* *‘**I don’t know** really what’s going on but’* *<laughter> **I don’t know** </laughter>* *‘**I don’t know**’* *‘**I don’t know** how to call it’* *‘**I don’t** really **know** about’* *‘**I don’t** really **know**’* *‘**I don’t know**’ (instead of I didn’t know)* *‘I just **do** </SC> I just **don’t know**’* *‘**I don’t know** why’*

The table shows that, in some instances*, ‘I don’t know’* is produced with internal repetition (*<R>*), slight pauses (*<?>*) and speaker correction (*<SC>*)*.* At a first glance, it would seem that this formula is not a PU for certain learners under analysis, as pauses and hesitations do not indicate holistic production (Myles & Cordier 2017). These instances, however, are produced by learners who have also pronounced the formula with phonological coherence previously, shown below in (12) through specific learner file examples (where the discontinuous instances are highlighted in bold).

(12)



File 00534

'I don't know the name'
'I don't <?>know</?> <OL> actually</OL>'

File 00525

But I don't know, but one man
<R>I<R>I don't know in detail but

File 00643

'I don't know'
'<R>I<?R>I don't know <laughter>'
<SC>im not s</SC> I don't know
<F>uh</F> I <?>don't</?>know<R> what</R>
'<R?>I don't</R?>I don't know what the party is'

File 00816

'I really don't know sometimes'
'I don't know'
'I don't know really what's going on but'
<laughter> I don't know </laughter>
'I don't know'
'<R>I</R>I don't know well'

File 01216

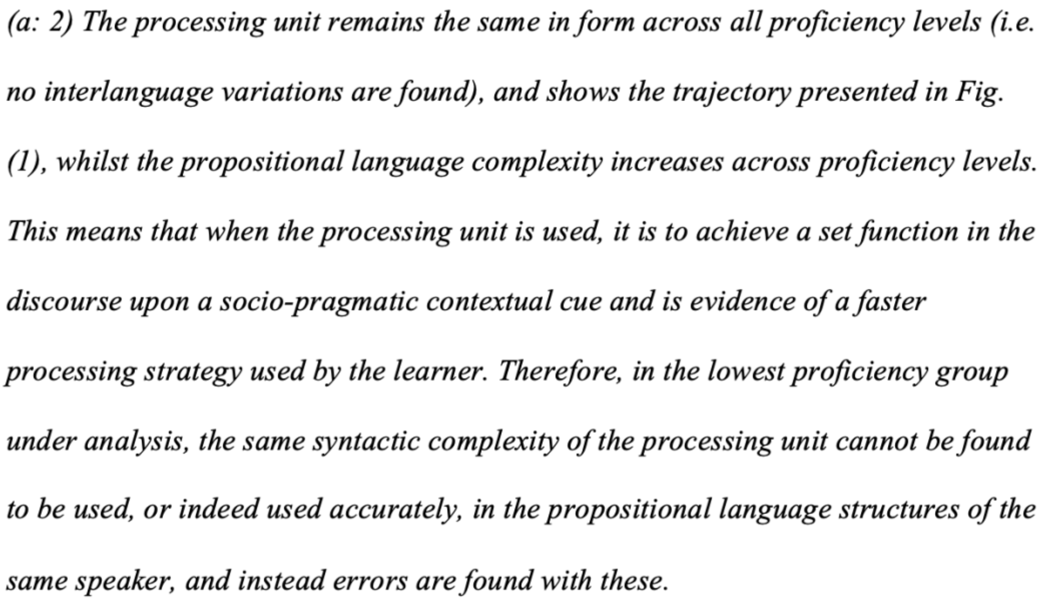
'I don't know why'
'I just do </SC> I just don't know'

File 01254

'I don't know'
'<R>I don't</R>I don't know any'

We can say then that hypothesis (a:1) is indeed fulfilled, as every learner produces the PU with phonological coherence at (at least) one point during their interview, a factor that was essentially predicted by virtue of the learner's being selected as candidates for using '*I don't know'* as a PU. The second part of hypothesis (a) is given below.

-



(a: 2) The processing unit remains the same in form across all proficiency levels (i.e. no interlanguage variations are found), and shows the trajectory presented in Fig. (1), whilst the propositional language complexity increases across proficiency levels. This means that when the processing unit is used, it is to achieve a set function in the discourse upon a socio-pragmatic contextual cue and is evidence of a faster processing strategy used by the learner. Therefore, in the lowest proficiency group under analysis, the same syntactic complexity of the processing unit cannot be found to be used, or indeed used accurately, in the propositional language structures of the same speaker, and instead errors are found with these.

The tables presented below in (13) allow us to see whether the learners' propositional language increases in complexity with proficiency level. The table shows each learner's use and accuracy of the syntactic properties associated with *'I don't know'* outside of the formula in their novel language productions. Figures are presented along with relevant descriptive statistics, in adhering to the conventions of corpus studies of a similar kind (Cheng 2011; McEnery & Hardie 2012). Accuracy rate is given in a relative percentage to the rate of usage under the (%) column, where 'inaccuracy' of a property refers to either absence of the property in structures where it is needed, or misuse of the property in these environments.

(13)

Learner Level 2

		DO-support			Cliticization			Cat NEGP			[EPP,TNS,AGG]			Generalisation		
		Use	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**
File 01127		0	0	0	1	1	100%	2	0	0%	17	5	29%	0	0	0%
File 01051		0	0	0	2	1	50%	1	0	0%	30	21	70%	0	0	0%
File 00584		0	0	0	2	1	50%	0	0	n/a	18	10	56%	0	0	0%
File 00418		0	0	0	1	1	100%	0	0	n/a	19	14	74%	0	0	0%
File 00233		3	3	100%	1	1	100%	1	1	100%	22	17	77%	1	1	100%
Average		**0.6**	**0.6**	**20%**	**1.4**	**1**	**80%**	**0.8**	**0.2**	**33%**	**21.2**	**13.4**	**61%**	**0.2**	**0.2**	**20%**

Learner Level 5

		DO-support			Cliticization			Cat NEGP			[EPP,TNS,AGG]			Generalisation		
		Use	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**
File 01028		10	6	60%	17	16	94%	8	6	75%	89	81	91%	6	4	67%
File 00247		8	7	88%	18	18	100%	10	9	90%	96	84	88%	7	7	100%
File 00534		10	10	100%	24	24	100%	7	7	100%	120	103	86%	6	6	100%
File 00525		11	10	91%	25	25	100%	13	12	92%	111	105	95%	9	8	89%
File 00643		5	5	100%	27	26	96%	15	13	87%	120	111	93%	2	1	50%
Average		**8.8**	**7.6**	**88%**	**22.2**	**21.8**	**98%**	**10.6**	**9.4**	**89%**	**107**	**96.8**	**90%**	**6**	**5.2**	**81%**

Learner Level 8

		DO-support			Cliticization			Cat NEGP			[EPP,TNS,AGG]			Generalisation		
		Use	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**
File 00816		12	12	100%	54	54	100%	32	32	100%	221	215	97%	9	9	100%
File 00933		11	11	100%	43	43	100%	24	24	100%	182	180	99%	8	8	100%
File 01253		4	3	75%	52	52	100%	13	13	100%	179	177	99%	0	0	0%
File 01216		12	12	100%	25	25	100%	17	17	100%	221	218	99%	9	8	89%
File 01254		10	10	100%	34	34	100%	20	20	100%	129	128	99%	7	7	100%
Average		**9.8**	**9.6**	**95%**	**41.6**	**41.6**	**100%**	**21.2**	**21.2**	**100%**	**186**	**184**	**99%**	**6.6**	**6.4**	**78%**

The tables in (13) show that the learners' propositional language complexity as a whole does increase as proficiency level rises, seen through the fact that both the average use and accuracy rate of the selected syntactic properties under analysis increases with the learner-levels. Tables (11) and (13) can be viewed in conjunction to show that when the learners of proficiency level 2 use the PU, the same syntactic properties are rarely found to be used, or indeed used *accurately*, in their corresponding propositional language. Apart from one learner (file 00233)[17], there is no use of DO-Support found in any novel language structure of level 2 learners, and a maximum of 2 instances per learner of cliticization and category NEGP, the majority of which are produced/appropriated inaccurately. Their use of agreement, tense and overt corresponding subject ([EPP]) is also considerably low in frequency and accuracy when compared to the higher learner levels. To better see this, tables have been created which show the individual learners' production of the PU and their accuracy of related syntactic properties together. Those of proficiency level 2 can be seen below.

[17] Note that the 2 accurate productions out of the 3 instances of DO-support for this learner was the phrase '*do you have +NP'*. We could argue perhaps that this construction is also formulaic (with the NP as an open slot), therefore meaning that it is not an accurate representation of their competence of this syntactic property. This is given further plausibility when we consider that their other instantiation of DO-support was an inaccurate negation construction- '*I don't <laughter>plan</laughter> after <?>do</?> this interview'*.

(14)

File 00233- Learner Level 2

Formula

‘I don’t know’
‘I don’t know’

DO-support			Cliticization			Cat NEGP			Features			Generalisation		
Use	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**
3	3	100%	1	1	100%	1	1	100%	22	17	77%	1	1	100%

(15)

File 01127- Learner Level 2

Formula

‘I don’t <laughter>know<laughter>’,
‘I don’t know’

DO-support			Cliticization			Cat NEGP			Features			Generalisation		
Use	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**
0	0	0	1	1	100%	2	0	0%	17	5	29%	0	0	0%

(16)

File 01051- Learner Level 2														
	Formula													
		I don’t know'												
DO-support			Cliticization			Cat NEGP			Features			Generalisation		
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%
0	0	0	2	1	50%	1	0	0%	30	21	70%	0	0	0%

(17)

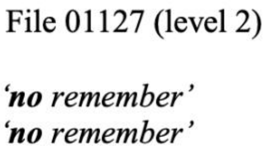
File 00584- Learner Level 2														
	Formula													
			I don’t know' ***'I don’t know'*** ***'I don’t know'***											
DO-support			Cliticization			Cat NEGP			Features			Generalisation		
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%
0	0	0	2	1	50%	0	0	0	18	10	56%	0	0	0%

(18)

File 00418- Learner Level 2														
	Formula													
		‘<F>Er</F> I don’t know’												
DO-support			**Cliticization**			**Cat NEGP**			**Features**			**Generalisation**		
Use	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**
0	0	0	1	1	100%	0	0	n/a	19	14	74%	0	0	0%

Such a lack of correlation can also be seen through the extremely low rates of generalisation in this level; tables (15) and (16) show how learners can produce ‘*I don’t know*’ coherently but fail to extend its associated syntactic properties to similar structures of negation, where instead *no* is used as a pre-verbal negation particle (i.e. NEG +V), a common trait of lower stages of both developmental sequences and acquisition order in L2 learning (eg. Cazden et al. 1975; Schumman 1979; Pienemann 1998). These attempts at negation are shown in (19) and (20).

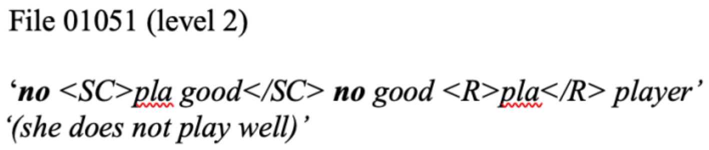
(19)



File 01127 (level 2)

*‘**no** remember’*
*‘**no** remember’*

(20)



File 01051 (level 2)

*'**no** <SC>pla good</SC> **no** good <R>pla</R> player'*
'(she does not play well)'

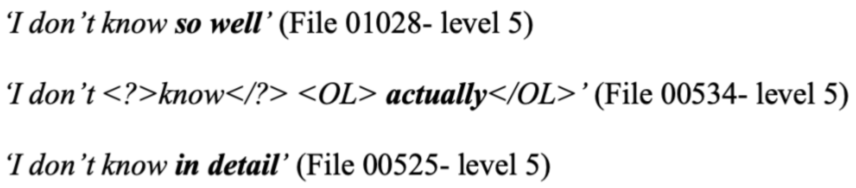
Such evidence supports studies who have branded formulaic language as accountable for the overrepresentation of linguistic knowledge in L2 learning, where learners use unanalysed multimorphemic sequences which go well beyond their grammatical competence (Myles 2004; Wray & Fitzpatrick 2008). It is therefore a strong indication that the PU for these level 2 learners is a memorized routine used as a functional filler only, which is stored more holistically whilst their propositional language develops independently. These results from learner-level 2 seem to support the second part of the hypothesis and are indicative of more general stances of formulaicity in L2 development taken by the likes of Hanania & Gradman (1977), Krashen & Scarcella (1979) and Bohn (1986) (see section II: 3).

However, this hypothesis cannot be extended to all the learners under investigation, as we see that *'I don't know'* does not just remain as a single-clause functional filler throughout the proficiency levels but is actually sensitive to both internal modification and integration into more complex clausal constructions. Table (11) shows that, as proficiency level increases, we see the formula modified by intensifiers (*I don't know* ***so well****/ I don't* ***really*** *know* etc) and used in subordination with complement clauses to form matrix constructions (*I don't know [**what** to say this]/I don't know [**what**'s going on but]'*). Although there is no evidence for 'interlanguage' variations of the formula i.e. ungrammatical/underdeveloped variations in lower proficiency levels, the use of the PU for these particular learners in the corpus does not reflect the trajectory predicted by Fig. (1), where certain acquisitional formulas are 'learned early, score high for accuracy and do not substantially change with proficiency level' (Bardovi- Harlig & Stringer 2017: 73). Instead, the PU in this analysis does change, and the variations of such that we see with these learners are actually 'advanced'/ 'complex' modifications both internally to the formula (*I [**really**] don't know'*) and externally *('I don't know [**so well**]'*). The PU therefore seems to move from being a single-clause functional filler produced in isolation to integrated into the learners' propositional language structures as proficiency level increases, supporting the original claim from Krashen & Scarcella (1979) that 'in some situations propositional language may "catch up" with automatic speech; that is, the language acquisition process may "reanalyse" patterns and routines as creative constructions' (p. 284).

Further support for this concept is that instances of speaker-pauses and self-correction are all at the intermediate and high proficiency levels (5 & 8), when it is considered that these are more likely to occur in propositional rather than automatic speech (Godlman-Eisler 1964) as they 'suggest that learners are engaged in syntactic processing' (Bardovi-Harlig & Stringer 2017: 77). That is, when a learner produces a string haltingly, it demonstrates that is has been put together online rather than 'retrieved as one unit' (Myles & Cordier 2017: 5). This infers that as *'I don't know'* becomes part of propositional language, in such structures it loses its holistic conceptualisation and by virtue any processing advantages that come with this, which could explain why some intermediate and higher-level productions are more discontinuous in nature (see Kanno 1993 for a review of such a concept).

What is interesting from this perspective though is that the productions of the PU at specific learner-levels are somewhat in conformity; it is only learner-level 2 where we see sole use of '*I don't know*' as a simple clause functional filler, and levels 5 and 8 also show homogeneity with regard to their more complex modifications. Whilst both of these higher levels show '*I don't know'* used as part of matrix clause constructions with complementisers such as *what* and *how*, the productions in level 5 show modification by intensifiers to the right adjunction of the main verb [V] only (Jackendoff 1972). Examples of such are taken from the table in (11) and repeated below.

(21)



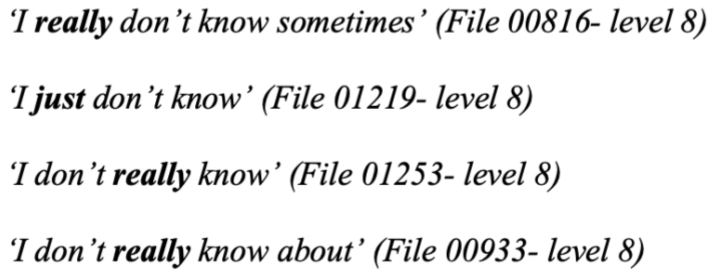
*'I don't know **so well**'* (File 01028- level 5)

*'I don't <?>know</?> <OL> **actually**</OL>'* (File 00534- level 5)

*'I don't know **in detail**'* (File 00525- level 5)

In this sense, the PU is still produced as a 'whole' at this stage of proficiency. It is only at learner-level 8 that we see instances of intensifier modification at the [I] and [VP] level through the use of sentential (S) adverbs (Jackendoff 1972; Potsdam 1998), shown respectively below.

(22)



*'I **really** don't know sometimes' (File 00816- level 8)*

*'I **just** don't know' (File 01219- level 8)*

*'I don't **really** know' (File 01253- level 8)*

*'I don't **really** know about' (File 00933- level 8)*

The modifications here are more ‘internal’ in nature, as these appear within the simple clause structure itself, where the PU is broken up intrasententially.

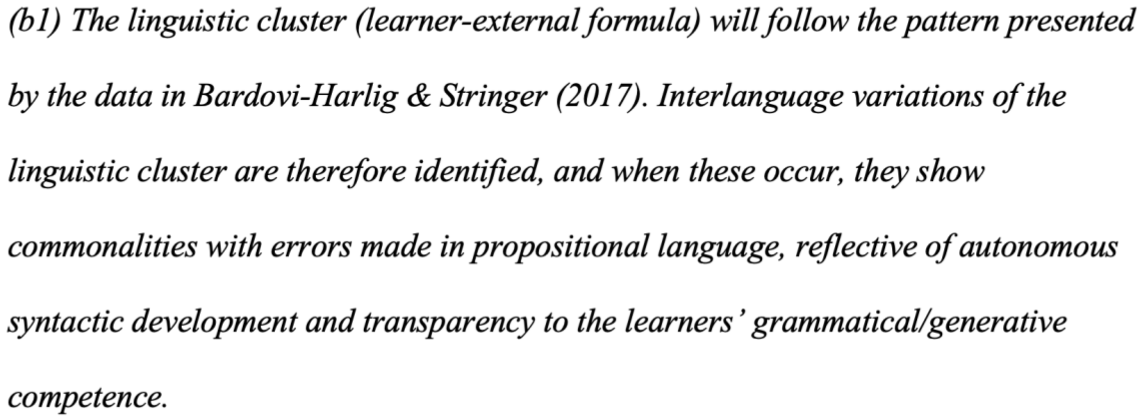
Such congruence between the nature of the production/modification of the PU and the level of proficiency could indicate that these are somewhat constrained by- and by virtue transparent to- grammatical/generative competence. Although this was not predicted of the PU, the claim that the conventional expressions under analysis in Bardovi-Harlig & Stringer’s (2017) study showed transparency to the interlanguage grammar could be extended to the present analysis of ‘*I don’t know’*, if we allow more advanced modifications to be included under the umbrella of ‘interlanguage variations’. Extending the notion of ‘interlanguage variations’ in this way could in turn explain why modification of the PU only starts to take place at learner level 5, that is, when the syntactic properties of the formula are appropriately developed/acquired[18]. This infers that as the learners’ generative competence becomes more complex, so too does the PU, as this is reparsed ‘once independent morphosyntactic development makes reanalysis possible’ (Bardovi-Harlig & Stringer 2017: 83).

[18] The tables in (13) clearly show that the associated syntactic properties are significantly more used, and indeed used more accurately, in proficiency levels 5 and 8.

2) Linguistic cluster- ‘How much is it’

The hypothesis for the results of the LC can also be split into two parts (b1: b2), the first of which is presented below.

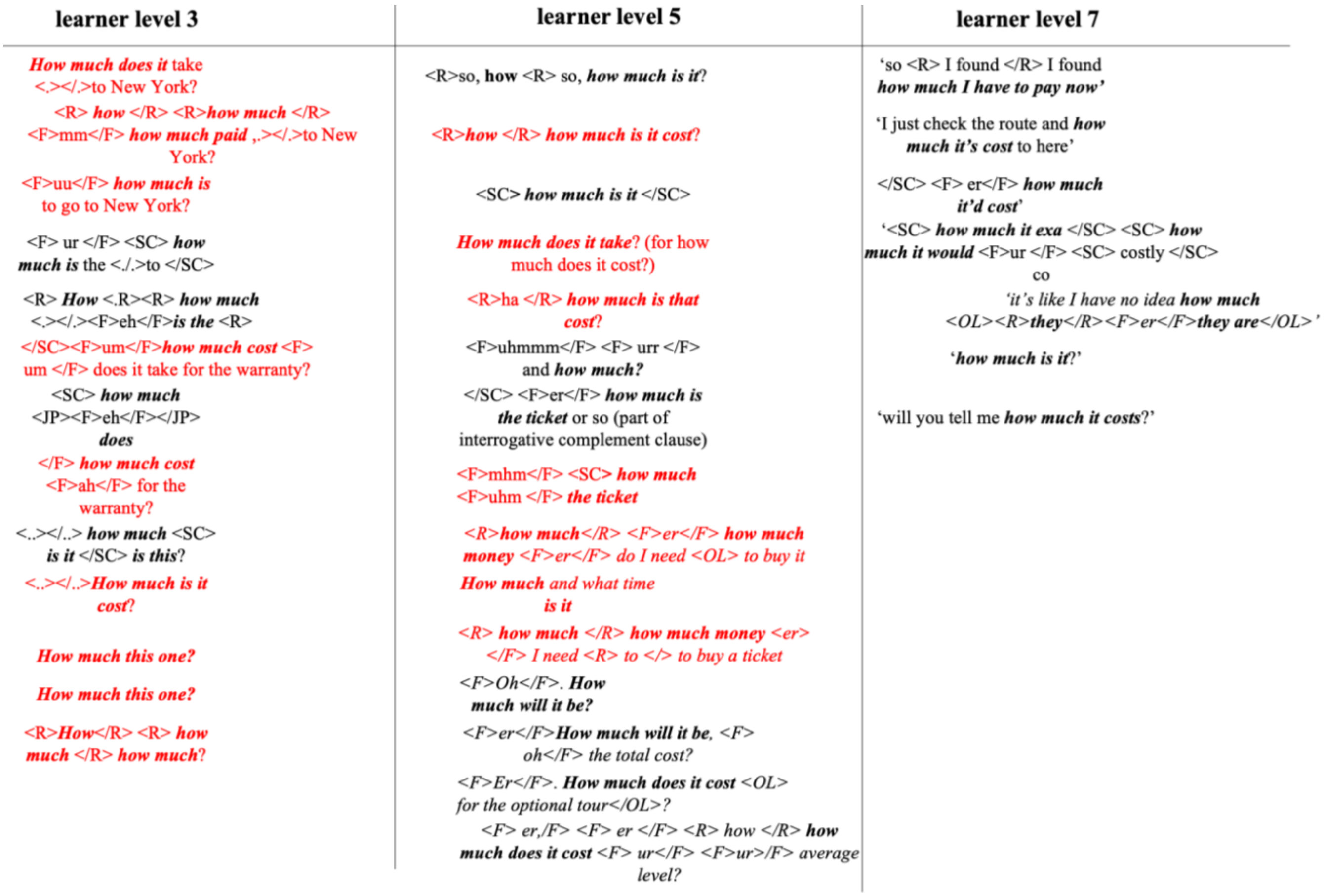
-



(b1) The linguistic cluster (learner-external formula) will follow the pattern presented by the data in Bardovi-Harlig & Stringer (2017). Interlanguage variations of the linguistic cluster are therefore identified, and when these occur, they show commonalities with errors made in propositional language, reflective of autonomous syntactic development and transparency to the learners' grammatical/generative competence.

The table in (23) shows how interlanguage variations of the LC are found, and its overall accuracy of production does indeed increase across proficiency levels (where inaccurate instantiations are highlighted in red, and accurate ones in black).

(23)

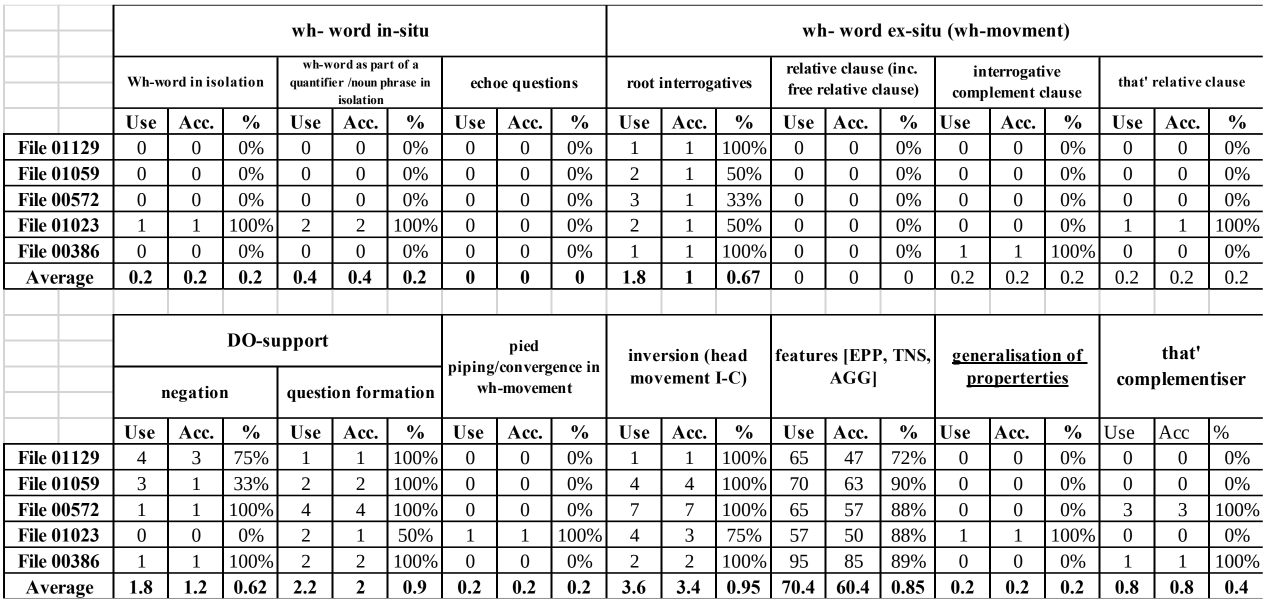


learner level 3	learner level 5	learner level 7
How much does it take <.></.>to New York?	<R>so, **how** <R> so, ***how much is it***?	‘so <R> I found </R> I found ***how much I have to pay now’***
<R> ***how*** </R> <R>***how much*** </R> <F>mm</F> ***how much paid*** ,.></.>to New York?	<R>***how*** </R> ***how much is it cost***?	‘I just check the route and ***how much it’s cost*** to here’
<F>uu</F> ***how much is*** to go to New York?	<SC> ***how much is it*** </SC>	</SC> <F> er</F> ***how much it’d cost***’
<F> ur </F> <SC> ***how much is*** the <./.>to </SC>	***How much does it take***? (for how much does it cost?)	‘<SC> ***how much it exa*** </SC> <SC> ***how much it would*** <F>ur </F> <SC> costly </SC> co
<R> ***How*** <.R><R> ***how much*** <.></.><F>eh</F>***is the*** <R>	<R>ha </R> ***how much is that cost***?	*‘it’s like I have no idea **how much** <OL><R>**they**</R><F>er</F>**they are**</OL>’*
</SC><F>um</F>***how much cost*** <F> um </F> does it take for the warranty?	<F>uhmmm</F> <F> urr </F> and ***how much?***	‘***how much is it***?’
<SC> ***how much*** <JP><F>eh</F></JP> ***does***	</SC> <F>er</F> ***how much is the ticket*** or so (part of interrogative complement clause)	‘will you tell me ***how much it costs***?’
</F> ***how much cost*** <F>ah</F> for the warranty?	<F>mhm</F> <SC> ***how much*** <F>uhm </F> ***the ticket***	
<..></..> ***how much*** <SC> ***is it*** </SC> ***is this***?	*<R>**how much**</R> <F>er</F> **how much money** <F>er</F> do I need <OL> to buy it*	
<..></..>***How much is it cost***?	***How much*** *and what time* ***is it***	
How much this one?	*<R> **how much** </R> **how much money** <er> </F> I need <R> to </> to buy a ticket*	
How much this one?	*<F>Oh</F>.* ***How much will it be?***	
<R>***How***</R> <R> ***how much*** </R> ***how much***?	*<F>er</F>**How much will it be**, <F> oh</F> the total cost?*	
	*<F>Er</F>. **How much does it cost** <OL> for the optional tour</OL>?*	
	*<F> er,/F> <F> er </F> <R> how </R> **how much does it cost** <F> ur</F> <F>ur>/F> average level?*	

As with *‘I don’t know’*, table (23) above can be used in conjunction with those in (24) below, which present each learners use and accuracy of related syntactic properties in their novel language constructions (note that in these tables, unrelated properties to *‘how much is it’* are also shown i.e. *DO-support*, as these are required/referred to in later stages of the analysis).

(24)

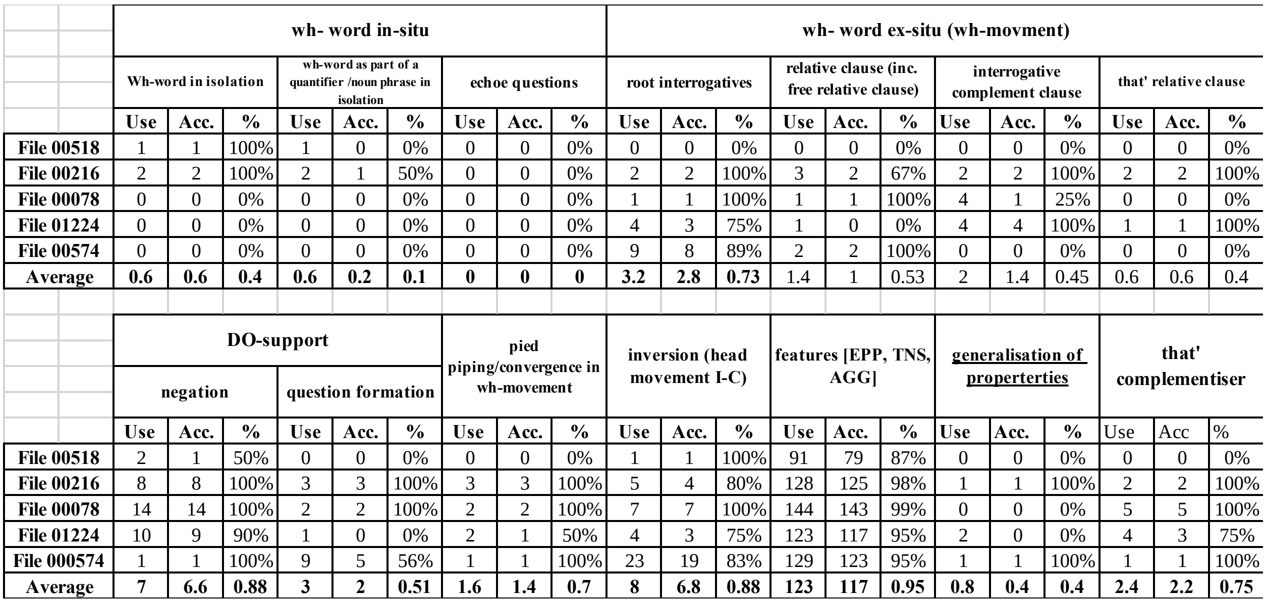
Learner Level 3



	wh- word in-situ									wh- word ex-situ (wh-movment)											
	Wh-word in isolation			wh-word as part of a quantifier /noun phrase in isolation			echoe questions			root interrogatives			relative clause (inc. free relative clause)			interrogative complement clause			that' relative clause		
	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%
File 01129	0	0	0%	0	0	0%	0	0	0%	1	1	100%	0	0	0%	0	0	0%	0	0	0%
File 01059	0	0	0%	0	0	0%	0	0	0%	2	1	50%	0	0	0%	0	0	0%	0	0	0%
File 00572	0	0	0%	0	0	0%	0	0	0%	3	1	33%	0	0	0%	0	0	0%	0	0	0%
File 01023	1	1	100%	2	2	100%	0	0	0%	2	1	50%	0	0	0%	0	0	0%	1	1	100%
File 00386	0	0	0%	0	0	0%	0	0	0%	1	1	100%	0	0	0%	1	1	100%	0	0	0%
Average	**0.2**	**0.2**	**0.2**	**0.4**	**0.4**	**0.2**	**0**	**0**	**0**	**1.8**	**1**	**0.67**	0	0	0	0.2	0.2	0.2	0.2	0.2	0.2

	DO-support						pied piping/convergence in wh-movement			inversion (head movement I-C)			features [EPP, TNS, AGG]			generalisation of properterties			that' complementiser		
	negation			question formation																	
	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc	%
File 01129	4	3	75%	1	1	100%	0	0	0%	1	1	100%	65	47	72%	0	0	0%	0	0	0%
File 01059	3	1	33%	2	2	100%	0	0	0%	4	4	100%	70	63	90%	0	0	0%	0	0	0%
File 00572	1	1	100%	4	4	100%	0	0	0%	7	7	100%	65	57	88%	0	0	0%	3	3	100%
File 01023	0	0	0%	2	1	50%	1	1	100%	4	3	75%	57	50	88%	1	1	100%	0	0	0%
File 00386	1	1	100%	2	2	100%	0	0	0%	2	2	100%	95	85	89%	0	0	0%	1	1	100%
Average	**1.8**	**1.2**	**0.62**	**2.2**	**2**	**0.9**	**0.2**	**0.2**	**0.2**	**3.6**	**3.4**	**0.95**	**70.4**	**60.4**	**0.85**	**0.2**	**0.2**	**0.2**	**0.8**	**0.8**	**0.4**

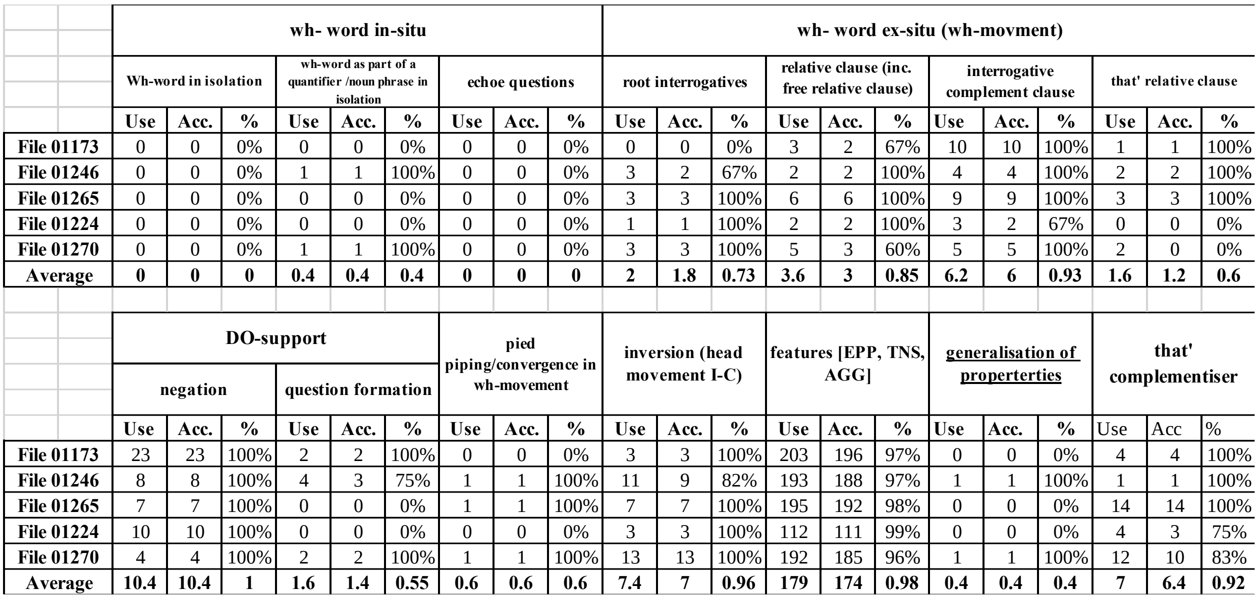
Learner Level 5



	wh- word in-situ									wh- word ex-situ (wh-movment)											
	Wh-word in isolation			wh-word as part of a quantifier /noun phrase in isolation			echoe questions			root interrogatives			relative clause (inc. free relative clause)			interrogative complement clause			that' relative clause		
	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%
File 00518	1	1	100%	1	0	0%	0	0	0%	0	0	0%	0	0	0%	0	0	0%	0	0	0%
File 00216	2	2	100%	2	1	50%	0	0	0%	2	2	100%	3	2	67%	2	2	100%	2	2	100%
File 00078	0	0	0%	0	0	0%	0	0	0%	1	1	100%	1	1	100%	4	1	25%	0	0	0%
File 01224	0	0	0%	0	0	0%	0	0	0%	4	3	75%	1	0	0%	4	4	100%	1	1	100%
File 00574	0	0	0%	0	0	0%	0	0	0%	9	8	89%	2	2	100%	0	0	0%	0	0	0%
Average	**0.6**	**0.6**	**0.4**	**0.6**	**0.2**	**0.1**	**0**	**0**	**0**	**3.2**	**2.8**	**0.73**	1.4	1	0.53	2	1.4	0.45	0.6	0.6	0.4

	DO-support						pied piping/convergence in wh-movement			inversion (head movement I-C)			features [EPP, TNS, AGG]			generalisation of properterties			that' complementiser		
	negation			question formation																	
	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc	%
File 00518	2	1	50%	0	0	0%	0	0	0%	1	1	100%	91	79	87%	0	0	0%	0	0	0%
File 00216	8	8	100%	3	3	100%	3	3	100%	5	4	80%	128	125	98%	1	1	100%	2	2	100%
File 00078	14	14	100%	2	2	100%	2	2	100%	7	7	100%	144	143	99%	0	0	0%	5	5	100%
File 01224	10	9	90%	1	0	0%	2	1	50%	4	3	75%	123	117	95%	2	0	0%	4	3	75%
File 000574	1	1	100%	9	5	56%	1	1	100%	23	19	83%	129	123	95%	1	1	100%	1	1	100%
Average	**7**	**6.6**	**0.88**	**3**	**2**	**0.51**	**1.6**	**1.4**	**0.7**	**8**	**6.8**	**0.88**	**123**	**117**	**0.95**	**0.8**	**0.4**	**0.4**	**2.4**	**2.2**	**0.75**

Learner Level 7

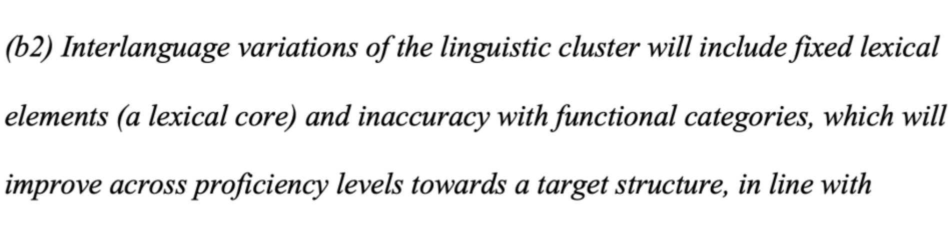


	wh- word in-situ									wh- word ex-situ (wh-movment)											
	Wh-word in isolation			wh-word as part of a quantifier /noun phrase in isolation			echoe questions			root interrogatives			relative clause (inc. free relative clause)			interrogative complement clause			that' relative clause		
	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%
File 01173	0	0	0%	0	0	0%	0	0	0%	0	0	0%	3	2	67%	10	10	100%	1	1	100%
File 01246	0	0	0%	1	1	100%	0	0	0%	3	2	67%	2	2	100%	4	4	100%	2	2	100%
File 01265	0	0	0%	0	0	0%	0	0	0%	3	3	100%	6	6	100%	9	9	100%	3	3	100%
File 01224	0	0	0%	0	0	0%	0	0	0%	1	1	100%	2	2	100%	3	2	67%	0	0	0%
File 01270	0	0	0%	1	1	100%	0	0	0%	3	3	100%	5	3	60%	5	5	100%	2	0	0%
Average	**0**	**0**	**0**	**0.4**	**0.4**	**0.4**	**0**	**0**	**0**	**2**	**1.8**	**0.73**	**3.6**	**3**	**0.85**	**6.2**	**6**	**0.93**	**1.6**	**1.2**	**0.6**

	DO-support						pied piping/convergence in wh-movement			inversion (head movement I-C)			features [EPP, TNS, AGG]			generalisation of properterties			that' complementiser		
	negation			question formation																	
	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc	%
File 01173	23	23	100%	2	2	100%	0	0	0%	3	3	100%	203	196	97%	0	0	0%	4	4	100%
File 01246	8	8	100%	4	3	75%	1	1	100%	11	9	82%	193	188	97%	1	1	100%	1	1	100%
File 01265	7	7	100%	0	0	0%	1	1	100%	7	7	100%	195	192	98%	0	0	0%	14	14	100%
File 01224	10	10	100%	0	0	0%	0	0	0%	3	3	100%	112	111	99%	0	0	0%	4	3	75%
File 01270	4	4	100%	2	2	100%	1	1	100%	13	13	100%	192	185	96%	1	1	100%	12	10	83%
Average	**10.4**	**10.4**	**1**	**1.6**	**1.4**	**0.55**	**0.6**	**0.6**	**0.6**	**7.4**	**7**	**0.96**	**179**	**174**	**0.98**	**0.4**	**0.4**	**0.4**	**7**	**6.4**	**0.92**

The tables in (24) show how the development of key related syntactic properties (i.e. *wh-movement* and *inversion*) are reflective of the development of the LC itself, in that these also increase in overall accuracy and use as proficiency level rises, suggesting that the development of *'how much is it'* is indeed reflective and transparent to syntactic competence. The second part of the hypothesis concerns the nature of the interlanguage variations and is repeated below.

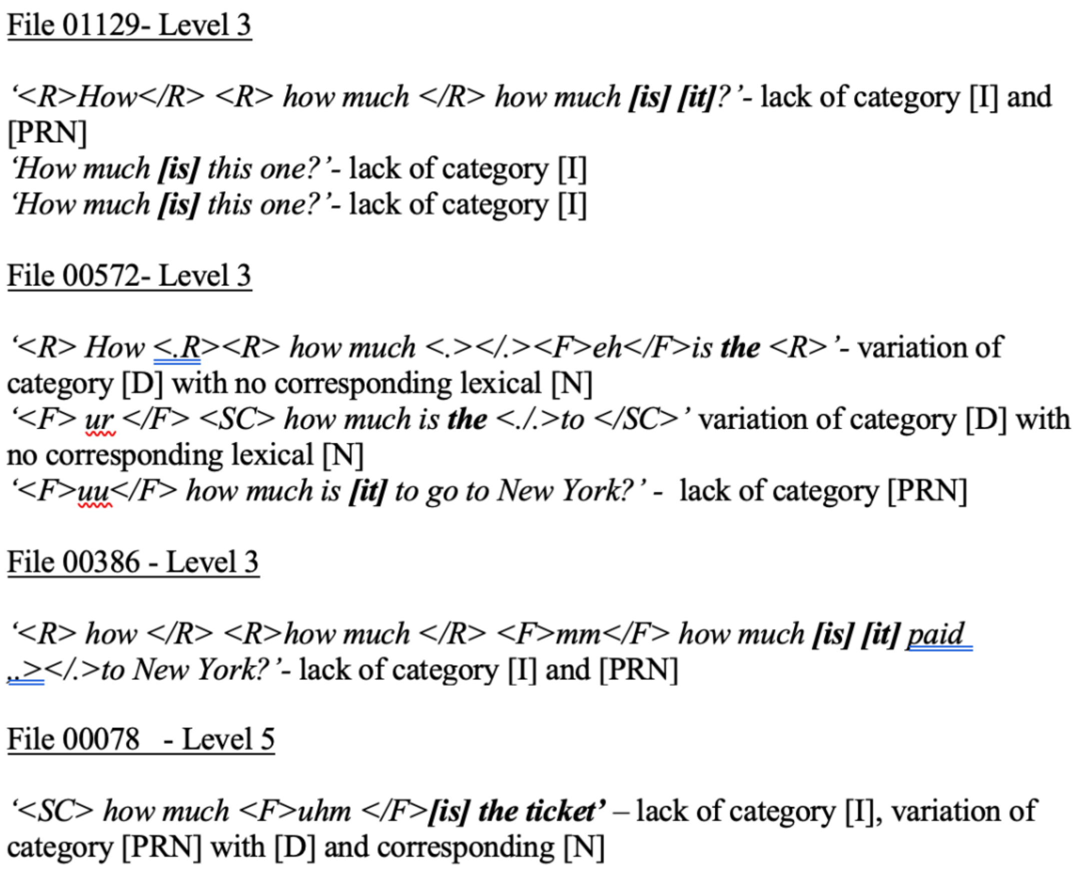
-



(b2) Interlanguage variations of the linguistic cluster will include fixed lexical elements (a lexical core) and inaccuracy with functional categories, which will improve across proficiency levels towards a target structure, in line with propositional language development following the trajectory in Fig. (2).

Traditional grammar and indeed the Minimalist Programme assume the dichotomy of lexical and functional categories (Radford 2004a), where the former refers to categories containing words which have substantive lexical-semantic content and can therefore denote specific objects or ideas, and the latter to categories of words which essentially serve to mark grammatical properties (Radford 2009). Lexical categories include noun (N), verb (V), adjective (ADJ), adverb (ADV) and preposition (PP) (Jackendoff 2002: 153), whilst functional categories include determiner (D), quantifier (Q), pronoun (PRN), auxiliary verb (I) and complementiser (C) (Radford 2009: 4-7). If we look at the table in (23), it is clear that the majority of interlanguage variations of the linguistic cluster involve accurate, fixed lexical categories (Adv *how* and by virtue Q *much*) and variation/inaccuracy of the functional ones, replicating the pattern seen from the development of conventional expressions in Bardovi-Harlig & Stringer (2017) and thus predicted in the present study by hypothesis (b2). Some inaccurate interlanguage productions are repeated below in (25), where the inaccuracy involves either variation at the functional categories (highlighted in bold), or a lack of these where structurally required (highlighted in bold and represented though square brackets).

(25)



File 01129- Level 3

‘<R>How</R> <R> how much </R> how much ***[is] [it]****? ’*- lack of category [I] and [PRN]
‘How much ***[is]*** *this one? ’*- lack of category [I]
‘How much ***[is]*** *this one? ’*- lack of category [I]

File 00572- Level 3

‘<R> How <.R><R> how much <.></.><F>eh</F>is ***the*** *<R> ’*- variation of category [D] with no corresponding lexical [N]
‘<F> ur </F> <SC> how much is ***the*** *<./.>to </SC>’* variation of category [D] with no corresponding lexical [N]
‘<F>uu</F> how much is ***[it]*** *to go to New York? ’* - lack of category [PRN]

File 00386 - Level 3

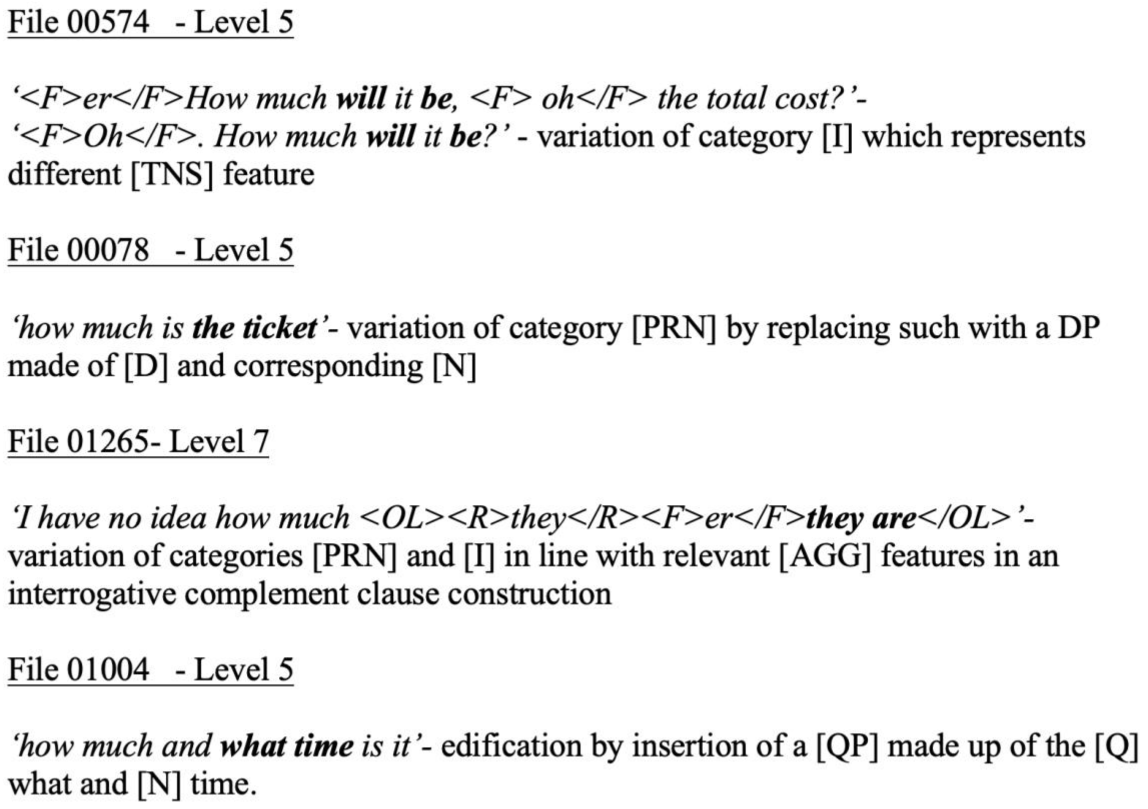
‘<R> how </R> <R>how much </R> <F>mm</F> how much ***[is] [it]*** *paid ..></.>to New York? ’*- lack of category [I] and [PRN]

File 00078 - Level 5

*‘<SC> how much <F>uhm </F>****[is] the ticket’*** – lack of category [I], variation of category [PRN] with [D] and corresponding [N]

As with the PU, as proficiency level increases, *‘how much is it’* shows advanced/complex modifications which also predominantly involve internal changes at the functional categories as well as integration of the LC into more complex clausal constructions. These are represented below in (26), where the functional variations are in bold.

(26)



File 00574 - Level 5

'<F>er</F>How much ***will*** *it* ***be****, <F> oh</F> the total cost?'-*
'<F>Oh</F>. How much ***will*** *it* ***be****?'* - variation of category [I] which represents different [TNS] feature

File 00078 - Level 5

'how much is ***the ticket****'*- variation of category [PRN] by replacing such with a DP made of [D] and corresponding [N]

File 01265- Level 7

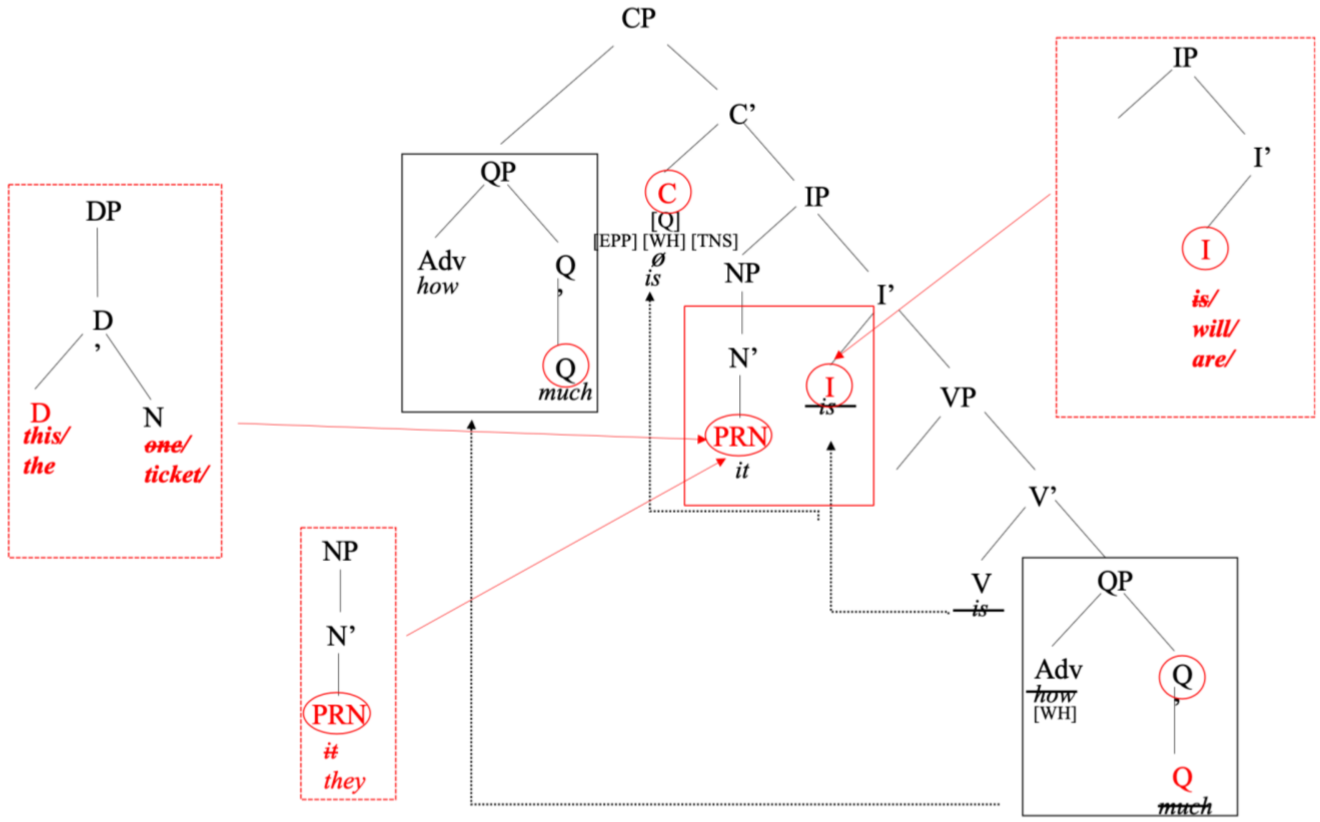
*'I have no idea how much <OL><R>they</R><F>er</F>****they are****</OL>'*- variation of categories [PRN] and [I] in line with relevant [AGG] features in an interrogative complement clause construction

File 01004 - Level 5

'how much and ***what time*** *is it'*- edification by insertion of a [QP] made up of the [Q] what and [N] time.

This congruence of both inaccurate (25) and complex (26) interlanguage variations involving functional categories can be represented structurally on the tree-diagram originally presented in IV: 3.2 (10), given below in (27).

(27)

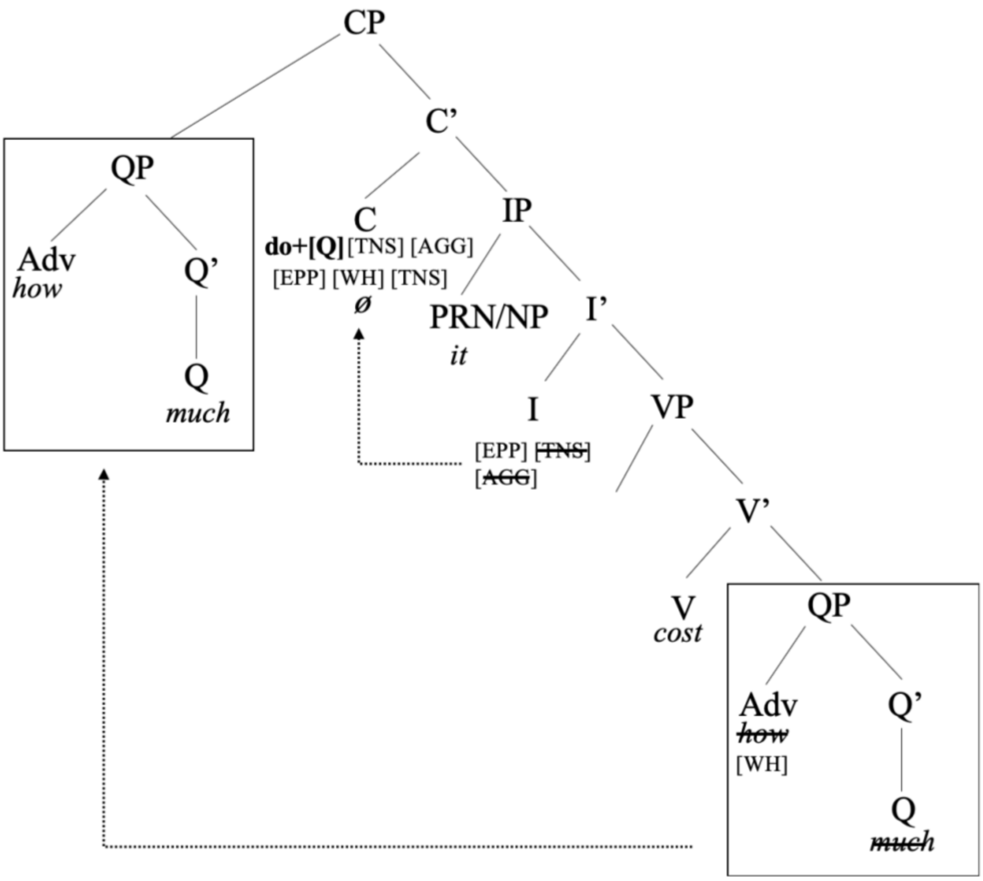


We can say then that the analysis of the LC in this investigation corresponds fairly directly with that of the conventional expressions in Bardovi-Harlig & Stringer (2017), and supports the hypothesis presented in (b).

The hypothesis is not completely realised though, as we see from the table in (23), '*how much is it'* does not seem to be the 'target expression' for all learners fulfilling the contextual (task) function of ‘enquiring about the price’. Rather, in proficiency levels 5 and 7, the majority of the learners seem to be aiming for the expression ‘*how much does it cost’*, the syntactic structure of which can be seen below in (29). Note that this expression requires DO-support, as ‘affix hopping’[19] cannot apply since the complement of the C constituent which contains the tense affix is not an appropriate host (i.e. it is not a VP headed by an overt verb but instead an IP headed by a null I) (Radford 2009).

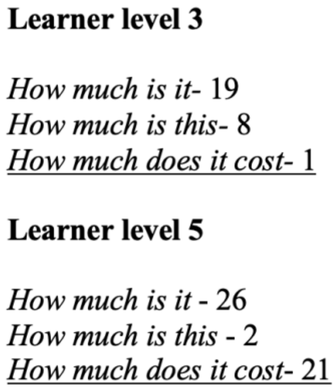
[19] ‘Affix hopping’ is a phenomenon whereby a constituent contains an unattached affix (Af), which in the PF component is lowered onto the head of its complement (provided the head is an appropriate host for the affix to attach to) (Radford 2009: 104).

(28)



Interestingly, the preference for the target *'how much does it cost'* correlates with the development of DO-support in propositional language; we see in table (24) that DO-support occurs an average of 4 times in the learners of level 3 proficiency compared to an average of 10 in level 5 and 12 in level 7. This idea is supported when we examine accurate productions of variations of the LC in learner-levels 3 and 5 across the whole corpus, which are stated below.

(29)



Learner level 3

How much is it- 19
How much is this- 8
How much does it cost- 1

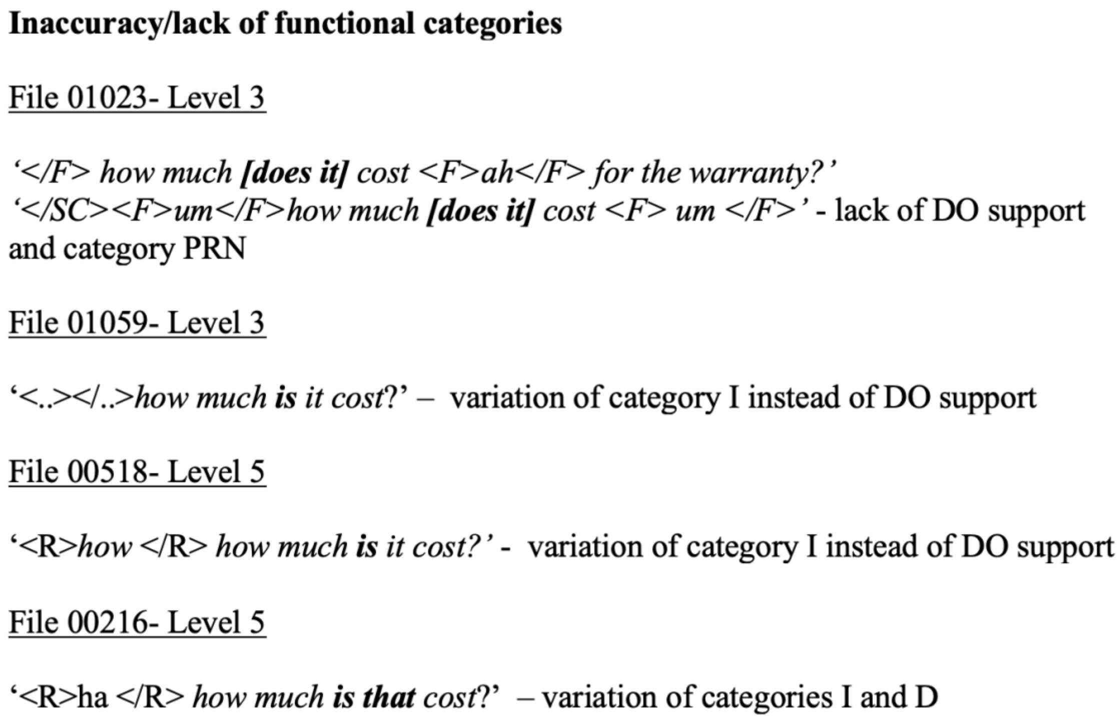
Learner level 5

How much is it - 26
How much is this - 2
How much does it cost- 21

The figures in (29) show that *'how much does it cost'* makes up only 4% of the accurate production variations in level 3, compared with 42% in level 5, as DO-support begins to show more use and accuracy in propositional language. As with those who conceptualised *'how much is it'* as the target expression, *'how much does it cost'* follows a similar pattern of development, where modifications at the functional categories and fixed lexical elements[20] derive interlanguage variations corresponding with proficiency level. Some of these are shown respectively in (30) and (31) below.

[20] The fixed lexical elements referred to here are the ADV *how* (and by virtue the Q *much*), and the V *cost*.

(30)



Inaccuracy/lack of functional categories

<u>File 01023- Level 3</u>

*‘</F> how much **[does it]** cost <F>ah</F> for the warranty?’*
*‘</SC><F>um</F>how much **[does it]** cost <F> um </F>’* - lack of DO support and category PRN

<u>File 01059- Level 3</u>

‘<..></..>*how much **is** it cost*?’ – variation of category I instead of DO support

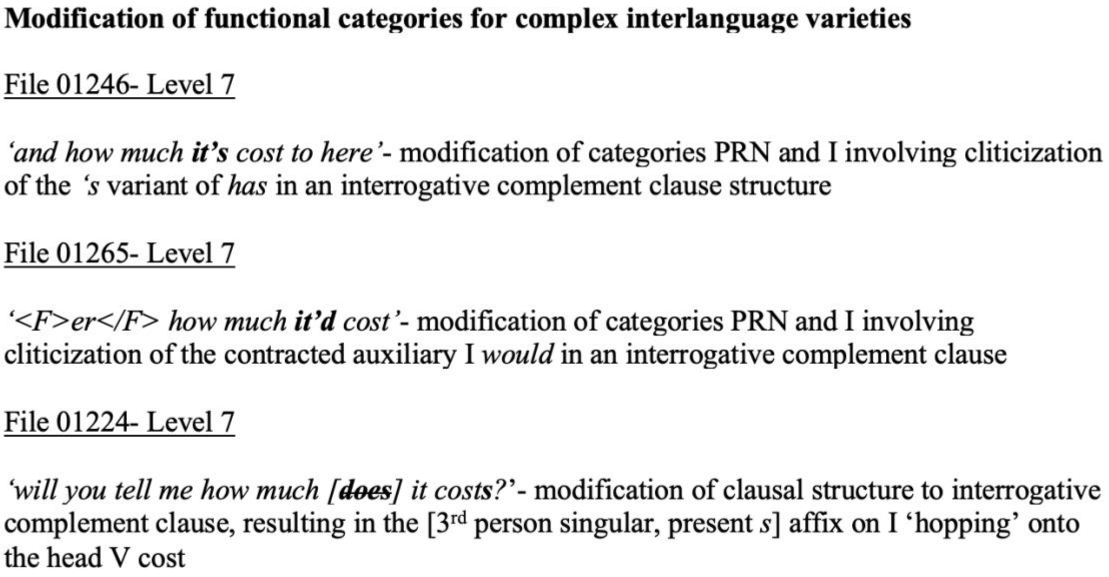
<u>File 00518- Level 5</u>

‘<R>*how* </R> *how much **is** it cost?’* - variation of category I instead of DO support

<u>File 00216- Level 5</u>

‘<R>ha </R> *how much **is that** cost*?’ – variation of categories I and D

(31)



Modification of functional categories for complex interlanguage varieties

File 01246- Level 7

*'and how much **it's** cost to here'*- modification of categories PRN and I involving cliticization of the *'s* variant of *has* in an interrogative complement clause structure

File 01265- Level 7

*'<F>er</F> how much **it'd** cost'*- modification of categories PRN and I involving cliticization of the contracted auxiliary I *would* in an interrogative complement clause

File 01224- Level 7

*'will you tell me how much [**~~does~~**] it cost**s**?'*- modification of clausal structure to interrogative complement clause, resulting in the [3rd person singular, present *s*] affix on I 'hopping' onto the head V cost

Results from the present experiment seem to show that no matter what functional LC the learners are aiming for (i.e. *how much is it*, *how much does it cost*), the development and structural nature of their interlanguage variations corresponds to those of the conventional expressions under analysis in Bardovi-Harlig & Stringer (2017). Whilst the authors note in their study that interlanguage forms of FL have lacked principled focus of investigation in the literature[21], the arrested development of functional categories compared to lexical ones is a recognised and somewhat consensual concept in generative SLA and L2 learning research. For example, Vainikka & Young- Scholten (1996) and Hawkins (2001) view the L2 development process as a hierarchical one; the former posing the 'Minimal Trees' analysis

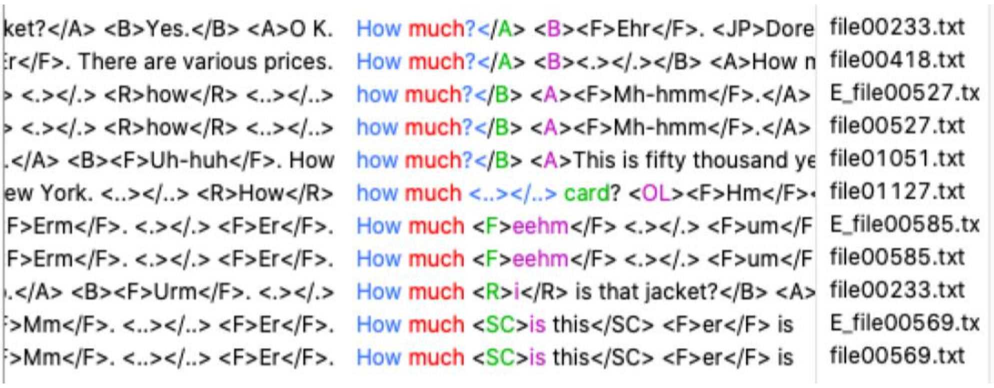
[21] Bardovi- Harlig & Stringer (2017) note that there are exceptions; Osborne (2008) reports instances of pluralized adjectives in FL produced by advanced learners, and the pragmatics literature has presented examples of intensifier omission/variation (eg. Foster 2001).

and the latter the ‘Shallow Structure Hypothesis’, both of which fundamentally state how lexical phrases (i.e. FL) start out as syntactically underspecified and involve accurate lexical categories only, before syntactic structure is gradually built as the functional categories begin to appear. Clahsen & Felser (2006) also follow from such and relate this to the mental representations of second language learners, postulating the ‘Shallow Structure Hypothesis’ which infers that L2 learners interpret strings of words (i.e. sentences) in a minimal semantic representation without mapping detailed, functional syntactic representations onto these. Such concepts are the foundations for models of SLA that place the principle difficulty of L2 learning to lie at the development of functional categories and their respective realisation (see for example Lardiere 1998b; White 2003; Slabakova 2008).

What is interesting is that all learners under investigation in the present study, and that of Bardovi-Harlig & Stringer (2017), were able to identify a ‘target’ expression that was appropriate for the discourse/pragmatic context, but were unable to realise this accurately, in other words, lexical representations allowed for ‘appropriate usage in advance of target like morphosyntactic knowledge’ (p. 81). Such a concept has even been recognised in research which has identified with usage-based proposals of acquisition; Hakuta (1972) for example stated how learners can have no internal structural knowledge of speech segments but do have the knowledge as to which particular situations call for what patterns. Bardovi-Harlig & Stringer (2017) use Jackendoff’s model of ‘parallel architecture’ to account for such a phenomenon, which takes the lexical item as a relation between different types of mental representation; minimally, a phonological structure (PS) a syntactic structure (SS) and a conceptual structure (CS) (2002: 165-77). They state that it is the redundancy in mapping of the SS and CS representations not just at the phrasal level but also at the whole, which allows learners to produce an appropriate, or ’target’, lexical core which is fixed, with functional slots open to various degrees (Bardovi- Harlig & Stringer 2017). Such a model can indeed be applied to the data in the present study and could perhaps be extended to include the advanced variations used by the learners, if we assume that the functional slots are the ones which remain 'open' for modification at latter stages of development.

As well as the development of the formulas themselves, what is perceived by the learners as a 'target' LC seems to be somewhat constrained by their independent syntactic competence, as we see from the propensity to aim for *'how much does it cost'* as proficiency level- and consequently use and accuracy of DO-support- increases. It could also be possible to extend Jackendoff's divisionary model of the mental lexicon outlined above in an attempt to explain such an observation. For example, in a similar fashion to Vainikka & Young-Scholten (1996) and Hawkins (2001), Myles (2004) has suggested that in L2 learning, semantic representations (what we can link to Jackendoff's CS level) are initially mapped onto phonological strings (PS) without assigning syntactic structure (SS). She uses the juxtaposition of the CS mappings [ask name] and [boy] to explain the overextension of [*comment t'appelles-tu] (what is your name)* by an L2 French learner in their production of the phrase *[[*comment t'appelles-tu] [le garcon]] (*what is your name the boy i.e. what is the boy's name*) (p. 155). If we can assume the semantic representation (CS) of *'how much is it/does it cost'* as [ask price], the results from these learners suggest that, although syntactic structure (SS) may not be assigned, or rather linked, *successfully* (hence underdeveloped interlanguage variations), the learners' syntactic/generative competence may somewhat constrain the choice of phonological string (PS) which can be primed by the conceptual structure (CS) mapping. This perhaps explains why, for the LC, we see no attempt of *'how much does it cost'* where there is a lack of 'DO- support' in propositional language. Such a concept is further supported when we look at all attempts of phonological mapping onto the CS [ask price] in instances of learners from proficiency level 2, where, as we have seen from section 5.1 of this chapter, DO-support is virtually non-existent. The KWIC view screenshot below in (32) shows no attempts at *'how much does it cost'*, with the majority of productions instead being *'how much'* used in isolation, or inaccurate/discontinuous attempts at the target *'how much is it/this'*.

(32) **All instances of learners 'asking the price' in transcripts from proficiency level 2**



```
ket?</A> <B>Yes.</B> <A>O K.      How much?</A> <B><F>Ehr</F>. <JP>Dore   file00233.txt
r</F>. There are various prices.  How much?</A> <B><.></.></B> <A>How n  file00418.txt
> <.></.> <R>how</R> <..></..>    how much?</B> <A><F>Mh-hmm</F>.</A>    E_file00527.tx
> <.></.> <R>how</R> <..></..>    how much?</B> <A><F>Mh-hmm</F>.</A>    file00527.txt
.</A> <B><F>Uh-huh</F>. How       how much?</B> <A>This is fifty thousand ye  file01051.txt
ew York. <..></..> <R>How</R>     how much <..></..> card? <OL><F>Hm</F>  file01127.txt
F>Erm</F>. <.></.> <F>Er</F>.     How much <F>eehm</F> <.></.> <F>um</F  E_file00585.tx
F>Erm</F>. <.></.> <F>Er</F>.     How much <F>eehm</F> <.></.> <F>um</F  file00585.txt
.</A> <B><F>Urm</F>. <.></.>      How much <R>i</R> is that jacket?</B> <A>  file00233.txt
>Mm</F>. <..></..> <F>Er</F>.     How much <SC>is this</SC> <F>er</F> is  E_file00569.tx
>Mm</F>. <..></..> <F>Er</F>.     How much <SC>is this</SC> <F>er</F> is  file00569.txt
```

Whilst Myles & Cordier (2017) place emphasis on the individualistic nature of PU's, it is possible that this could also be extended to LC's; that is, a learner's individual generative competence may be what ultimately determines the syntactic structures (SS) available in the CS/PS mapping, and consequently limits how learners automatize the formulaicity of the language around them.

Transparency of the LC to the learner's interlanguage grammar is not just reflected in correspondence between its production and the use and accuracy of syntactic properties as a whole, but also in how both of these are realised across various clausal constructions. Table (23) shows that, in proficiency levels 3 and 5, the LC's are realised predominantly in their 'target' like, root interrogative form. We can see how this corresponds to wh-movement in their propositional language, which manifests predominantly in root interrogative constructions, repeated and highlighted below in (33).

(33)

Learner Level 3

wh- word ex-situ (wh-movment)											
root interrogatives			**relative clause (inc. free relative clause)**			**interrogative complement clause**			**that' relative clause**		
Use	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**
1	1	100%	0	0	0%	0	0	0%	0	0	0%
2	1	50%	0	0	0%	0	0	0%	0	0	0%
3	1	33%	0	0	0%	0	0	0%	0	0	0%
2	1	50%	0	0	0%	0	0	0%	1	1	100%
1	1	100%	0	0	0%	1	1	100%	0	0	0%
1.8	**1**	**0.67**	0	0	0	0.2	0.2	0.2	0.2	0.2	0.2

Learner Level 5

wh- word ex-situ (wh-movment)											
root interrogatives			relative clause (inc. free relative clause)			interrogative complement clause			that' relative clause		
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%
0	0	0%	0	0	0%	0	0	0%	0	0	0%
2	2	100%	3	2	67%	2	2	100%	2	2	100%
1	1	100%	1	1	100%	4	1	25%	0	0	0%
4	3	75%	1	0	0%	4	4	100%	1	1	100%
9	8	89%	2	2	100%	0	0	0%	0	0	0%
3.2	**2.8**	**0.73**	1.4	1	0.53	2	1.4	0.45	0.6	0.6	0.4

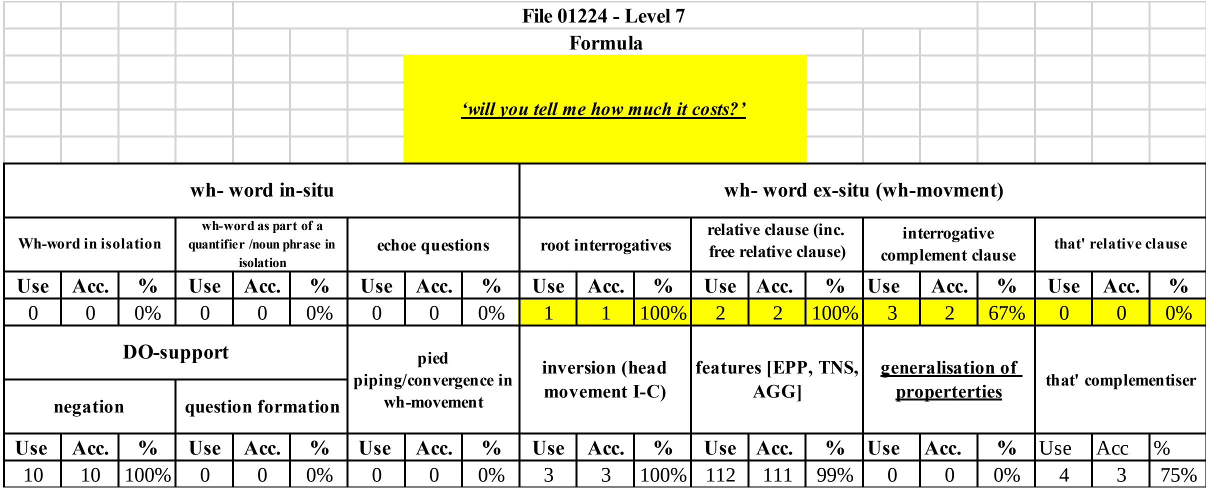
As we reach learner-level 7, however, table (34) shows the majority of learner productions involve the integration of the LC's into relative and interrogative complement clauses, which mirrors the manifestation of wh-movement in their propositional language structures. This table can be compared below along with instances of individual learner examples.

(34)

Learner Level 7

wh- word ex-situ (wh-movment)											
root interrogatives			relative clause (inc. free relative clause)			interrogative complement clause			that' relative clause		
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%
0	0	0%	3	2	67%	10	10	100%	1	1	100%
3	2	67%	2	2	100%	4	4	100%	2	2	100%
3	3	100%	6	6	100%	9	9	100%	3	3	100%
1	1	100%	2	2	100%	3	2	67%	0	0	0%
3	3	100%	5	3	60%	5	5	100%	2	0	0%
2	**1.8**	**0.73**	**3.6**	**3**	**0.85**	**6.2**	**6**	**0.93**	**1.6**	**1.2**	**0.6**

(35)



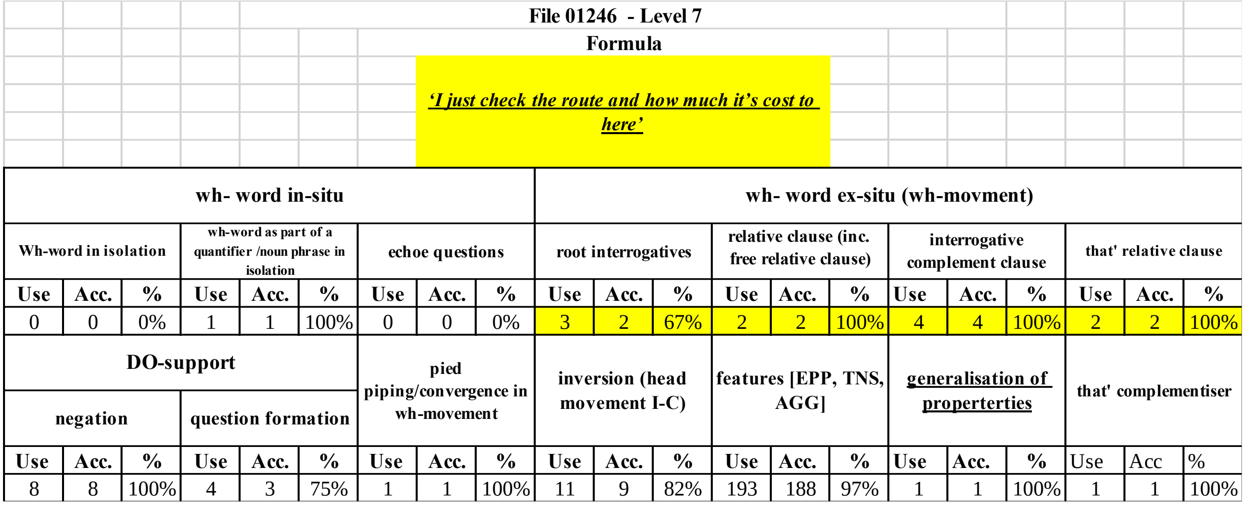
File 01224 - Level 7

Formula

‘will you tell me how much it costs?’

wh- word in-situ									wh- word ex-situ (wh-movment)											
Wh-word in isolation			wh-word as part of a quantifier /noun phrase in isolation			echoe questions			root interrogatives			relative clause (inc. free relative clause)			interrogative complement clause			that' relative clause		
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%
0	0	0%	0	0	0%	0	0	0%	1	1	100%	2	2	100%	3	2	67%	0	0	0%
DO-support						pied piping/convergence in wh-movement			inversion (head movement I-C)			features [EPP, TNS, AGG]			generalisation of properterties			that' complementiser		
negation			question formation																	
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc	%
10	10	100%	0	0	0%	0	0	0%	3	3	100%	112	111	99%	0	0	0%	4	3	75%

(36)



File 01246 - Level 7

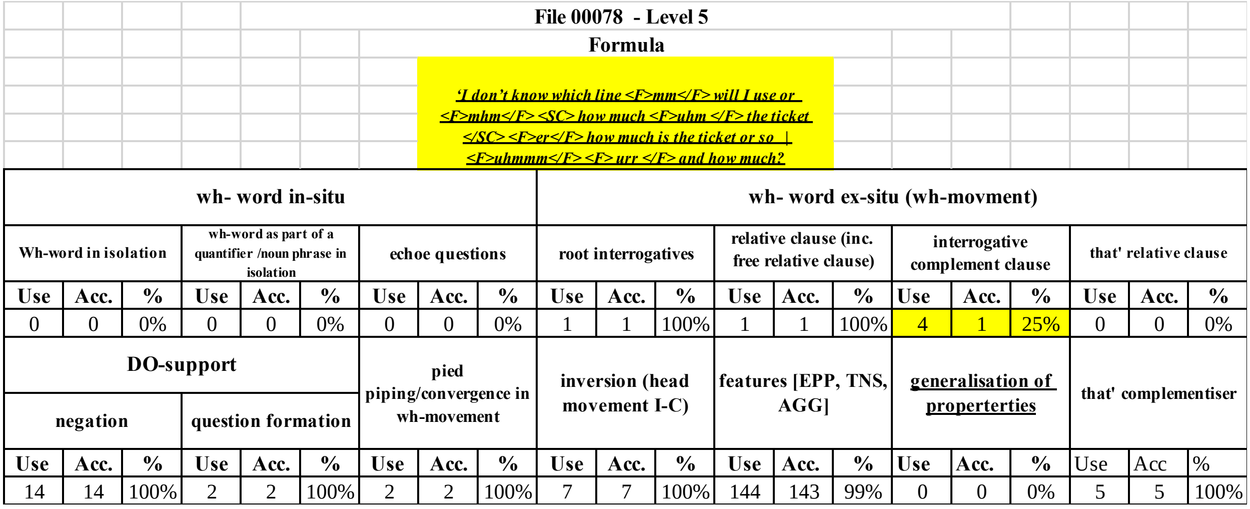
Formula

'I just check the route and how much it's cost to here'

wh- word in-situ									wh- word ex-situ (wh-movment)												
Wh-word in isolation			wh-word as part of a quantifier /noun phrase in isolation			echoe questions			root interrogatives			relative clause (inc. free relative clause)			interrogative complement clause			that' relative clause			
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	
0	0	0%	1	1	100%	0	0	0%	3	2	67%	2	2	100%	4	4	100%	2	2	100%	
DO-support						pied piping/convergence in wh-movement			inversion (head movement I-C)			features [EPP, TNS, AGG]			generalisation of properterties			that' complementiser			
negation			question formation																		
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc	%	
8	8	100%	4	3	75%	1	1	100%	11	9	82%	193	188	97%	1	1	100%	1	1	100%	

Further, the one learner of level 5 who attempts to integrate *'how much is it'* into an interrogative complement clause (as shown in table 23), can also be seen to have attempted this respectively in their propositional language (and equally unsuccessfully).

(37)



File 00078 - Level 5

Formula

'I don't know which line <F>mm</F> will I use or <F>mhm</F> <SC> how much <F>uhm </F> the ticket </SC> <F>er</F> how much is the ticket or so | <F>uhmmm</F> <F> urr </F> and how much?

wh- word in-situ									wh- word ex-situ (wh-movment)												
Wh-word in isolation			wh-word as part of a quantifier /noun phrase in isolation			echoe questions			root interrogatives			relative clause (inc. free relative clause)			interrogative complement clause			that' relative clause			
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	
0	0	0%	0	0	0%	0	0	0%	1	1	100%	1	1	100%	4	1	25%	0	0	0%	
DO-support						pied piping/convergence in wh-movement			inversion (head movement I-C)			features [EPP, TNS, AGG]			generalisation of properterties			that' complementiser			
negation			question formation																		
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc	%	
14	14	100%	2	2	100%	2	2	100%	7	7	100%	144	143	99%	0	0	0%	5	5	100%	

We see therefore complete transparency between the learners' use of the LC and their independent generative competence, not only at the underlying property and clausal level, but also with regard to what learners perceive as a 'target' expression upon a socio-pragmatic contextual cue.

Chapter VII now looks at the development of both formulas together and offers some implications that can be drawn from the results regarding the overarching role FL plays throughout the SLA process.

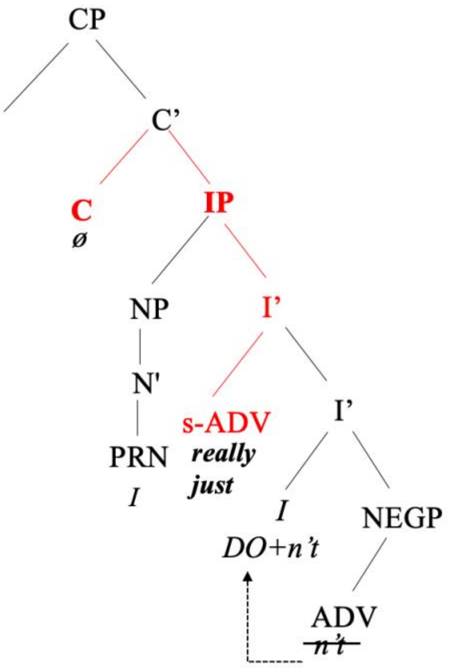
VI Further discussion and wider implications

1) Comparing the two formulas

In order to systematically address any overarching implications regarding the role of FL in L2 acquisition, it is necessary to first draw on the similarities and discrepancies between the two types of formulas under investigation. The main difference is that the PU is less transparent to the learners interlanguage grammar in early stages of acquisition, that is, learners from level 2 are shown to produce '*I don't know'* accurately with no other accurate use- or indeed occurrences- of its syntactic properties in their propositional language. This is different for the LC, which shows development in line with generative competence from the initial stages, that is, we do see underdeveloped interlanguage variations of *'how much is it/does it cost'*, and no learners, for example, were shown to produce the formula accurately when syntactic properties were correspondingly lacking in their grammar of novel language constructions. Such a difference in development ultimately supports the dichotomy proposed by Myles & Cordier (2017), in that the two types of formulaicity under analysis do indeed appear to be distinct phenomena (at least at the initial stages of acquisition), with the PU seemingly allowing lower-level learners to benefit from a more holistic production, whilst the LC less so.

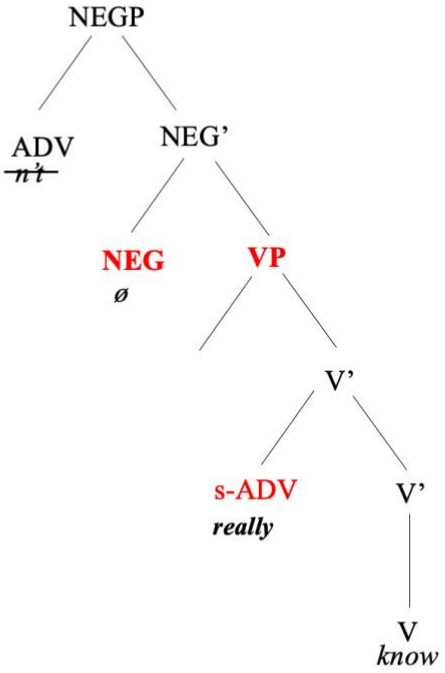
We can however draw parallels between the two types of formulas. Each seems to be sensitive to internal and external modifications that can be classed as 'advanced/complex' varieties, at both the grammatical and clausal level. Section 5.2 of this chapter stated how for the LC, these mainly involve changes to the functional categories whilst the lexical ones are fixed, a concept which could be extended to the PU if we examine the nature of the internal modifications in more detail. Figure (22) showed how more advanced modifications of *'I don't know'* involved s-adverb adjunction at the [I] and [VP] level, used only by learners of proficiency level 8. If we follow Chomsky's (1986a) Adjunction Prohibition, which allows adjunction to complements of *functional* heads only, we can see how this restriction seems to be accountable for the derivations of these more advanced learner productions. For example, the tree in (38) shows how the s-ADVs (*really, just*) in learner productions such as *'I* ***really*** *don't know'* and *'I* ***just*** *don't know'* are adjunctions to the complement IP of the functional head C. Similarly, (39) shows how in the learner production *'I don't **really** know'*, the s-ADV (*really)* is an adjunction to the complement VP of the functional head NEG[22].

(38)



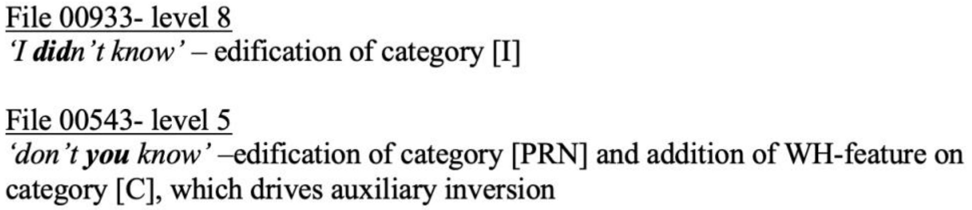
[22] NEG can be classified as a functional head of NEGP on the basis that functional elements, whether expressed in the morphology or the syntax, include 'markers of tense, subject-verb agreement…and negation' (Hegarty, 2011: 14).

(39)



On this basis, although the intensifiers themselves belong to the lexical category ADV, their distribution internal to the PU can be seen as somewhat constrained by the functional categories which govern them. Further, through a closer look at their propositional language, there is evidence that some learners also modify the PU at functional categories C, I and PRN, examples of such are given below in (40).

(40)

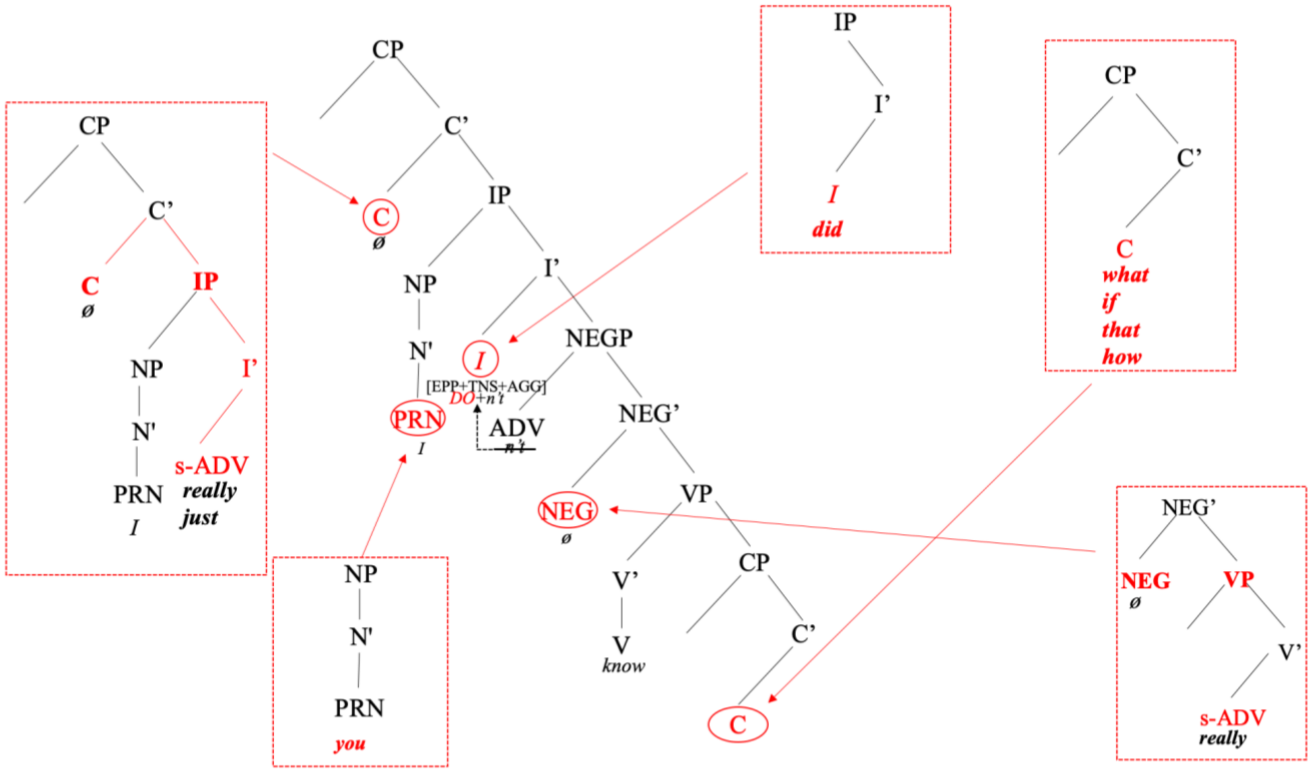


File 00933- level 8
*‘I **didn**’t know’* – edification of category [I]

File 00543- level 5
*‘don’t **you** know’* –edification of category [PRN] and addition of WH-feature on category [C], which drives auxiliary inversion

It is therefore possible to draw a parallel between both types of formulas and their advanced modifications at the functional categories, as they are gradually integrated into the propositional language of individual speakers in line with their generative competence. We could posit the structural nature of such modifications for the PU as figure (40) below, which corresponds to that of the LC in (27).

(41)



2) Wider implications for the role of formulaic language in second language acquisition

As discussed in section 2 of chapter II, a prominent view of usage-based proposals of SLA is that formulas are the seeds from which L2 learners extract (and by virtue acquire) syntax, with intuition of grammar existing only via the statistical interpretations which are drawn from patterns of frequency from the input. The results in this investigation confirm the position of Myles & Cordier (2017), who suggest that a 'learner-internal' approach to formulaicity is favourable for any model which aims to investigate such a concept. That is, *'I don't know'* seems a better candidate as a formula for which syntax could be extracted; the PU clearly has more psycholinguistic reality in the learners' minds as it is produced fluently even at the very early stages of proficiency, whereas *'how much is it/does it cost'* replicates the grammatically transparent development of conventional expressions presented in Bardovi-Harlig & Stringer (2017).

The claim however that learners extract syntax from the 'patterns' of advanced, model formulas implies that the syntactic properties of these formulas are able to be extended across novel language productions. As mentioned previously, this is not what our results suggest; instead the learners of level 2 show significantly low rates of use and accuracy of syntactic properties related to the PU (including a complete absence of DO-support), whilst they produce *'I don't know'* fluently, indicating that the syntax at this stage has not been appropriately extracted and/or acquired. Perhaps more indicative are the rates of generalisation of the syntactic properties used in combination outside of the PU. If *'I don't know'* acts as a model pattern, we would expect that replication of this pattern as a whole (i.e. the syntactic properties in combination) across identical structures to be a desirable and accessible production strategy for the learners. Again, this is not the case, as exemplified in novel attempts at negation by learners in level 2 presented previously in (19) and (20) (i.e. ***no*** *remember*). What we see instead, is that the rates of generalisation actually *increase* with proficiency level, that is, the more developed the propositional language, the more the syntactic properties of the PU are used in combination outside of it. From these results, we could infer that it seems unlikely that learners are using *'I don't know'* as a template from which to extract syntax, but rather other propositional constructions of a similar syntactic complexity are possible only when the respective generative competence is appropriately developed. A natural criticism/counter to this idea would be that such syntactic properties are developed gradually as the learners have continuous access to the patterns of PU's, which are essentially the syntactic properties exemplified in a preserved, target form. This means that, for learners in level 2, although they haven't yet extracted the properties of *'I don't know'*, they continuously work on these throughout the acquisition process[23] until they are able to realise them at later stages in their development. However, this again seems unlikely, as we have seen that learners from levels 5 and 8 do not seem to be conceptualising *'I don't know'* as a target; it instead shows sensitivity to modification and effectively becomes integrated into their propositional language constructions. Although this study does not have access to the individual's own language development and can therefore only provide limited implications on this, the broader picture that emerges from these sample learners is that independent development of generative competence is what determines the learners use and accuracy of syntax, rather than fluent productions of an unedited, target PU.

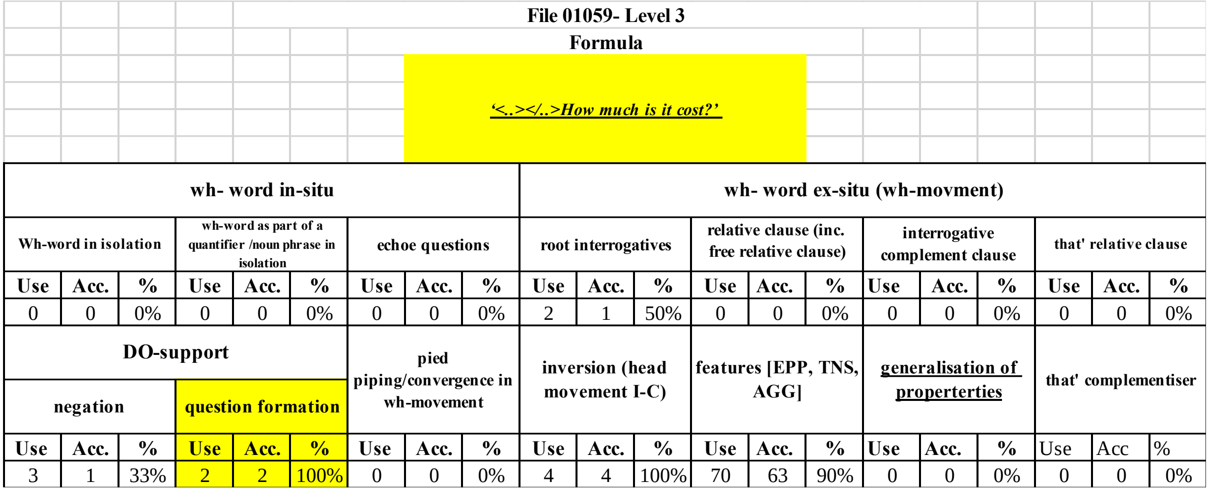
The role of the PU in this study for these particular learners then supports the views represented in section II: (3), as it seems to be limited to the communicative benefits which come with quick, holistic production of a more advanced and contextually appropriate phrase for learners at lower levels of proficiency. The results show that indeed, the prosody of a PU can be in place well in advance of an appropriate syntactic representation (Peters, 1977), a

[23] see Myles 2004 for a similar outline of development.

basis for which Carroll (2010) implies is one of the fundamental issues with usage-based proposals that view FL as input for syntactic development. As our PU is by definition a phonologically coherent string of sounds, the idea that morpho-syntax can emerge from the breakdown of such is difficult to pertain, when it is considered that 'grouping processes are not equivalent to identification processes' (Caroll 2010: 231). It is indeed hard to see how a sound form segmented from the input of speech can develop into a grammatical structure, in other words, 'syntactic structure cannot be picked up for free simply through paying attention to the sound stream' (Bardovi-Harlig & Stringer 2017).

The results from the present investigation also suggest that such a holistic storage of formulas may not be in fact as advantageous to the learner as studies have tended to claim. Following from Jackendoff's (2002) parallel architecture is Wray's (2008) Morpheme Equivalent Unit (MEU), which postulates a similarly heteromorphic mental lexicon in which formulas are processed like a morpheme, that is, 'without recourse to any form-meaning matching of any sub-parts it may have' (p. 12). Within this model, Wray builds on Peters (1983) and states how the content of the lexicon is determined through a Needs-Only-Analysis (NOA) where formulas are not broken down into their individual component parts unless there is an individual need for this, essentially deeming 'chunking' as the default strategy with syntactic analysis a secondary one (2008: p. 17). We can see in our results that such a lack of analysis/decomposition is actually what leads to some inaccuracies in learner productions. The table in (42) shows how a learner of level 5 has extended '*how much is it'* when aiming for the target *'how much does it cost'*, even though DO-support in their propositional question formation structures shows a 100% accuracy rate.

(42)



File 01059- Level 3

Formula

'<..></..>How much is it cost?'

wh- word in-situ									wh- word ex-situ (wh-movment)											
Wh-word in isolation			wh-word as part of a quantifier /noun phrase in isolation			echoe questions			root interrogatives			relative clause (inc. free relative clause)			interrogative complement clause			that' relative clause		
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%
0	0	0%	0	0	0%	0	0	0%	2	1	50%	0	0	0%	0	0	0%	0	0	0%
DO-support						pied piping/convergence in wh-movement			inversion (head movement I-C)			features [EPP, TNS, AGG]			generalisation of properterties			that' complementiser		
negation			question formation																	
Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc.	%	Use	Acc	%
3	1	33%	2	2	100%	0	0	0%	4	4	100%	70	63	90%	0	0	0%	0	0	0%

Similarly, with the PU, a level 5 learner extends the verb segment of *'I don't know'* which results in errors of [AGG] features, even though the accuracy of these are substantially high in their propositional language (88%).

(43)

	File 00247- Learner Level 5															
			Formula													
				'I don't know what to say this' *'some sales person don't know very well'*												
		DO-support			**Cliticization**			**Cat NEGP**			**Features**			**Generalisation**		
		Use	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**
		8	7	88%	18	18	100%	10	9	90%	96	84	88%	7	7	100%

Related is a learner of proficiency level 8, who overextends the [TNS] features of the PU (-past) where instead this should be (+past), even though their propositional language shows 99% accuracy rate of [TNS] features.

(44)

File 01253- Learner Level 8														
Formula														
		'I don't know what was happening' ***'I don't really know'***												
DO-support			**Cliticization**			**Cat NEGP**			**Features**			**Generalisation**		
Use	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**	**Use**	**Acc.**	**%**
4	3	75%	52	52	100%	13	13	100%	179	177	99%	0	0	0%

We could suggest that in these situations, if the default processing strategy of formulas was complete syntactic analysis rather than holistic production, then such errors would have been bypassed, as the learners' propositional language shows they have the generative capacity to realise the corresponding syntactic properties accurately. Indeed, such a proposal has been tested previously in the literature with research aiming to determine whether compound nouns were faster or slower to process than single nouns. For example, Fiorentino & Poppel (2007) found that compound words with high frequency internal morphemes were processed faster than singular nouns in learner decision tasks, a concept that Bardovi-Harlig & Stringer (2017) extend to their data of developmental conventional expressions, in that 'L2 speakers will automatically assign structure whenever possible, in accordance with their interlanguage grammars' (p. 83). Results from the present study suggest that the concept of 'whenever possible' is dependent on the learner's generative competence; perhaps structural decomposition of formulas only becomes possible when this is appropriately developed. This is hence why it is only higher-level learners who seem to assign structure to the PU as it becomes integrated into their propositional language.

VII Conclusions

This dissertation has built on recent formalist approaches to the role of FL in SLA, the results of which ultimately support the dichotomoy of formulaicity as suggested by Myles & Cordier (2017) and the grammatically transparent development of linguistic clusters as shown in Bardovi-Harlig & Stringer (2017). Drawing on similarities between the formulas, it has been suggested that generative competence is what drives both the integration of formulas into propositional language at higher developmental levels, as well as the choice of 'target formula' available to the learners in the language they are acquiring, both of which could be interpreted using Jackendoff's (2002) model of parallel architecture. The implications for the role of FL in the overall acquisition process for these particular learners under investigation then can be understood through a distinction between their performance (usage) and linguistic/generative competence (grammar). In taking a learner-internal approach, the role of the PU seems to be limited to the faster processing advantages that come with the use of such upon a socio-pragmatic cue, which allows lower-level learners to engage in more advanced linguistic performance. Such performance is not an accurate representation of their generative competence however, and generalisation/extraction of the PU's compositional syntactic properties does not seem to be a strategy used by these learners. Generalisation of syntactic properties instead increases with proficiency level as the development of generative competence makes this possible. The suggestion is therefore made that a complete breakdown analysis of a formula's compositional syntactic parts, rather than holistic production, could be a favourable strategy to overcome the learners' overextension of formulas in grammatically inappropriate contexts.

The limitations of this pilot study have been addressed throughout, and it should be added here that any implications of the present study can bear evidence to Japanese L1 learner productions only. It can be assumed that properties such as inversion and wh-movement are very different to their L1, and this dissertation did not allow for room to address any issues of transfer. Whilst this study does therefore not warrant any sweeping generalisations or conclusions, the data can be seen as encouraging of insightful implications that could be built upon through addressing its limitations and by carrying out a similar investigation on a large-scale basis.

VIII Glossary of some terms

Although this study makes a distinction between a **processing unit** and a **linguistic cluster**, the term **'formula'** and '**formulaic language (FL)**' is often used anaphorically throughout to refer to both of these. This is a stylistic choice only.

Interlanguage in its traditional sense refers to the system of mental representations influenced by both the learners L1 and L2 (Archibald, 2011), resulting in a grammar specific to the individual learner.

Processing is used throughout to refer to how speakers convert language input into relevant phonological, semantic and syntactic structure to form meaning and consequently enable production.

Propositional language contrasts to ‘automatic’ and ‘gestalt’ language and refers to creative language generated by rule-governed, grammatical processes (Krashen & Scarcella 1978). The terms ‘propositional’ and ‘novel’ language are used interchangeably throughout.

Word count: 15,729

IX References

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL*, 7-13.

Archibald, J. (2011). Second Language Acquisition. In O’Grady, W., Archibald, J., & Katamba, F. *Contemporary Linguistics: An introduction* (ed. 2). Pearson Education Limited. 394-428.

Baker, P (2006). *Using corpora in discourse analysis*. A&C Black.

Bardovi-Harlig, K & Vellenga, HE. (2012). The effect of instruction on conventional expressions in L2 pragmatics. *System* 40: 77–89.

Bardovi-Harlig, K. (2014). Awareness of meaning of conventional expressions in second language pragmatics. *Language Awareness* 23: 41–56.

Bardovi-Harlig, K., & Stringer, D. (2017). Unconventional expressions: Productive syntax in the L2 acquisition of formulaic language. *Second Language Research*, *33*(1), 61-90.

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, *8*(4), 243-257.

Biber, D., Conrad, S., & Cortes, V. (2004) "If you look at . . .": Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*, 371–405.

Boersma, P., & Weenink, D. (2005). Praat. *Doing phonetics by computer (Version 5.1)*.

Bohn, O. S. (1986). Formulas, frame structures, and stereotypes in early syntactic development: Some new evidence from L2 acquisition. *Linguistics*, *24*(1), 185-202.

Borjars, K. & Burridge, K. (2013). *Introducing English Grammar*. Routledge.

Carroll, SE. (2010). Explaining how learners extract ‘formulae’ from L2 input. *Language, Interaction and Acquisition* 1: 229–50.

Cazden, C., Cancino, H., Rosansky, E., & Schumann, J. (1975). Second language acquisition in children, adolescents and adults. *Final Report. US Department of Health, Education and Welfare*.

Cheng, W. (2011). *Exploring corpus linguistics: Language in action*. Routledge.

Chomsky, N. (1981). *Lectures on Government and Binding*. Foris, Dordrecht.

Chomsky, N. (1982). *Some Concepts and Consequences of the Theory of Government and Binding*, MIT Press, Cambridge, MA.

Chomsky, N. (1986a). *Knowledge of Language: Its Nature, Origin and Use*. Praeger, New York.

Chomsky, N. (1995). *The Minimalist Program*. MIT Press, Cambridge, MA.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied psycholinguistics*, *27*(1), 3.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*.

Coulmas, F. (1994). Formulaic language. In R. Asher (Eds.), *Encyclopedia of language and*

Divjak & Cadwell-Harris, C.L. (2015). Frequency and entrenchment. In E. Dbrowska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics* 53-74. Berlin and New York, NY: Mouton de Gruyter.

Ebeling, S. O., & Hasselgård, H. (2015). Learner corpora and phraseology. *The Cambridge handbook of learner corpus research*, 207-230.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287–314). Ann Arbor: University of Michigan Press.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual review of applied linguistics*, *32*, 17-44.

Ellis, N. C., & Wulff, S. (2015). *Second language acquisition*. De Gruyter Mouton.

Eskildsen, S.W.,& Cadierno, T. (2007). Are recurring multi-word expressions really syntactic freezes? Second language acquisition from the perspective of usage-based linguistics. In M. Nenonen, & S. Niemi (Eds.), *Collocations and idioms 1: Papers from the first Nordic conference on syntactic freezes* 86–99. Joensuu: Joensuu University Press.

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. Stuttgart, Germany: University of Stuttgart.

Fillmore, C.J. (1979). On fluency. In C.J Fillmore, D. Kempler & S.- Y.W.Wang (eds.) *Individual differences in language ability & language behaviour*. New York: Academic Press, 85- 101.

Fillmore, L. W. (1976). *The second time around: Cognitive and social strategies in second language acquisition* (Vol. 1). Dept. of Linguistics.

Fiorentino, R. & Poeppel, D. (2007). Compound words and structure in the lexicon. *Language and Cognitive Processes* 12: 953–1000.

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In *Discourse in the professions: Perspectives from corpus linguistics*, *11*, 33.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In: Bygate M, Skehan P and Swain M (eds) *Researching pedagogical tasks: Second language learning, teaching and testing*. Harlow: Longman, pp. 75–93.

Goldman-Eisler, F. (1964). Hesitation, information, and levels of speech production. In A. De Reuck & M. O'Connor (Eds.), *Disorders of language* (pp. 96–111). London: J & A Churchill.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. *Phraseology: Theory, analysis, and applications*, *145*(160), 3-18.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. *Phraseology: An interdisciplinary perspective*, *27*, 49.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. *Phraseology: An interdisciplinary perspective*, 3-25.

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition 1. *Language learning*, *24*(2), 287-297.

Hanania, E. A., & Gradman, H. L. (1977). Acquisition of English structures: A case study of an adult native speaker of Arabic in an English-speaking environment. *Language Learning*, *27*(1), 75-91.

Hawkins, R. (2001). *Second language syntax: A generative introduction*. Wiley-Blackwell.

Hegarty, M. (2011). *A feature-based syntax of functional categories: The structure, acquisition and specific impairment of functional systems* (Vol. 79). Walter de Gruyter.

Hopper, P. (1987). Emergent grammar. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 13, pp. 139-157).

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

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge: MIT Press.

Kanno, Y. (1993). Do formulaic utterances cease to be ‘chunks’ when they are analyzed? *MITA Working Papers in Linguistics* 3: 75–92.

Kecskes, I (2019). Formulaic language and its place in intercultural pragmatics. In *Understanding formulaic language: A second language acquisition perspective*. 132-150. New York: NY: Routledge.

Krashen, S., & Scarcella, R. (1978). On routines and patterns in language acquisition and performance 1. *Language learning*, *28*(2), 283-300.

Lardiere, D. (1998b). Dissociating syntax from morphology in a divergent L2 end-state grammar. In *Second Language Research* 14, 359-375.

Leech, G. (1992). Corpora and theories of linguistic performance. In *Directions in corpus linguistics*, *1992*, 105-122.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In *Corpus linguistics and the web*.133-149. Brill Rodopi.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburgh University Press.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Mellow, J. D. (2008). The emergence of complex syntax: A longitudinal case study of the ESL development of dependency resolution. *Lingua*, *118*(4), 499-521.

Myles, F. (2004). From data to theory: The over-representation of linguistic knowledge in SLA. *Transactions of the Philological Society*, *102*(2), 139-168.

Myles, F., & Cordier, C. (2017). Formulaic sequence (FS) cannot be an umbrella term in SLA: Focusing on psycholinguistic FSs and their identification. *Studies in Second Language Acquisition*, *39*(1), 3-28.

Myles, F., Hooper, J., & Mitchell, R. (1998) Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language learning*, *48*(3), 323-364.

O'Grady, W. (2011). Syntax: The analysis of sentence structure. In O'Grady, W., Archibald, J., & Katamba, F. *Contemporary Linguistics: An introduction* (ed. 2). Pearson Education Limited., 153-197.

Oakes, M. (1998). *Statistics for corpus linguistics*. Edinburgh, UK: Edinburgh University Press.

Osborne, J. (2008). Phraseology effects as a trigger for errors in L2 English: The case of more advanced learners. In: Meunier F and Granger S (eds) *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins, pp. 67–83.

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* 191–225. London, UK: Longman.

Peters, A. M. (1977). Language learning strategies: Does the whole equal the sum of the parts?. *Language*, 560-573.

Peters, A. M. (1983). *The units of language acquisition* (Vol. 1). CUP Archive.

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory.* Amsterdam: John Benjamins.

Potsdam, E. (1998). A syntax for adverbs. In *The proceedings of the twenty-seventh western conference on linguistics* (397-411).

Radford, A. (2004a). *Minimalist Syntax: Exploring the Structure of English*. Cambridge University Press.

Radford, A. (2009). *Analysing English sentences: A minimalist approach*. Cambridge University Press.

Radford, A., Atkinson, M., Britain, D., Clahsen, H. and Spencer, A. (1999). *Linguistics: An introduction*. Cambridge University Press.

Rehbein, J. (1987). On fluency in second language speech. In H. Dechert & M. Raupach (Eds.), *Psycholinguistic models of production* 97–105. Norwood, NJ: Ablex.

Rizzi, L. (1990). *Relativised Minimality*. MIT Press, Cambridge, MA.

Ross, J. R. (1967). Constraints on Variables in Syntax, PhD dissertation, MIT (published as Infinite Syntax! by Ablex Publishing Corporation, Norwood, NJ, 1986).

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid. *Formulaic sequences: Acquisition, processing and use*, 127-151.

Schumann, J. (1979). The acquisition of English negation by speakers of Spanish: A review of the literature. In *The acquisition and use of Spanish and English as first and second languages*, 3-32.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*(2), 251-272.

Slabakova, R. (2008) *Meaning in the second language*. (Vol. 34). Walter de Gruyter.

Tono, Y., Kaneko, T., Isahara, H., Saiga, T. & E. Izumi. (2001). The Standard Speaking Test Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. In S. Lee, ed., *ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary*.(pp. 257-262). The Second Asialex International Congress, August 8-10, 2001, Yonsei University, Korea.

Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus linguistics and linguistic theory*, *1*(2), 225-261.

Vainikka, A. & Young-Scholten, M. (1996). Gradual development of L2 phrase structure. *Second Language Research* 12, 7-39.

Weinert R (2010). Formulaicity and usage-based language: Linguistic, psycholinguistic and acquisitional manifestations. In: Wood D (ed.) *Perspectives on formulaic language: Acquisition and communication*. London: Continuum, pp. 1–20.

White, L. (2003). Fossilization in steady state L2 grammars: Persistent problems with inflectional morphology. In *Bilingualism: language and cognition*, *6*(2), 129-141.

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

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford University Press.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, *32*, 231-254.

Wray, A., & Fitzpatrick, T. (2008). Why can't you just leave it alone? Deviations from memorized language as. *Phraseology in foreign language learning and teaching*, *123*.

Wulff, S. (2019) Acquisition of formulaic language from a usage-based perspective. In *Understanding formulaic language: A second language acquisition perspective*, 19-37. New York: NY: Routledge.

Yang, C. D. (1999). Unordered Merge and its linearization. *Syntax* 1: 38 64.